EDUCACIÓN
SECRETARÍA DE EDUCACIÓN PÚBLICA

Instituto Politécnico Nacional
"La Técnica al Servicio de la Patria"

Research in Computing Science

Vol. 152 No. 9
September 2023

Research in Computing Science

Series Editorial Board

Editors-in-Chief:

Grigori Sidorov, CIC-IPN, Mexico
Gerhard X. Ritter, University of Florida, USA
Jean Serra, Ecole des Mines de Paris, France
Ulises Cortés, UPC, Barcelona, Spain

Associate Editors:

Jesús Angulo, Ecole des Mines de Paris, France
Jihad El-Sana, Ben-Gurion Univ. of the Negev, Israel
Alexander Gelbukh, CIC-IPN, Mexico
Ioannis Kakadiaris, University of Houston, USA
Petros Maragos, Nat. Tech. Univ. of Athens, Greece
Julian Padget, University of Bath, UK
Mateo Valero, UPC, Barcelona, Spain
Olga Kolesnikova, ESCOM-IPN, Mexico
Rafael Guzmán, Univ. of Guanajuato, Mexico
Juan Manuel Torres Moreno, U. of Avignon, France

Editorial Coordination:
Griselda Franco Sánchez

Research in Computing Science, Año 22, Volumen 152, No. 9, septiembre de 2023, es una publicación mensual, editada por el Instituto Politécnico Nacional, a través del Centro de Investigación en Computación. Av. Juan de Dios Bátiz S/N, Esq. Av. Miguel Othon de Mendizábal, Col. Nueva Industrial Vallejo, C.P. 07738, Ciudad de México, Tel. 57 29 60 00, ext. 56571. https://www.rcs.cic.ipn.mx. Editor responsable: Dr. Grigori Sidorov. Reserva de Derechos al Uso Exclusivo del Título No. 04-2019-082310242100-203. ISSN: en trámite, ambos otorgados por el Instituto Politécnico Nacional de Derecho de Autor. Responsable de la última actualización de este número: el Centro de Investigación en Computación, Dr. Grigori Sidorov, Av. Juan de Dios Bátiz S/N, Esq. Av. Miguel Othon de Mendizábal, Col. Nueva Industrial Vallejo, C.P. 07738. Fecha de última modificación 01 de septiembre de 2023.

Las opiniones expresadas por los autores no necesariamente reflejan la postura del editor de la publicación.

Queda estrictamente prohibida la reproducción total o parcial de los contenidos e imágenes de la publicación sin previa autorización del Instituto Politécnico Nacional.

Research in Computing Science, year 22, Volume 152, No. 9, September 2023, is published monthly by the Center for Computing Research of IPN.

The opinions expressed by the authors does not necessarily reflect the editor's posture.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without prior permission of Centre for Computing Research of the IPN.

Volume 152(9)

Advances in Artificial Intelligence

Hiram Calvo (ed.)

Instituto Politécnico Nacional, Centro de Investigación en Computación
México 2023

ISSN: in process

Copyright © Instituto Politécnico Nacional 2023
Formerly ISSNs: 1870-4069, 1665-9899

Instituto Politécnico Nacional (IPN)
Centro de Investigación en Computación (CIC)
Av. Juan de Dios Bátiz s/n esq. M. Othón de Mendizábal
Unidad Profesional "Adolfo López Mateos", Zacatenco
07738, México D.F., México

http://www.rcs.cic.ipn.mx
http://www.ipn.mx
http://www.cic.ipn.mx

The editors and the publisher of this journal have made their best effort in preparing this special issue, but make no warranty of any kind, expressed or implied, with regard to the information contained in this volume.

All rights reserved. No part of this publication may be reproduced, stored on a retrieval system or transmitted, in any form or by any means, including electronic, mechanical, photocopying, recording, or otherwise, without prior permission of the Instituto Politécnico Nacional, except for personal or classroom use provided that copies bear the full citation notice provided on the first page of each paper.

Indexed in LATINDEX, DBLP and Periodica

Electronic edition

ISSN 1870-4069

Table of Contents

Page

ISSN 1870-4069

Detección temprana de la enfermedad de Parkinson mediante técnicas de aprendizaje profundo

Jaime Hernández-Ramírez[1], Giner Alor-Hernández[1], Nancy Aracely Cruz-Ramos[1], José Luis Sánchez-Cervantes[2], Lisbeth Rodríguez-Mazahua[1]

[1] Tecnológico Nacional de México, Instituto Tecnológico de Orizaba, México

[2] Consejo Nacional de Ciencia y Tecnología, Tecnológico Nacional de México, Instituto Tecnológico de Orizaba, México

dci.ncruz@ito-depi.edu.mx,{M21011175,giner.ah, jose.sc, lisbeth.rm }@orizaba.tecnm.mx

Resumen. La enfermedad de Parkinson (EP) es un trastorno neurodegenerativo progresivo que afecta a más de 10 millones de personas en todo el mundo, representando una importante disminución de la calidad de vida del paciente y sus familiares. Es por ello que, el diagnóstico temprano representa un papel fundamental en el tratamiento de los síntomas. Sin embargo, a pesar de los esfuerzos realizados para la recolección de información clínica, muchos de los métodos no proporcionan la precisión y sensibilidad suficiente para la detección temprana de la EP. En este trabajo, se desarrolló un módulo web para la detección temprana de la EP mediante ejercicios de trazabilidad utilizando técnicas de Aprendizaje Profundo. Los resultados presentan estadísticas favorables en precisión y eficacia, destacando los valores obtenidos por el algoritmo Random Forest. Además, este enfoque se muestra como herramienta de gran potencial en el apoyo al sector médico en la toma de decisiones para mejorar la calidad de vida del paciente.

Palabras clave: Aprendizaje profundo, enfermedades neurodegenerativas, inteligencia artificial, Parkinson.

Early Detection of Parkinson's Disease Using Deep Learning Techniques

Abstract. Parkinson's disease (PD) is a progressive neurodegenerative disorder that affects over 10 million people worldwide, representing a significant decrease in the patient's quality of life and their families. Therefore, early diagnosis plays a crucial role in symptom treatment. However, despite efforts made for clinical information collection, many methods do not provide sufficient accuracy and sensitivity for early detection of PD. In this work, a web module for early

detection of PD using traceability exercises and Deep Learning techniques was developed. The results present favorable statistics in accuracy and efficacy, highlighting the values obtained by the Random Forest algorithm. Furthermore, this approach shows great potential as a tool to support the medical sector in decision-making to improve the patient's quality of life.

Keywords: Deep learning, neurodegenerative diseases, artificial intelligence, Parkinson's disease.

1. Introducción

La enfermedad de Parkinson (EP) es un trastorno neurodegenerativo progresivo que, de acuerdo con Dorsery et al. [1], afecta a más de 6 millones de personas en todo el mundo y se prevé que esta cifra para el año 2040, sea duplicada, donde además, la mayoría de personas que cuentan con este padecimiento no obtiene un diagnóstico positivo, sino hasta una edad posterior a los 50 años.

Por su parte, el Instituto Nacional de Neurología y Neurocirugía [2], considera la confirmación de 150 a 200 casos por cada 100,000 habitantes al año en diversas partes del mundo, donde tan solo en México 50 de cada 100,000 habitantes son posibles candidatos de este padecimiento.

La EP se caracteriza por ser una enfermedad que influye directamente al sistema nervioso del paciente que lo padece, esto se debe, a la muerte de células nerviosas en el cerebro encargadas de la producción de dopamina, un neurotransmisor clave en el control del movimiento y la coordinación muscular.

Algunos de los síntomas presentados en una edad temprana incluyen, temblor, rigidez, dificultar para caminar, hablar y problemas de equilibrio, resultando con el tiempo, la disminución de la calidad de vida de los pacientes y sus familiares.

A pesar de que no existe cura para la EP, el diagnóstico temprano, es fundamental para el tratamiento y monitoreo de los síntomas, es así que, se cuenta con investigaciones que dedican esfuerzos en el análisis y recolección de información clínica con el fin de brindar herramientas de apoyo al personal médico y pacientes en la detección y seguimiento de la EP, aunque, actualmente los diagnósticos basan sus resultados en la observación clínica y la evaluación de síntomas, la mayoría de estos métodos no proveen la precisión y sensibilidad suficiente para detectar la enfermedad en sus primeras etapas, delimitando tratamiento adecuado del paciente donde la enfermedad ya cuenta con un amplio desarrollo.

Por otro lado, la Inteligencia Artificial (IA) ha demostrado, en los últimos años, ser una herramienta de gran relevancia en el sector de la salud, protagonizando, entre diversos proyectos, la detección temprana, seguimiento y toma de decisiones sobre diversas enfermedades, tales como, el Parkinson, Alzheimer, el cáncer y SARS-CoV-2 (COVID-19).

Se han desarrollado propuestas que también cuentan con el impulso de técnicas de Aprendizaje Profundo (AP), rama de la IA que tiene como principal participe, prometedores modelos en sistemas de clasificación de resultados, destacando en detección temprana de la EP por sus estudios.

En este trabajo se desarrolló un módulo web para la detección temprana de la EP mediante ejercicios de trazabilidad utilizando técnicas de AP, brindando así, una

herramienta de diagnóstico temprano que busca brindar apoyo en la toma de decisiones por parte del sector médico, mejorando así, la calidad de vida del paciente.

Este trabajo presenta la siguiente estructura, la sección 2 aborda los trabajos de literatura relacionados con la detección de la EP mediante técnicas de AP. La sección 3 presenta la arquitectura del módulo propuesto durante la investigación realizada. La sección 4 presenta un caso de estudio como prueba de concepto. Finalmente, en la sección 5 se presentan las conclusiones alcanzadas durante el desarrollo de la investigación.

2. Trabajos relacionados

En esta sección, se presentan diversas investigaciones de proyectos relacionados con la detección temprana de la EP mediante AP, enfocadas en diferentes herramientas, técnicas y pruebas de diagnóstico. Como grupo principal, se presentan a continuación investigaciones que propusieron herramientas digitales para el diagnóstico y seguimiento de pacientes con la EP.

Worasawate et al. [3], analizaron los datos de grabación de voz de un teléfono inteligente como posible herramienta de autodiagnóstico médico, los resultados demostraron precisiones de clasificación con valores estimados del 97% al 99%. Similarmente, Punarselvam et al. [4] propusieron un sistema de monitorización de la señal de la voz mediante elementos de Internet de las Cosas (IoT), los resultados demostraron métricas de precisión, sensibilidad y especificidad por encima del 95%.

Por su parte, Fraiwan et al. [5] presentaron un sistema de detección del movimiento rítmico involuntario mediante una aplicación móvil enlazada con los sensores incorporados, los resultados del estudio demostraron valores de precisión de hasta 95%. De manera similar, Zhang et al. [6] enfocaron su estudio en el desarrollo de DeepVoice, una aplicación de recolección de la huella de voz en dominio del espectrograma fonético, la precisión promedio obtenida por el modelo fue de 90.45%.

Considerando la introducción de herramientas digitales, Kuosmanen et al. [7] desarrollaron STOP, una aplicación de seguimiento de síntomas y gestión de la ingesta médica, donde los resultados demostraron una aceptación positiva al uso de herramientas para el tratamiento de la EP.

Similarmente, ocurrió con los trabajos de Hu et al. [8], Martin et al. [9] y Linares-del Rey et al. [10], donde el enfoque se dirigió con respecto a la inspección sistemática sobre el uso de herramientas digitales para la gestión, seguimiento y ejecución de tareas relacionadas al estudio de la EP, dando como resultado común, factores de aceptación positivos por parte de los pacientes y personal médico.

En segunda instancia, se analizaron estudios de proyectos enfocados en la detección de la EP mediante la aplicación de ejercicios de trazabilidad y supervisión de la actividad motriz, pruebas dedicadas en evaluar el rendimiento continuo de los pacientes a través de los resultados obtenidos, estas investigaciones se mencionan a continuación. Kotsavasiloglou et al. [11] estudiaron el uso de dispositivos digitalizadores para analizar las diferencias entre el movimiento y coordinación manuscrita.

Como resultado demostraron una precisión de análisis de 91% mediante clasificadores de redes bayesianas. Camps et al. [12] proponen un método de análisis por congelación de la marcha mediante monitoreo espectral de la inercia, proceso

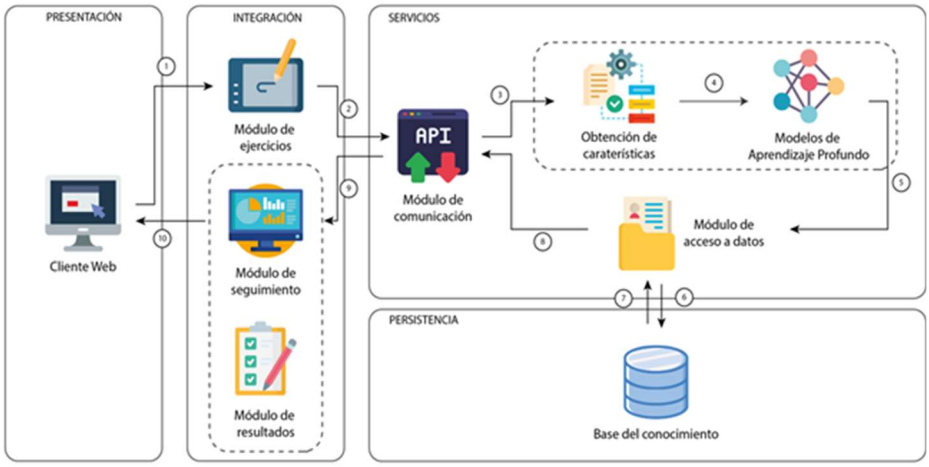


Fig. 1. Arquitectura del módulo web para la detección temprana de la EP.

impulsado por el algoritmo Support Vector Machine (SVM) con resultados de precisión de 87%.

Alkhatib et al. [13] desarrollaron un algoritmo de alto nivel enfocado en segmentar grupos de estudio en función de la distribución de carga durante la marcha, los resultados presentaron valores de precisión en un 95%.

Desde otra perspectiva, Rovini et al. [14] implementaron la recopilación de estadísticas de la actividad motriz utilizando dispositivos de seguimiento portátiles, como resultado observaron mejor precisión (95%) al implementar RF como principal algoritmo de clasificación.

Así mismo, Moshkova et al. [15] implementaron técnicas de AP para la clasificación de datos obtenidos mediante sensores Leap Motion, los clasificadores analizados destacaron a Random Forest (RF) en 90.6% como la precisión de mayor rango.

Por otro lado, investigaciones también reconocen la detección de la EP mediante el diagnostico de diversas pruebas, tales como, el análisis de temblores en el estado de reposo, muestras de estudio del lenguaje natural y seguimiento en la afección causada en neuronas dopaminérgicas, estos estudios se presentan a continuación.

Yao et al. [16] presentaron un enfoque de estudio mediante el análisis de temblores en el estado de reposo, durante la comparación de resultados, los algoritmos RF y XGBoost obtuvieron el mejor resultado de precisión (>70%). Anand et al. [17] diagnosticaron la probabilidad de padecimiento mediante el estudio de lenguaje natural, los clasificadores representados en el modelo de AP destacaron a los algoritmos KNN y RF con mejores métricas de rendimiento.

Con enfoque distinto, Wodzinski et al. [18] realizaron el estudio de espectrogramas vocales mediante una arquitectura ResNet, donde el modelo propuesto expuso valores de precisión de 91% en el uso de algoritmos como RF y SVM.

Mientras que, Prashanth et al. [19] dirigieron la investigación al estudio de la perdida de neuronas dopaminérgicas, durante el seguimiento las estadísticas resultantes, se demostró que el algoritmo SVM contó con el mayor grado de precisión obtenido.

Tabla 1. Resultados comparativos de los algoritmos de clasificación.

#	Algoritmo	Precisión	Sensibilidad	Especificidad	F1 Score
1	KNN	0.808	0.913	0.783	0.857
2	Decision Tree	0.679	0.826	0.609	0.745
3	SVM	0.792	0.826	0.783	0.809
4	GB	0.818	0.783	0.826	0.800
5	LightGBM	0.792	0.826	0.783	0.809
6	XGBoost	0.760	0.826	0.739	0.792
7	RF	0.909	0.870	0.913	0.889
8	CatBoost	0.739	0.739	0.739	0.739
9	AdaBoost	0.739	0.739	0.739	0.739

A su vez, con el enfoque determinante en la generación de estadísticas y desarrollo de nuevos modelos, se exploraron investigaciones que ofrecieron como resultados, valores de rendimiento obtenidos en la comparación de diversas técnicas y procesos de análisis, observando lo siguiente.

Arora et al. [20] realizaron pruebas de clasificación con base en registros de pacientes mediante su expediente clínico estandarizado, como resultado, RF presentó óptimas métricas de rendimiento con resultados de sensibilidad en 96.2% y 96.9%. Por su parte, Ali et al. [21] propusieron el desarrollo de un sistema de aprendizaje escalonado, demostrando una mejora de rendimiento frente al uso de clasificadores independientes, este conjunto destacó AdaBoost con un incremento de mejora en 3.3%.

De manera similar, Karapinar et al. [22] propusieron diagnosticar el rendimiento obtenido por distintos algoritmos de clasificación en la detección de la EP, la tabla de resultados demostró SVM con el mejor desempeño obtenido en valores de precisión. Del mismo modo, Haq et al. [23] durante su investigación, compararon las distintas ventajas del rendimiento que ofrecen diversas metodologías de AP en el campo de la EP, resultados destacaron a SVM como una de las principales alternativas.

De manera semejante, Pahuja et al. [24] decidieron realizar la comparación de diversas tecnologías para la detección de la EP, incorporando entre los algoritmos de mejor rendimiento en valor de precisión hallado, Levenberg-Marquardt y SVM.

Adicionalmente, La investigación Nilashi et al. [25] propuso el desarrollo de un modelo híbrido para la detección de la EP mediante múltiples algoritmos de clasificación, los resultados permitieron reconocer las ventajas tanto individuales como globales de cada uno, así como también el alto impacto obtenido al incorporar herramientas de AP en el campo de la detección de la EP.

3. Arquitectura del módulo para la detección temprana de la enfermedad de Parkinson

En este trabajo, se desarrolló un módulo web enfocado en el proceso de identificación de la EP mediante el uso de ejercicios de trazabilidad y un clasificador impulsado por una red neuronal convolucional (CNN). Este módulo tiene como función brindar una conexión entre la plataforma y el modelo de clasificación entrenado en indicar el valor probabilidad de padecimiento, de acuerdo con determinadas reglas de

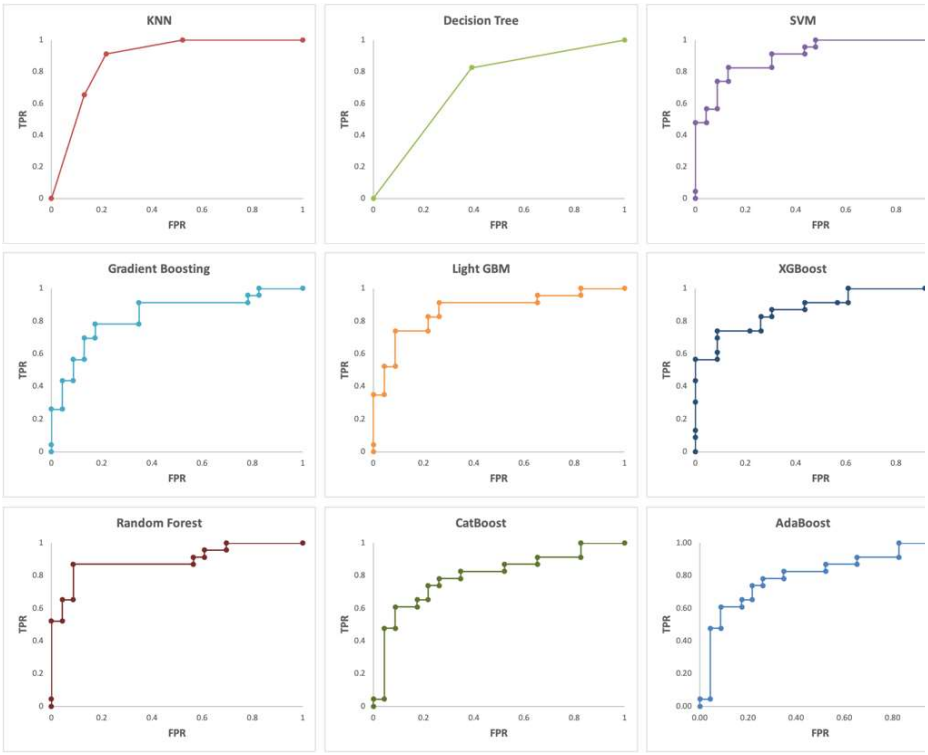


Fig. 1. Resultados de la curva ROC.

referencia. El esquema global del módulo, comienza con la plataforma web, la cual es el punto de partida dedicado al personal médico y con el objetivo de realizar pruebas de detección mediante ejercicios de trazabilidad con el paciente.

Esta plataforma cuenta con diversas secciones operativas, entra las cuales, se considera, la aplicación de nuevos estudios y revisión de la evolución del paciente, enfocado en el historial de pruebas realizadas. Durante la aplicación de nuevos estudios el módulo recopila, procesa, interpreta y recupera los resultados obtenidos para su posterior seguimiento y monitorización.

La obtención de datos es un proceso guiado mediane una tableta digitalizadora que permite al paciente ilustrar un conjunto de ejercicios de trazos con base en plantillas predefinidas. Los datos obtenidos mediante la tableta se determinan con base en las propiedades del trazo realizado, tales como la precisión, la presión, rapidéz e inclinación ejecutada por el usuario durante el seguimiento del ejercicio.

Posteriormente el resultado se convierte en un conjunto de ilustraciones y propiedades cuyo objetivo es ser enviados a un servidor externo, donde el clasificador se enfoca en el procesamiento e interpretación de datos, finalmente, los valores obtenidos se registran para seguimiento y consulta médica.

El módulo está distribuido en un modelo arquitectónico basado en capas para brindar una solución de mejor escalabilidad, robustez y de fácil mantenimiento. Durante el proceso de desarrollo se consideró el uso de diversas tecnologías, con motivo de

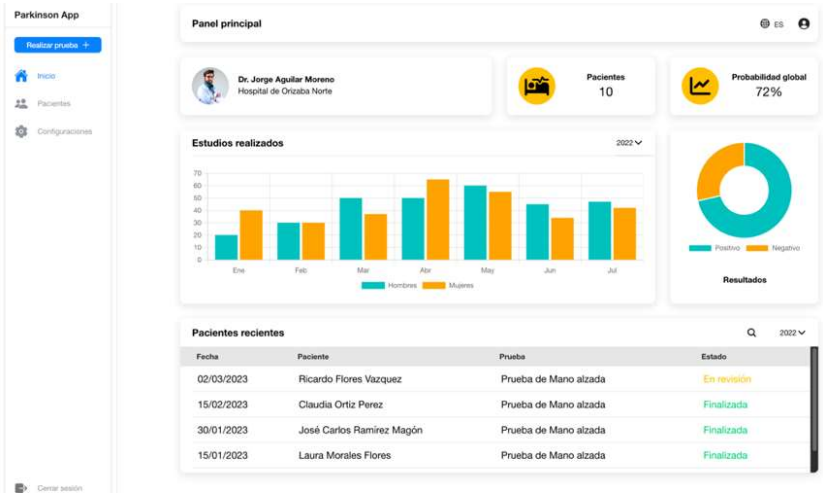


Fig. 2. Módulo web - panel principal.

garantizar un mejor rendimiento entre los procesos involucrados, así como también medidas de seguridad para garantizar la integridad de los datos.

La arquitectura del módulo desarrollado se describe en la Fig. 1, donde además se encuentra señalado a través de etiquetas enumeradas (1-10) el flujo de trabajo principal "Clasificación de una prueba de trazabilidad". A continuación, se describe cada una de las capas integradas en la arquitectura.

- **Capa de presentación:** es la aplicación web principal, encargada de proporcionar al usuario una interfaz gráfica para la recopilación y seguimiento de las pruebas, esta capa está diseñada de manera óptima en concepto de usabilidad, con el objetivo de brindar una mejor experiencia de usuario por el personal médico y su interacción con el paciente.
- **Capa de integración:** está compuesta por diferentes módulos encargados de realizar funciones específicas dentro del sistema. El módulo de ejercicios tiene por motivo, obtener los ejercicios realizados por el paciente mediante una tableta digitalizadora. El módulo de seguimiento incorpora un conjunto de estadísticos referentes al historial de resultados globales de cada paciente. El módulo de resultados tiene como objetivo presentar al médico los resultados obtenidos por el clasificador de forma clara y concisa. Finalmente, todos los datos interpretados por esta capa se envían a la capa de servicios mediante su interfaz de acceso, módulo de comunicación.
- **Capa de servicios:** compuesta también por distintos módulos determina los casos de uso de predicción de estudios y consulta de datos. Referente al proceso de predicción se establecieron los módulos de obtención de características, enfocado en la lectura y análisis de instancias gráficas; y modelos de AP, encargados de la interpretación de instancias y clasificación de resultados mediante los algoritmos previamente entrenados para su uso. El módulo de acceso a datos proporciona una interfaz de acceso a la capa de persistencia para almacenar los datos obtenidos. Finalmente, el módulo de comunicación se integra como único punto de entrada y salida, encargado de la comunicación con la capa de integración.

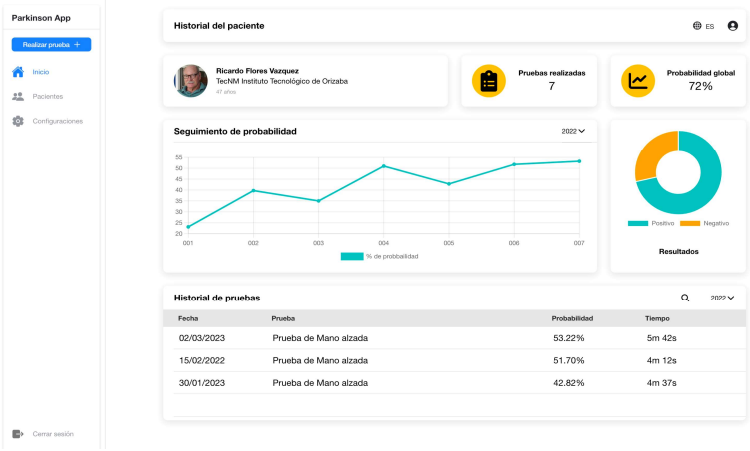


Fig. 3. Módulo web - Panel expediente del paciente.

– **Capa de persistencia:** incorpora la base de datos principal del módulo, encargada de almacenar la información necesaria para su posterior análisis y seguimiento.

4. Caso de estudio: Identificación temprana de síntomas de la enfermedad de Parkinson

En esta sección se describe el caso de estudio desarrollado para la identificación temprana de síntomas de la EP mediante ejercicios de trazabilidad y técnicas de AP.

4.1. Características de la aplicación

Como punto de partida en selección de tecnologías para el desarrollo de la aplicación web, se implementó el marco de trabajo de IONIC impulsado por la tecnología Angular, el cual proporciona amplia facilidad de uso y soporte, lo que permite una integración rápida y eficiente en el desarrollo de interfaces gráficas para el usuario, así como también, el despliegue híbrido para nuevas áreas de oportunidad como el uso de plataformas móviles [26. 27].

Asía también, se utilizó Flask de Python que es un micro marco de trabajo debido a su sencillez y flexibilidad, así como también destaca sus capacidades de integración con nuevas tecnologías, es por ello que, se determinó como herramienta ideal para la construcción del núcleo de la aplicación [28].

A su vez, para el módulo de inteligencia, encargado de la distribución de los algoritmos de clasificación, se seleccionó Scikit-Learn que es una biblioteca de AP con base de código en el lenguaje Python que proporciona herramientas simples y eficientes para la minería y análisis de datos [29]. Mediante esta biblioteca, se realizó la integración de diversos algoritmos, así como también, su comparación de métricas de rendimiento para la selección óptima en caso de estudio.

Finalmente, como base de conocimiento, se determinó MySQL para el almacenamiento de datos y control de registros, herramienta seleccionada debido a su

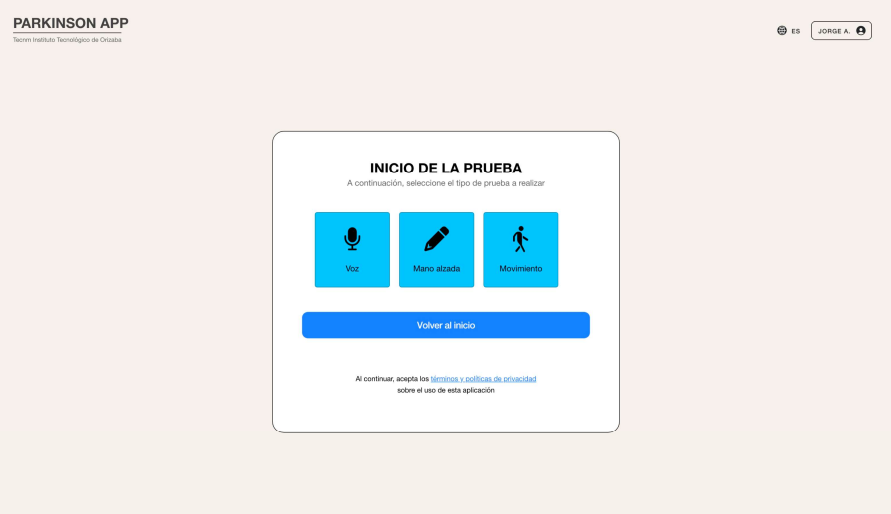


Fig. 4. Módulo web - Catálogo de pruebas.

optimización de recursos, amplio soporte y fácil integración con las herramientas de infraestructura del servidor, facilitando así, la correcta gestión de la información.

4.2. Selección de algoritmos de clasificación

Como parte fundamental del proyecto, se realizó la investigación comparativa de nueve algoritmos de clasificación de AP, mismos que se encontraron con mejores resultados dentro de la literatura, cada uno de los algoritmos se puso a prueba bajo un entorno preparado a través de un mismo conjunto de datos y métricas de estudio.

Como resultados obtenidos durante la comparación, se demostró que el clasificador RF obtuvo los valores óptimos en términos de precisión y especificidad, obteniendo así un puntaje promedio de 90.9% y 91.3% respectivamente en las pruebas de detección, como segundo clasificado, se observó a Gradient Boosting (GB) con una puntuación de precisión en 81.8% y especificidad de 82.6%. En la Tabla se observa a detalle los valores obtenidos por los clasificadores durante su análisis.

Se observa que los resultados obtenidos en esta comparación muestran que, durante la determinación de los clasificadores, RF representó los mejores puntajes de rendimiento, determinando así, el principal modelo a implementar, seguido por el algoritmo GB en términos de precisión.

Además, se llevaron a cabo pruebas comparativas entre los clasificadores mediante el estudio de la curva ROC, la cual representa la relación entre la sensibilidad y especificidad de cada modelo.

Los resultados de estas pruebas permitieron confirmar los resultados anteriores, a continuación, en la Fig. 1 se muestran los resultados obtenidos durante su análisis.

Como se observa, durante el análisis RF continúa representando un mejor balance en el resultado obtenido, seguido en este caso por el algoritmo SVM, y en tercera posición GB, dejando de lado entre demás opciones, el algoritmo KNN, que, si bien

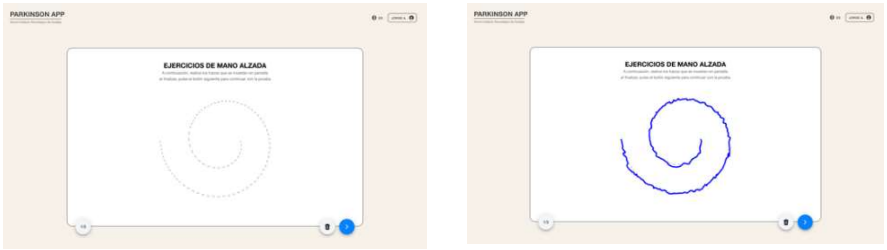


Fig. 5. Módulo web - Aplicación de la prueba.

mantiene valores óptimos de precisión y sensibilidad, no proporciona el mejor balance en el análisis de la curva ROC, y por ello, no se consideró para su implementación.

Como resultado concluyente, la comparación se realizó mediante cinco métricas y nueve algoritmos de clasificación distintos, donde se obtuvo las opciones de RF, GB y SVM, como los principales clasificadores a implementar en el módulo de detección temprana de la EP mediante técnicas de AP.

4.3. Proceso de detección

En este apartado se presenta el proceso de detección a través del módulo desarrollado, el escenario a considerar fue mediante un paciente masculino de 60 años que presenta un diagnóstico positivo de la EP.

Para ello, con asistencia médica el paciente utilizó el módulo web realizando pruebas de trazabilidad para posteriormente enviarlas al clasificador de AP, y finalmente, obtener resultados que permitieron al personal médico evaluar de forma objetiva la condición del paciente. Para el proceso de detección, el módulo cuenta con varias interfaces y pasos que permiten una aplicación clara y sencilla de las pruebas de trazabilidad, descritas a continuación:

Como punto de partida, se encuentra el panel principal del módulo (

Fig. **2**), que permite al personal médico acceder a las distintas opciones disponibles, entre ellas, el listado de pacientes registrados y el catálogo de pruebas.

En el listado de pacientes registrados, se muestra un resumen de la información de cada registro, permitiendo seleccionar al paciente con el que se encuentra trabajando. Cada paciente, cuenta su expediente, donde se registran los resultados de las pruebas realizadas y los datos clínicos relevantes, como se observa en la Fig. 3.

Determinada la selección del paciente, en el catálogo de pruebas (Fig. 4), se presentan las distintas categorías disponibles para la evaluación de la EP, enfocado para el caso de estudio las pruebas de trazabilidad.

Posterior a la selección de la prueba, el módulo presenta en pantalla un lienzo en blanco conformado por una serie de ejercicios, los cuales, se representan en plantillas de modelos de trazos en espiral y ondas, durante este proceso el paciente debe replicar la figura presentada con la mayor precisión posible tal como se observa en la Fig. 5, finalizando la prueba al concluir el último ejercicio aplicado.

Una vez finalizado el proceso de aplicación de la prueba, el médico obtiene un folio de seguimiento para la revisión de resultados, los cuales son accesibles desde el panel

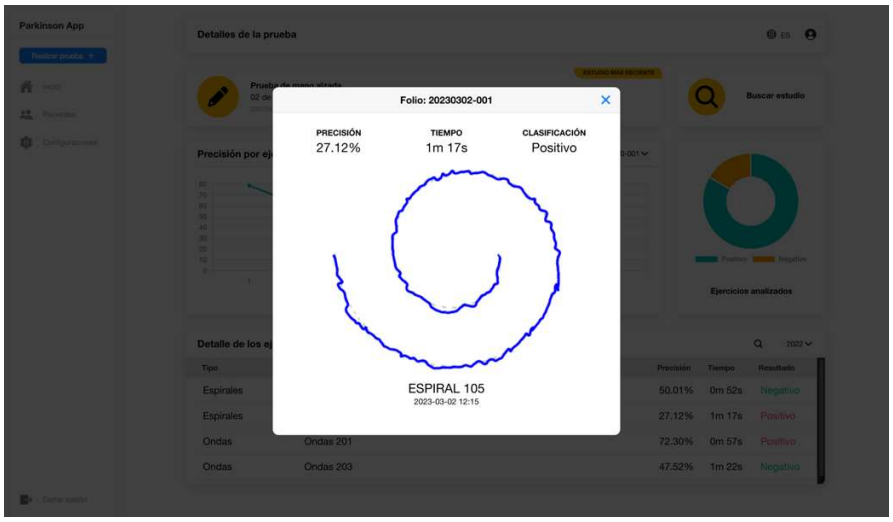


Fig. 6. Módulo web - Resultados de la prueba.

principal o panel del paciente, representando la vista previa del ejercicio y los resultados obtenidos por el clasificador, observado en la Fig. 6.

Es importante destacar que los resultados obtenidos son adecuados, esto se debe a la implementación de técnicas de AP que incorporan los procesos de clasificación para proporcionar resultados con mayor precisión, además, los algoritmos utilizados fueron seleccionados con base en métricas de rendimiento analizadas durante el proceso de estudio, determinando, así como modelo principal a los clasificadores: RF, SVM y GB.

Por otro lado, y como herramienta de apoyo en la toma de decisiones, cabe destacar que el valor de clasificación obtenido representa la probabilidad del padecimiento y no un resultado de diagnóstico determinante.

5. Conclusiones y trabajo a futuro

El enfoque en la detección temprana de la EP mediante técnicas de AP y con el desarrollo de un módulo web que permita incorporar los procesos de detección mediante ejercicios de trazabilidad, demostraron ser una herramienta efectiva, permitiendo a profesionales de salud, evaluar y clasificar con mayor precisión a los pacientes que presentan esta enfermedad.

Así mismo, la precisión obtenida por los algoritmos implementados (RF, SVM, y GB) permiten considerar al modelo como un recurso de gran importancia en el diagnóstico y seguimiento de la EP.

Como conclusiones se destacó que la utilización de técnicas de AP permite obtener resultados de mayor precisión y eficacia en comparación con métodos tradicionales. Además, el módulo desarrollado resultó ser una herramienta de gran utilidad para la detección temprana de la EP, así como también para el seguimiento en la evolución de los pacientes a lo largo del tiempo.

En cuanto a trabajo a futuro, se prevé expandir la capacidad del modelo presentado, mejorando su rendimiento para realizar estudios de clasificación a través de nuevos sistemas de información, tales como, el análisis de datos del lenguaje natural y la supervisión de la actividad motriz, así mismo, el modelo presentado se mantendrá como una plataforma enfocada en la identificación, diagnóstico y seguimiento de enfermedades crónicas y neurodegenerativas.

Finalmente, cabe mencionar que la incorporación de tecnologías de la información y técnicas de aprendizaje profundo en el sector de la salud, representan una herramienta valiosa para la detección de enfermedades, lo que brinda apertura y expanción a futuro de mejoras que permitan agilizar los procesos de diagnóstico de diversas enfermedades neurodegenerativas.

Agradecimientos. Este trabajo de investigación fue patrocinado por el Consejo Nacional de Ciencia y Tecnología de México (CONACYT) y la Secretaría de Educación Pública (SEP) de México a través del programa PRODEP. Los autores también agradecen al Tecnológico Nacional de México (TecNM) por apoyar este proyecto.

Referencias

1. Dorsey, E. R., Bloem, B. R.: The Parkinson andemic-A call to action. JAMA Neurology, vol. 75, no. 1, pp. 9–10 (2018) doi: 10.1001/jamaneurol.2017.3299
2. Instituto Nacional de Neurología y Neurocirugía: Enfermedad de Parkinson. INNN (2022) www.innn.salud.gob.mx/interna/medica/padecimientos/parkinson.html
3. Worasawate, D., Asawaponwiput, W., Yoshimura, N., Intarapanich, A., Surangsrirat, D.: Classification of Parkinson's disease from smartphone recording data using time-frequency analysis and convolutional neural network. Technology and Health Care, vol. 31, no. 2, pp. 705–718 (2023) doi: 10.3233/THC-220386
4. Punarselvam, E.: A pragmatic approach of Parkinson disease detection using hybrid case-based reasoning neuro-fuzzy classification system over mobile edge computing. Journal of Intelligent and Fuzzy Systems, vol. 44, no. 5, pp. 7653–7668 (2023) doi: 10.3233/JIFS-220941
5. Fraiwan, L., Khnouf, R., Mashagbeh, A. R.: Parkinson's disease hand tremor detection system for mobile application. Journal of Medical Engineering and Technology, vol. 40, no. 3, pp. 127–134 (2016) doi: 10.3109/03091902.2016.1148792
6. Zhang, H., Wang, A., Li, D., Xu, W.: DeepVoice: A voiceprint-based mobile health framework for Parkinson's disease identification. In: 2018 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI), pp. 214–217 (2018) doi: 10.1109/BHI.2018.8333407
7. Kuosmanen, E., Kan, V., Visuri, A., Vega, J., Nishiyama, Y., Dey, A., Harper, S., Ferreira, D.: Mobile-based monitoring of Parkinson's disease. In: Proceedings of the 17th International Conference on Mobile and Ubiquitous Multimedia, pp. 441–448 (2018) doi: 10.1145/3282894.3289737
8. Hu, J. Yuan, D. Z., Zhao, Q. Y., Wang, X. F., Zhang, X. T., Jiang, Q. H., Luo, H. R., Li, J., Ran, J. H., Li, J. F.: Acceptability and practicability of self-management for patients with Parkinson's disease based on smartphone applications in China. BMC Medical Informatics Decision Making, vol. 20, no. 183 (2020) doi: 10.1186/s12911-020-01187-x
9. Estévez, S., Cambronero, M. E., García-Ruiz, Y., Llana-Díaz, L.: Mobile applications for people with Parkinson's disease: A systematic search in app stores and content review.

JUCS - Journal of Universal Computer Science, vol. 25, no. 7, pp. 740–763 (2019) doi: 10.3217/jucs-025-07-0740

10. Linares-del-Rey, M., Vela-Desojo, L., Cano-Cuerda, R.: Aplicaciones móviles en la enfermedad de Parkinson: Una revisión sistemática. Neurología, vol. 34, no. 1, pp. 38–54 (2019) doi: 10.1016/j.nrl.2017.03.006
11. Kotsavasiloglou, C., Kostikis, N., Hristu-Varsakelis, D., Arnaoutoglou, M.: Machine learning-based classification of simple drawing movements in Parkinson's disease. Biomedical Signal Processing and Control, vol. 31, pp. 174–180 (2017) doi: 10.1016/j.bspc.2016.08.003
12. Camps, J., Samá, A., Martín, M., Rodríguez-Martín, D., Pérez-López, C., Arostegui, J. M., Cabestany, J., Catalá, A., Alcaine, S., Mestre, B., Prats, A., Crespo-Maraver, M. C., Counihan, T. J., Browne, P., Quinlan, L. R., Laighin, G., Sweeney, D., Lewy, H., Vainstein, G., Costa, A., et al.: Deep learning for freezing of gait detection in Parkinson's disease patients in their homes using a waist-worn inertial measurement unit. Knowledge-Based Systems, vol. 139, pp. 119–131 (2018) doi: 10.1016/j.knosys.2017.10.017
13. Alkhatib, R., Diab, M. O., Corbier, C., Badaoui, M. E.: Machine learning algorithm for gait analysis and classification on early detection of Parkinson. IEEE Sensors Letters, vol. 4, no. 6, pp. 1–4 (2020) doi: 10.1109/lsens.2020.2994938
14. Rovini, E., Maremmani, C., Moschetti, A., Esposito, D., Cavallo, F.: Comparative motor pre-clinical assessment in Parkinson's disease using supervised machine learning approaches. Annals of Biomedical Engineering, vol. 46, no. 12, pp. 2057–2068 (2018) doi: 10.1007/s10439-018-2104-9
15. Moshkova, A., Samorodov, A., Voinova, N., Volkov, A., Ivanova, E., Fedotova, E.: Parkinson's disease detection by using machine learning algorithms and hand movement signal from LeapMotion sensor. In: 2020 26th Conference of Open Innovations Association, pp. 321–327 (2020) doi: 10.23919/fruct48808.2020.9087433
16. Yao, L., Brown, P., Shoaran, M.: Resting tremor detection in Parkinson's diseasewith machine learning and Kalman filtering. In: 2018 IEEE Biomedical Circuits and Systems Conference, pp. 1–4 (2018) doi: 10.1109/BIOCAS.2018.8584721
17. Anand, A., Haque, M. A., Alex, J. S., Venkatesan, N.: Evaluation of machine learning and deep learning algorithms combined with dimentionality reduction techniques for classification of Parkinson's disease. In: 2018 IEEE International Symposium on Signal Processing and Information Technology, pp. 342–347 (2018) doi: 10.1109/isspit.2018.8642776
18. Wodzinski, M., Skalski, A., Hemmerling, D., Orozco-Arroyave, J. R., Noth, E.: Deep learning approach to Parkinson's disease detection using voice recordings and convolutional neural network dedicated to image classification. In: 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 717–720 (2019) doi: 10.1109/embc.2019.8856972
19. Prashanth, R., Roy, S. D., Mandal, P. K., Ghosh, S.: High-accuracy detection of early Parkinson´s disease through multimodal features and machine learning. International Journal of Medical Informatics, vol. 90, pp. 13–21 (2016) doi: 10.1016/j.ijmedinf.2016.03.001
20. Arora, S., Venkataraman, V., Zhan, A., Donohue, S., Biglan, K., Dorsey, E., Little, M.: Detecting and monitoring the symptoms of Parkinson´s disease using smartphones: A pilot study. Parkinsonism and Related Disorders, vol. 21, no. 6, pp. 650–653 (2015) doi: 10.1016/j.parkreldis.2015.02.026
21. Ali, L., Zhu, C., Golilarz, N. A., Javeed, A., Zhou, M., Liu, Y.: Reliable Parkinson's disease detection by analyzing handwritten drawings: Construction of an unbiased cascaded learning system based on feature selection and adaptive boosting model. IEEE Access, vol. 7, pp. 116480–116489 (2019) doi: 10.1109/access.2019.2932037

22. Senturk, Z. K.: Early diagnosis of Parkinson's disease using machine learning algorithms. Medical Hypotheses, vol. 138 (2020) doi: 10.1016/j.mehy.2020.109603
23. Haq, A. U., Li, J., Memon, M. H., Khan, J., Din, S. U., Ahad, I., Sun, R., Lai, Z.: Comparative analysis of the classification performance of machine learning classifiers and deep neural network classifier for prediction of Parkinson disease. In: 2018 15th International Computer Conference on Wavelet Active Media Technology and Information Processing, pp.101–106 (2018) doi: 10.1109/iccwamtip.2018.8632613
24. Pahuja, G., Nagabhushan, T. N.: A comparative study of existing machine learning approaches for Parkinson's disease detection. IETE Journal of Research, vol. 67, no. 1, pp. 4–14 (2018) doi: 10.1080/03772063.2018.1531730
25. Nilashi, M., Ibrahim, O., Ahmadi, H., Shahmoradi, L., Farahmand, M.: A hybrid intelligent system for the prediction of Parkinson's disease progression using machine learning techniques. Biocybernetics and Biomedical Engineering, vol. 38, no. 1, pp. 1–15 (2018) doi: 10.1016/j.bbe.2017.09.002
26. Griffith, C.: What is hybrid mobile app development? (2021) ionic.io/resources/articles/what-is-hybrid-app-development/
27. Griffith, C.: Mobile app development with ionic: Cross-platform apps with ionic, angular, and cordova. O'Reilly Media (2017)
28. Grinberg, M.: Flask web development: Developing web applications with Python. O'Reilly Media (2018)
29. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in python. Journal of Machine Learning Research, vol. 12, no. 85, pp. 2825–2830 (2011)

ISSN 1870-4069

Detección de antiespacios urbanos usando YOLO: Caso de estudio Mexicali

Jorge Adrián Martínez López, Héctor De la Torre Gutiérrez,
Francisco Javier Hernández López

Centro de Investigación en Matemáticas,
Unidad Aguascalientes,
México

{adrian.martinez,hector.delatorre,fcoj23}@cimat.mx

Resumen. El estudio de los antiespacios urbanos ha traído conclusiones interesantes en el campo de la dinámica urbana, pero el número de investigaciones son escasos puesto que la identificación y localización de estos antiespacios resulta ser una tarea laboriosa y complicada. Es por lo anterior que es indispensable contar con una herramienta que facilite a los tomadores de decisiones la detección de estos antiespacios. En este trabajo, se entrenan los modelos de detección de objetos YOLOv4 y YOLOv5, usando imágenes satelitales de alta resolución, para detectar vacíos urbanos y espacios abandonados. Los resultados muestran una precisión promedio del 0.66 y 0.71 utilizando YOLOv4 y YOLOv5 respectivamente. Sin embargo, la detección de cada antiespacio por separado muestra un buen desempeño en la detección de vacíos urbanos pero un desempeño pobre en la detección de espacios abandonados, lo que abre las puertas a futuras investigaciones con el objetivo de mejorar el desempeño de estos modelos.

Palabras clave: Antiespacios urbanos, detección de objetos, deep learning, YOLOv4, YOLOv5, redes neuronales, dinámica urbana.

Urban Antispaces Detection Using YOLO: Case Study Mexicali

Abstract. The urban antispace study has brought interesting conclusions in the urban dynamics area, but the number of investigations is scarce since identifying and locating these antispaces is laborious and complex. For this reason, a tool for decision-makers is necessary to facilitate the detection of these antispaces. In this work, YOLOv4 and YOLOv5 object detection models are trained using high-resolution satellite images to detect urban empty and abandoned spaces. The results show an average precision of 0.66 and 0.71 for YOLOv4 and YOLOv5, respectively. However, the individual detection of each antispace shows a good performance in detecting urban empty spaces but a poor performance in detecting abandoned spaces, which depicts further research aimed at improving the performance of these models.

Keywords: Urban antispaces, object detection, deep learning, YOLOv4, YOLOv5, neural networks, urban dynamics.

1. Introducción

De acuerdo con [6], un antiespacio puede asociarse a espacios perdidos, indecisos o desprovistos de función dentro de un espacio urbano. Estos se pueden clasificar en tres grandes grupos (ver Figura 1):

- **Vacío urbano:** Tiene como principal característica que es un espacio sin edificar y sin ningún uso.
- **Espacio abandonado:** Son aquellas edificaciones ya sea terminadas o sin terminar que se encuentran abandonadas y que se han degradado con el paso del tiempo.
- **Remanente urbano:** Son aquellos espacios que se derivan de un proyecto urbano de gran escala y los cuales al finalizar el proyecto dejan de cumplir un propósito.

Los vacíos urbanos han mostrado tener relevancia en el estudio de la dinámica urbana, pues en ciudades fronterizas de México se ha encontrado que su distribución está asociada con la escala de la ciudad, siendo las ciudades de menor escala las que presentan una distribución más homogénea [7]. Por otro lado, las ciudades de mayor escala presentan una correlación espacial entre los vacíos urbanos, la pobreza y las condiciones de seguridad.

Esto provoca que la población con más recursos busque habitar sectores más favorables rechazando estas zonas de la ciudad, ocasionando grandes diferencias en la morfología urbana, en concreto presentando un desequilibrio entre el espacio construido y no construido, propiciando la segregación social [8].

En ciudades fronterizas, como Mexicali, se tiene una alta presencia de vacíos urbanos, los cuales se podrían explicar debido a los altos flujos migratorios e intercambio de mercancía característicos de este tipo de ciudades [2], también por el gran crecimiento de la mancha urbana debido al $boom$ económico surgido del Tratado de Libre Comercio de América del Norte (TLCAN) [5]. Además de los hallazgos que han sido descubiertos con respecto a los vacíos urbanos, de acuerdo con [6], la identificación y análisis de antiespacios urbanos puede traer los siguientes beneficios:

- Pueden constituir una valiosa fuente de información que puede dar pie a proyectos de rehabilitación y regeneración urbana, al ser áreas de oportunidad para la reconfiguración de las estructuras urbanas.
- En el ámbito económico, se puede estudiar su influencia en el valor del suelo y plusvalía de la zona.
- En el ámbito legal, permitiría profundizar en estudios para determinar los mecanismos mediante los cuales el Estado podría tomar el control de aquellos vacíos urbanos o espacios abandonados de propiedad privada, que con el pasar de los años se han quedado sin un propietario legítimo que pueda reclamar los derechos de propiedad.

Por todo lo anterior, se tiene interés en el estudio de los antiespacios, en particular los vacíos urbanos, pero debido a que la tarea de identificación no es sencilla, estos se ven limitados. Así pues, surge la necesidad de desarrollar un modelo que facilite la identificación de los antiespacios urbanos sin intervención directa humana.

(a) (b) (c)

Fig. 1. Ejemplos de los tres antiespacios, marcados dentro del polígono rojo. (a) Vacío urbano. (c) Espacio abandonado. (d) Remanente urbano.

Con el avance de las técnicas de visión computacional, teledetección y aprendizaje supervisado, se abre la interrogante de si es posible utilizar estas técnicas para detectar de manera autónoma los vacíos urbanos. Se espera que utilizando imágenes satelitales de alta resolución y un modelo de detección de objetos como You Only Look Once v4 y v5 (YOLOv4 y YOLOv5) sea posible detectar los antiespacios urbanos.

El resto del manuscrito se organiza como se menciona a continuación. En la sección 2 se exponen las principales técnicas de deep learning y los detectores de objetos más populares así como su uso en trabajos relacionados con la teledetección urbana. En la sección 3 se describe la metodología aplicada en la realización de este trabajo. Los resultados se muestran y discuten en la sección 4. Por último, en la sección 5, se presentan las conclusiones y recomendaciones para trabajo futuro.

2. Marco teórico

La teledetección urbana de objetos ha avanzado a la par del desarrollo y mejora de sensores, la fotogrametría y las técnicas de procesamiento de imágenes. Así también, la teledetección urbano de objetos se ha visto potencializada por la incorporación de técnicas de deep learning y las redes neuronales convolucionales y recurrentes (CNN y RNN, respectivamente, por sus siglas en inglés)[9].

Generalmente, los modelos de detección se clasifican en dos grandes categorías: a dos etapas y a una etapa. Los detectores a una etapa más utilizados son YOLO y Single-Shot Detector (SSD), mientras que los de dos etapas son Faster y Mask Region Based Convolutional Neural Networks (Faster R-CNN y Mask R-CNN) y Region Based Fully Convolutional Networks (R-FCN) [15].

Para este tipo de tareas, los modelos que utilizan imágenes satelitales han mostrado tener buena capacidad de detección. En [16] se utiliza un modelo de segmentación semántica para la identificación de espacios vacantes en áreas urbanas utilizando DeepLabv3 e imágenes satelitales de alta resolución.

Se entrenó el modelo dividiendo las ciudades en bloques parcialmente traslapados. De acuerdo con los resultados, el modelo en general consigue una precisión superior al 90 %. Por otro lado, muchos investigadores optan por una fusión de datos multimodal para mejorar la habilidad de detección [13].

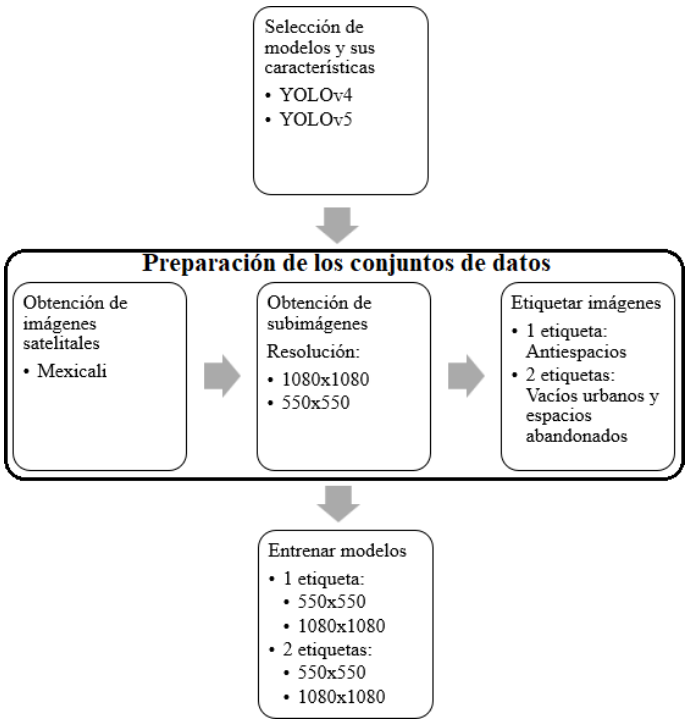


Fig. 2. Diagrama general de la metodología propuesta.

Para la segmentación y detección de áreas funcionales urbanas, se han utilizado imágenes satelitales de alta resolución, puntos sociales de interés e imágenes de alta resolución de las luces nocturnas [4]; además se ha propuesto un modelo de deep learning basado en CNN para imágenes satelitales fusionado con una red con memoria de largo y corto plazo (LSTM, por sus siglas en inglés) para datos de social sensing [1].

También, [3] propusieron un modelo para identificar asentamientos urbanos informales utilizando deep learning y datos provenientes de imágenes satelitales e imágenes de calles (Street-view). En [14] se utilizaron técnicas de deep learning para detectar edificios recientemente construidos que no han sido registrados en el catastro, se utilizaron datos de distintas fuentes con la finalidad de tener información en dos momentos en el tiempo, los cuales fueron entrenados utilizando una CNN.

De los modelos de detección de objetos YOLOv4 y YOLOv5 destacan por su desempeño y flexibilidad de implementación [10]. YOLOv4 destacó por mejorar en gran medida al modelo YOLOv3 al agregar CSPDarknet53 como backbone, Spatial Pyramid Pooling (SPP) y Path Aggregation Network (PAN) para aumentar el campo receptivo [20].

YOLOv5 de igual manera utiliza como backbone CSPDarknet53 y una capa de enfoque al inicio de la red y PAN para aumentar el campo receptivo, aunque en general su arquitectura no se considera tan innovadora comparada con la de YOLOv4, pero ofrece una gran flexibilidad para su implementación [10].

Fig. 3. Imágen satelital de la ciudad de Mexicali.

Además, YOLOv5 ofrece cinco tamaños de modelo distintos lo que permite ajustarse a las necesidades de la tarea a realizar [12], además de utilizar buenas técnicas de aumentación de datos: escalado, ajuste del espacio de color y aumentación de mosaico, este último está también presente en YOLOv4.

Comparaciones entre las distintas versiones de YOLO contra otros detectores muestran que el primero tiene un muy buen desempeño y velocidad [21], y se ha observado una ligera superioridad de YOLOv5 contra YOLOv4 para la detección autónoma de sitios de aterrizaje [17], la detección de aeronaves utilizando imágenes satelitales [11] y en el conteo de multitudes en tiempo real [19].

Pero, no se tiene una comparación entre YOLOv4 y YOLOv5 específicamente en la tarea de detección de antiespacios urbanos utilizando imágenes satelitales, por lo que es de interés investigar cuál de estos dos es mejor para esta tarea.

3. Metodología

En esta sección se explicará los pasos llevados a cabo para la realización de este trabajo. De manera resumida (ver Figura 2), el primer paso consistió en elegir las características de los modelos de YOLOv4 y YOLOv5 a utilizar. Después, con la meta de observar el efecto que tienen las resoluciones de las imágenes en el desempeño de estos modelos se prepararon dos conjuntos de datos, uno con imágenes de resolución de 1080x1080 y el otro con resolución de 550x550.

Una vez definidas las resoluciones, se obtuvieron las imágenes satelitales, las cuales fueron utilizadas para obtener las subimágenes que se etiquetaron con la localización de los antiespacios y finalmente se prepararon los conjuntos de datos para su entrenamiento con los modelos de YOLOv4 y YOLOv5.

Tabla 1. Número de imágenes destinadas al entrenamiento y validación para los dos conjuntos de datos.

Resolución	Entrenamiento	Validación	Total
550×550	780	334	1114
1080×1080	785	337	1122

3.1. Modelo detetección de objetos

YOLOv4: Para la implementación de este modelo se ajustaron los siguientes hiperparámetros: 6,000 épocas; tamaño de batch y mini-batch de 64 y 16 respectivamente; tamaño de la red de 896 para las imágenes de resolución de 1080x1080 y 544 para las imágenes de 550x550; tasa de aprendizaje de 0.001; momentum y decaimiento de pesos (weight decay) de 0.949 y 0.0005 respectivamente. Se utilizaron 9 cajas de anclaje, el tamaño de estas cajas se calcula utilizando k-medias con el objetivo de maximizar la intersección sobre la unión (IoU).

YOLOv5: Para su implementación primero es necesario elegir el tamaño del modelo, este ofrece cinco tamaños distintos del modelo: nano, pequeño, mediano, grande y extragrande; para este trabajo se seleccionó el tamaño mediano, el cual ofrece un buen desempeño además de que se adapta a los recursos de cómputo disponibles.

El entrenamiento se ajustó con 1000 épocas y un tamaño de batch de 32. Se eligió el conjunto de hiperparámetros para una alta aumentación de los datos, una tasa de aprendizaje de 0.01; momentum y decaimiento de pesos (weight decay) de 0.937 y 0.0005 respectivamente, los cuales son los predeterminados en YOLOv5.

3.2. Obtención de imágenes

Se cuenta con datos relacionados con la zona de exploración y la localización de los antiespacios urbanos en la ciudad de Mexicali, estos están en formato .shp. Estos datos derivaban de una investigación donde se analizó la distribución geoespacial de los antiespacios urbanos en ciudades fronterizas mexicanas [6].

Se utilizó el software SAS Planet versión 211230.10225 para obtener dos imágenes satelitales de alta resolución de tres canales (RGB) de la zona de Mexicali, (ver Figura 3), con resolución espacial de 0.25m/píxel y 0.5m/píxel para las imágenes de resolución 1080x1080 y 550x550 respectivamente.

Se obtuvieron las subimágenes a partir de estas dos imágenes de la zona. Partiendo de la idea presentada en [16], cada subimágen tiene un traslape con el objetivo de evitar en la medida de lo posible que un antiespacio se encuentre entre dos imágenes, permitiendo que al menos en una imágen el objeto se encuentre completo.

Una vez obtenidas las imágenes se dividieron las imágenes en un conjunto de entrenamiento y uno de validación a una razón de 70-30. Las características del conjunto de datos con las imágenes de alta resolución son:

– Resolución: 1080x1080 píxeles con tres canales (RGB).
– Resolución espacial: 0.25m/píxel.
– Traslape de las imágenes de 80 píxeles en todas las direcciones.

Tabla 2. Número de objetos (antiespacios urbanos) de cada clase de antiespacios en Mexicali.

	Vacíos	**Abandonados**	**Remanentes**
Número de objetos	6282	1524	123

Las características del conjunto de datos con imágenes de menor resolución son:

– Resolución: 550x550 píxeles con tres canales (RGB).
– Resolución espacial: 0.5m/píxel.
– Traslape de las imágenes de 50 píxeles en todas las direcciones.

El número de imágenes obtenidas, así como la cantidad destinada al conjunto de entrenamiento y validación, se muestran en la Tabla 1.

3.3. Etiquetado de imágenes

Se crearon las etiquetas en formato YOLO para cada imagen a partir de los archivos .shp con la localización de los antiespacios, se consideraron solo dos clases: vacío urbano y espacio abandonado. Se decidió omitir los remanentes urbanos puesto que el número de estos es muy pequeño (ver Tabla 2) y su localización no resulta ser de interés para tomadores de decisiones en temas ralacionados a la dinámica urbana.

Para las pruebas enfocadas en la comparación del desempeño de las distintas resoluciones y versiones de YOLO, se utilizó una sola etiqueta de antiespacios que englobaba los vacíos urbanos y los espacios abandonados. Por otro lado, se utilizaron dos etiquetas, cada una por las dos clases de antiespacios seleccionados, para comparar el desempeño que se tenía en cada clase. En la Figura 4 se muestra a manera de ejemplo como se observaría una imagen etiquetada con los dos antiespacios.

4. Resultados

4.1. Especificaciones del equipo

Los modelos se entrenaron en un equipo con las siguientes especificaciones:

– CPU: Intel Xeon Gold 5222 @3.8Ghz
– GPU: NVIDIA QUADRO RTX 8000 con 48GB GDDR6
– RAM: 48GB DDR4

Se entrenó YOLOv4 en el framework de Darknet, mientras que YOLOv5 se entrenó en el framework de PyTorch. El equipo utilizaba el sistema operativo Ubuntu 20.04.5.

4.2. Métricas de evaluación utilizadas

Se comparó el rendimiento entre ambos modelos utilizando las métricas que estos modelos arrojaban al final de su entrenamiento, las cuales son $precision$, $recall$, la puntuación $F1$, la precisión promedio (AP) y la media de la precisión promedio (mAP) con un umbral de la intersección sobre la unión (IoU) de 0.50. A continuación se detalla el cálculo de estas métricas de acuerdo con [18]. Primero es necesario definir los siguientes conceptos básicos:

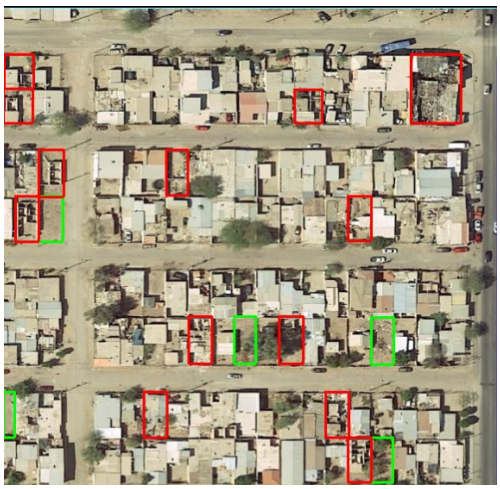


Fig. 4. Etiquetado de los antiespacios. Vacíos urbanos (color verde) y Espacios abandonados (color rojo).

– Verdadero positivo (VP): Una detección correcta de una etiqueta verdadera.
– Falso negativo (FN): Una etiqueta verdadera no detectada.
– Falso positivo (FP): Detección incorrecta de parte del modelo de un objeto no existente.

Es esencial tener un criterio para decidir si el modelo clasifica correcta o incorrectamente cada etiqueta verdadera (antiespacio) dentro de una imagen. El método más común es el IoU.

Considerando que cada etiqueta verdadera tiene una "área" o "caja" asociada B_v y el modelo realiza una predicción de esta "caja" B_p. Entonces el IoU mide el área de traslape de B_v y B_p y lo divide sobre la unión de estas dos, de la siguiente manera:

$$IoU = \frac{\text{area}(B_p \cap B_v)}{\text{area}(B_p \cup B_v)}. \tag{1}$$

Así pues, se puede elegir un umbral u, tal que si IoU $\geq u$ la predicción se considera correcta y si IoU $< u$ la predicción se considera incorrecta. Esto nos permite conocer el número total de VP, FN y FP, para calcular las métricas de precision P y recall R las cuales se definen de la siguiente manera:

$$P = \frac{VP}{VP + FP}, \tag{2}$$

$$R = \frac{VP}{TP + FN}, \tag{3}$$

donde P la capacidad del modelo de detectar únicamente etiquetas verdaderas, R la capacidad del modelo de detectar todas las etiquetas verdaderas y la puntuación $F1$ es la media armónica de las dos métricas anteriores:

Fig. 5. Detección de antiespacios: (a) Etiquetas verdaderas, en rojo espacios abandonados, en verde vacíos urbanos; (b) YOLOv4; (c) YOLOv5.

$$F_1 = 2\frac{P * R}{P + R}. \tag{4}$$

La precisión promedio (AP) se calcula por cada clase y es una aproximación del área bajo la curva de la curva $P \times R$. Esta métrica indica la capacidad de los modelos de detección de objetos de mantener altos niveles de P y R, en [18] se detallan las distintas formas de cómo calcularlo. Finalmente, la media de la precisión promedio (mAP) es la media aritmética de las AP, y se define como sigue:

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i, \tag{5}$$

Siendo N la cantidad total de clases. Como ya se mencionó, en este trabajo se utilizará un umbral del IoU de 0.5, por lo tanto, se utilizará la notación $AP@50$ y $mAP@50$.

Tabla 3. Resultados comparando YOLOv4 y YOLOv5 entrenado con los dos conjuntos de datos (550x550 y 1080x1080).

Métrica	v4 550	v4 1080	v5 550	v5 1080
P	**0.76**	0.73	0.67	0.71
R	0.61	0.67	0.64	**0.69**
F1	0.68	**0.7**	0.65	**0.7**
AP@50	0.627	0.664	0.69	**0.705**

4.3. Comparación entre YOLOv4 y YOLOv5

Los resultados de ambos modelos con los dos conjuntos de imágenes de distinta resolución se muestran en la Tabla 3. Con respecto al conjunto de datos, con imágenes de resolución de 550x550, observamos que YOLOv5 obtiene un mejor AP y R, pero pierde ante YOLOv4 en P y el puntaje $F1$. Por otro lado, en el conjunto de datos con imágenes de resolución 1080x1080, de igual forma, YOLOv5 obtiene un mejor AP y R, pero se queda un poco atrás en P, aunque ambos obtienen el mismo puntaje $F1$.

En la Figura 5 se muestra un ejemplo de los objetos detectados por los modelos de YOLOv4 y YOLOv5 entrenados con el conjunto de datos con resolución 1080x1080 píxeles. Se aprecia en este caso que la salida obtenida con YOLOv5 detecta una mayor cantidad de los antiespacios verdaderos, pero en ambos modelos se aprecian algunos falsos positivos. En cuanto a la comparación de resultados de acuerdo con la resolución de las imágenes, resulta claro que una más alta resolución conlleva mejores resultados.

En YOLOv4, una mayor resolución mejora en buena medida el AP y el R, pero se redujo P, aunque a pesar de lo anterior, se obtiene un mejor puntaje $F1$. Lo anterior indica que el modelo entrenado con imágenes de mayor resolución obtiene menos inferencias exitosas, pero a cambio se identifica una mayor cantidad de los antiespacios.

En cuanto a YOLOv5, una mayor resolución tuvo como consecuencia una mejora en todas las métricas, especialmente en P, R y el puntaje $F1$, mientras que se obtuvo una mejora menor en el AP. En cuanto a la tarea de identificación de antiespacios, es de mayor interés detectar la mayor cantidad de antiespacios verdaderos (mayor R) que detectar una menor cantidad de éstos, pero procurando una alta certeza de que los que identificó son verdaderos (mayor P).

Lo anterior sucede pues es más sencillo y rápido para una persona descartar los objetos detectados por el modelo (falsos positivos), que realizar la tarea de analizar las imágenes en la búsqueda de los antiespacios que el modelo no fue capaz de detectar (falsos negativos). Por lo anterior, el modelo de YOLOv5 entrenado con imágenes de resolución de 1080x1080 es el de mayor utilidad.

4.4. Comparación entre la precisión promedio en cada clase

Con el fin de observar el desempeño en la identificación de vacíos urbanos y espacios abandonados por separado, se entrenaron modelos en YOLOv5 con estas dos clases y con los dos conjuntos de datos de imágenes de distinta resolución. Los resultados se muestran en la Tabla 4. Se vuelve a observar una mejora significativa en todos los resultados al tener una mayor resolución en las imágenes.

Tabla 4. Resultados del desempeño de las dos clases de antiespacios utilizados: vacíos urbanos y espacios abandonados.

Resolución	mAP@50	AP Vacíos	AP Abandonados
550	0.563	0.718	0.408
1080	**0.613**	**0.741**	**0.485**

Por otro lado, observamos un buen desempeño para los vacíos urbanos, llegando a tener una AP de 0.741 en la resolución de 1080x1080; pero se tiene un muy bajo desempeño para los espacios abandonado, llegando apenas a una AP de 0.485.

El problema de identificación de espacios abandonados radica en la complejidad y heterogeneidad de estos mismos al ser observados con solo imágenes satelitales, pues muchos de estos son construcciones terminadas pero que no han sido habitadas, lo que al ser vistas de forma aérea tienen características muy similares a cualquier otra construcción que no está abandonada (véase Figura 6).

Lo anterior implica que difícilmente se podría mejorar la detección de estos antiespacios usando solamente imágenes satelitales y, que sería necesario el uso de otras fuentes de datos, así generando modelo multimodal como lo han hecho otros autores [4, 1, 3, 14] en otros contextos. A pesar de lo anterior, se ha visto un mayor interés en los vacíos urbanos [7, 8] y estos modelos muestran un buen desempeño identificándolos.

5. Conclusiones y trabajo futuro

En este trabajo se buscaba detectar de manera autónoma la ubicación de antiespacios urbanos utilizando imágenes satelitales y un modelo de detección de objetos como YOLO. Después de una exploración inicial de los datos, se observó un número muy bajo de remanentes urbanos en el conjunto de datos; por lo que, aunado al poco interés que estos pueden tener en su estudio, se decidió por no integrarlos, estudiándose solamente los vacíos urbanos y espacios abandonados.

Se utilizó YOLOv4 y YOLOv5 para precisar si es posible detectar estos dos antiespacios y, al mismo tiempo determinar cuál de estos dos modelos tiene un mejor desempeño.

Se entrenaron los modelos de YOLOv4 y YOLOv5 utilizando dos conjuntos de datos con imágenes de la ciudad fronteriza de Mexicali, uno con imágenes a una resolución de 550x550 y el otro a una resolución 1080x1080 y, englobando a los dos antiespacios en una sola clase.

Los resultados mostraron que YOLOv5 obtuvo un mejor desempeño, principalmente utilizando las imágenes de mayor resolución, logrando una precisión promedio del 0.705, contra 0.664 obtenido en YOLOv4. Utilizando al modelo de YOLOv5 como base, se entrenó con los mismos conjuntos de datos pero ahora cada antiespacio tenía su propia clase, esto con el objetivo de observar el desempeño que el modelo tendría en la detección de cada uno de los dos antiespacios.

Se logró un buen desempeño en la detección de vacíos urbanos, pero pobre desempeño con los espacios abandonados. Se concluyó que la posible causa del pobre desempeño es la gran cantidad de espacios abandonados que observados desde una perspectiva aérea tienen características no muy distintas a las de cualquier otra construcción no abandonada.

Fig. 6. Ejemplos de espacios abandonados (en rojo) donde no se presentan características que los permitan diferenciar de las demás estructuras.

Este trabajo abre las puertas a trabajos futuros. Por un lado, si bien se exploró el efecto que tiene la resolución de las imágenes en el desempeño final del modelo, es necesario seguir explorando y utilizar otras configuraciones de YOLO con la meta de determinar las mejores condiciones que aseguren un mejor desempeño en la detección de los antiespacios.

También es de interés explorar el uso de otros tipos de detectores como SDD, o el uso de modelos de segmentación semántica como deeplabv3 el cuál demostró un buen desempeño en una tarea similar [16].

Por otro lado, el uso de imágenes satelitales es suficiente para tener un buen desempeño en la detección de vacíos urbanos, no se da el mismo caso para la detección de espacios abandonados, lo que invita al uso de otros tipos de datos además de las imágenes satelitales que nos permitan mejorar ese desempeño [13].

Finalmente, es de interés el desarrollo de una herramienta que permita la identificación de antiespacios urbanos, lo que facilitará su estudio en distintos ámbitos y podrá ofrecer información valiosa en beneficio de la rehabilitación y planeación urbana.

Referencias

1. Cao, R., Tu, W., Yang, C., Li, Q., Liu, J., Zhu, J., Zhang, Q., Li, Q., Qiu, G.: Deep learning-based remote and social sensing data fusion for urban region function recognition. Journal of Photogrammetry and Remote Sensing, vol. 163, pp. 82–97 (2020) doi: 10.1016/j.isprsjprs.2020.02.014
2. Ceniceros, B., Ettinger, C.: Paisaje urbano desde la frontera Juárez-El Paso. Mapeando manifestaciones de arte urbano desde el bordo. Revista Latinoamericana de Estudios Urbanos Regionales, vol. 46, pp. 181–201 (2020) doi: 10.4067/S0250-71612020000100181
3. Chen, B., Feng, Q., Niu, B., Yan, F., Gao, B., Yang, J., Gong, J., Liu, J.: Multi-modal fusion of satellite and street-view images for urban village classification based on a dual-branch deep neural network. International Journal of Applied Earth Observation and Geoinformation, vol. 109 (2022) doi: 10.1016/j.jag.2022.102794

4. Chen, S., Zhang, H., Yang, H.: Urban functional zone recognition integrating multisource geographic data. Remote Sensing, vol. 13, pp. 4732 (2021) doi: 10.3390/rs13234732
5. Coubes, M. L.: Evolución del empleo fronterizo en los noventa: Efectos del TLCAN y de la devaluación sobre la estructura ocupacional. Frontera Norte, vol. 15, pp. 33–64 (2017)
6. Curzio, C.: Análisis sobre la distribución geoespacial del anti-espacio urbano; estudio enfocado en ciudades fronterizas del norte de México (2021)
7. Curzio, C., de la Torre, H.: Análisis geoespacial sobre la distribución de los vacíos urbanos localizados en nuevo laredo, tamaulipas. DECUMANUS, vol. 6 (2021) doi: 10.20983/decumanus.2021.1.3
8. Curzio, C., de la Torre, H.: Vacíos urbanos y desigualdad socioeconómica: Temas que convergen en la frontera norte de México. Frontera Norte, vol. 33, pp. 1–26 (2021) doi: 10.33679/rfn.v1i1.2174
9. Gong, J., Liu, C., Huang, X.: Advances in urban information extraction from high-resolution remote sensing imagery. Science China Earth Sciences, vol. 63, pp. 463–475 (2020) doi: 10.1007/s11430-019-9547-x
10. Jiang, P., Ergu, D., Liu, F., Cai, Y., Ma, B.: A review of YOLO algorithm developments. Procedia Computer Science, vol. 199, pp. 1066–1073 (2022) doi: 10.1016/j.procs.2022.01.135
11. Jindal, M., Raj, N., Saranya, P., Sundarabalan, V.: Aircraft detection from remote sensing images using YOLOv5 architecture. In: 6th International Conference on Devices, Circuits and Systems, pp. 332–336 (2022) doi: 10.1109/ICDCS54290.2022.9780777
12. Jocher, G.: YOLOv5 by Ultralytics (2020)
13. Li, J., Hong, D., Gao, L., Yao, J., Zheng, K., Zhang, B., Chanussot, J.: Deep learning in multimodal remote sensing data fusion: A comprehensive review. International Journal of Applied Earth Observation and Geoinformation, vol. 112, pp. 102926 (2022) doi: 10.1016/j.jag.2022.102926
14. Li, Q., Taubenböck, H., Shi, Y., Auer, S., Roschlaub, R., Glock, C., Kruspe, A., Zhu, X. X.: Identification of undocumented buildings in cadastral data using remote sensing: Construction period, morphology, and landscape. International Journal of Applied Earth Observation and Geoinformation, vol. 112 (2022) doi: 10.1016/j.jag.2022.102909
15. Liu, L., Ouyang, W., Wang, X., Fieguth, P., Chen, J., Liu, X., Pietikäinen, M.: Deep learning for generic object detection: A survey. International Journal of Computer Vision, vol. 128, pp. 261–318 (2020) doi: 10.1007/s11263-019-01247-4
16. Mao, L., Zheng, Z., Meng, X., Zhou, Y., Zhao, P., Yang, Z., Long, Y.: Large-scale automatic identification of urban vacant land using semantic segmentation of high-resolution remote sensing images. Landscape and Urban Planning, vol. 222, pp. 104384 (2022) doi: 10.1016/j.landurbplan.2022.104384
17. Nepal, U., Eslamiat, H.: Comparing YOLOv3, YOLOv4 and YOLOv5 for autonomous landing spot detection in faulty UAVs. Sensors, vol. 22, pp. 464 (2022) doi: 10.3390/s22020464
18. Padilla, R., Netto, S. L., da Silva, E. A. B.: A survey on performance metrics for object-detection algorithms. pp. 237–242 (2020) doi: 10.1109/IWSSIP48289.2020.9145130
19. Ranjan, A., Pathare, N., Dhavale, S., Kumar, S.: Performance analysis of YOLO algorithms for real-time crowd counting. In: 2nd Asian Conference on Innovation in Technology, pp. 1–8 (2022) doi: 10.1109/ASIANCON55314.2022.9909018
20. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection (2015) doi: 10.48550/arXiv.1506.02640
21. Sultana, F., Sufian, A., Dutta, P.: A review of object detection models based on convolutional neural network (2019) doi: 10.1007/978-981-15-4288-6_1

ISSN 1870-4069

Aprendizaje computacional aplicado en medicina convencional y alternativa para la detección temprana de enfermedades basada en análisis ocular: Revisión y propuesta de arquitectura

Jorge Ernesto González-Díaz[1], Yara Anahí Jiménez-Nieto[2], Adolfo Rodríguez Parada[2], Daniel González-Díaz[3], José Luis Sánchez-Cervantes[1]

[1] Consejo Nacional de Ciencia y Tecnología, Instituto Tecnológico de Orizaba, México

[2] Universidad Veracruzana, Facultad de Negocios y Tecnología, Campus Ixtaczoquitlan, México

[3] Universidad Tecnológica del Centro de Veracruz, México

daniel.gonzalez@utcv.edu.mx, {yjimenez, adrodriguez}@uv.mx, {D04010291, jose.sc}@orizaba.tecnm.mx

Resumen. El aprendizaje computacional estudia la construcción de sistemas capaces de aprender a partir de datos. La detección temprana de enfermedades mejora las posibilidades de tratamiento oportuno y previene el deterioro de los órganos relacionados con la enfermedad. Este trabajo se centra en la revisión del estado del arte de trabajos de investigación reportados en las áreas de la medicina convencional y la medicina alternativa que emplean técnicas de aprendizaje computacional para la detección temprana de enfermedades mediante el análisis del ojo humano, se describe una propuesta de arquitectura de un sistema automatizado para la detección de características asociadas a una condición médica.

Palabras clave: Aprendizaje computacional, enfermedades del ojo, iridología, oftalmología.

Machine Learning Applied in Conventional and Alternative Medicine for Early Detection of Diseases based on Ocular Analysis: Review and Proposed Architecture

Abstract. Machine learning studies the construction of systems capable of learning from data. Early detection of disease improves the chances of timely treatment and prevents disease-related organ deterioration. This work focuses on

the review of the state of the art of research works reported in the areas of conventional medicine and alternative medicine related to the use of computational learning techniques applied to the early detection of diseases through the analysis of the human eye, and describes a proposed architecture of an automated system for the detection of features associated with a medical condition.

Keywords: Machine learning, eye diseases, iridology, ophthalmology.

1. Introducción

El aprendizaje computacional estudia la construcción de sistemas capaces de aprender a partir de datos [1]. La aplicación de esta técnica se ha empleado en diferentes dominios, tales como el financiero, industrial y médico, por mencionar algunos. En la literatura se documentan numerosos trabajos de investigación basados en aprendizaje computacional para la detección de diversas enfermedades y condiciones médicas para un gran número de especialidades médicas [2].

En la medicina convencional, particularmente en oftalmología, el análisis de imágenes de los ojos basado en técnicas de aprendizaje computacional se presenta como un recurso útil para la detección temprana de algunas enfermedades, tales como queratocono, catarata, glaucoma, retinopatía diabética, degeneración macular, entre otros [3], alcanzando altos índices de precisión en la identificación de características asociadas a una condición médica de hasta un 99.76% [4].

La medicina alternativa, también conocida como complementaria, se refiere a un grupo de tratamientos terapéuticos y disciplinas diagnósticas que existen en su mayoría fuera de las instituciones donde se enseña y se brinda atención médica convencional [5]. En este campo, la iridología se presenta como un método clínico que consiste en el análisis del iris del ojo con la finalidad de establecer una correlación de los patrones del iris tales como: los colores, la debilidad de los tejidos, la rotura y otras características, que revelan información sobre la salud sistémica del paciente [6].

Se han reportado en la literatura algunos avances en métodos y equipos de investigación iridológica tales como: dispositivos tecnológicos para adquisición de imágenes, la adaptación de cámaras de alta resolución, aparatos de observación, además de trabajos que lograron vincular una patología específica o una condición de salud sindrómica del cuerpo humano representado con el mapa reflejo del iris de los ojos [7].

Sin embargo, algunos estudios ponen en duda la efectividad de los diagnósticos basados en iridología, por ejemplo, E. Ernst [8] en su revisión sistemática de la literatura reportó diecisiete artículos clasificados como intentos de evaluar la validez diagnóstica de la iridología y afirma que la mayoría de estas investigaciones se realizaron sin un grupo de control y algunas (con o sin un grupo de control) no fueron enmascaradas por el evaluador, es decir, no se tomaron las prevenciones necesarias para garantizar que solo el investigador a cargo del estudio conociera el tratamiento o la intervención que recibieron los participantes hasta terminado el ensayo, lo que hace las investigaciones más propensas a sesgos.

Dado el contexto anterior este artículo tiene como propósito presentar los resultados de una revisión sistemática de la literatura de los últimos cinco años (2017-2022) de artículos de investigación reportados en la medicina convencional y alternativa

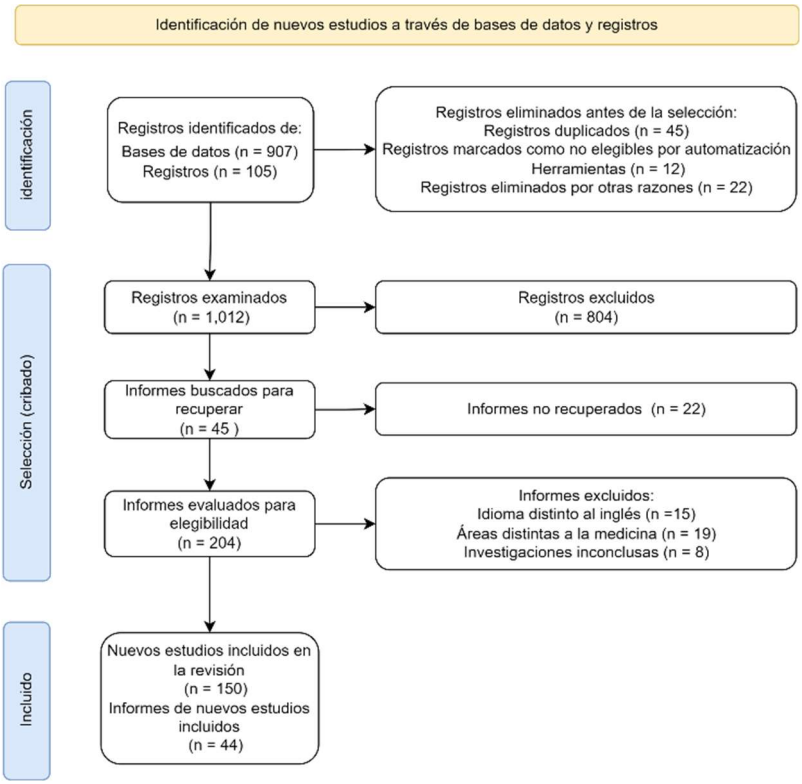


Fig. 1. Metodología de análisis.

relacionados con la detección de características relacionas con enfermedades y condiciones médicas a través del análisis del ojo humano, también se describe la propuesta de una arquitectura de un sistema automatizado para la detección de características asociadas a una condición médica y finalmente, se presentan las conclusiones obtenidas con base en el análisis realizado y los trabajos a futuro.

2. Metodología de análisis

El estudio y análisis del estado del arte, se basó en la revisión y análisis de 204 trabajos de investigación extraídos de las principales bibliotecas digitales disponibles (ACM digital library, Elsevier, IEEE, MDPI, SpringerLink, entre otras). Los principales términos de búsqueda empleados fueron: Automatic detection, disease detection systems, early detection, early diagnosis, eye detection diseases, eye diseases, eye image detection diseases, eye frameworks diagnosis, machine learning diagnosis; se incluyeron algunas combinaciones entre los términos antes mencionados con el propósito de ampliar el alcance de la búsqueda en las bibliotecas digitales.

La selección de artículos se basó en la metodología PRISMA, únicamente se eligieron trabajos publicados en los últimos cinco años (2017-2022) en revistas,

Tabla 1. Análisis comparativo de los trabajos analizados.

Año	Trabajo	CT	PCT	AMD	GL	HM	KT	DR	Técnica
2022	Abbas Q. et al. [9]	X	X	X	✓	X	X	✓	MDL
2021	A. Vyas, V. Khanduja [10]	✓	X	✓	✓	X	X	✓	DCNN
2021	A. Dash, C. Dehury [11]	✓	X	✓	X	X	X	✓	ANN
2020	Malik F. et al. [12]	X	X	✓	✓	X	X	✓	N/R
2020	Sengupta S. et al. [13]	X	X	✓	✓	X	X	✓	CNN
2019	Lopes B. et al. [14]	X	X	X	X	X	✓	X	ANN
2019	N. Bharti et al. [15]	X	X	X	✓	X	X	X	SVM

CT= Catarata; **PCT**= Catarata Pediátrica; **AMD**= Degeneración Macular; **GL**= Glaucoma; HM = Miopía Alta; **KT=** Queratocono; **DR**= Retinopatía Diabética; **ANN**= Artificial Neural Networks; **CNN** = Convolutional Neural Network; **DCNN** = Deep Convolutional Neural Network; **N/R** = No Reportado; **MDL** = Multilayer Deep Learning; **SVM**= Support Vector Machine.

Tabla 2. Condiciones médicas identificadas en la medicina convencional.

Enfermedad [16]	**Abreviatura**	**Artículos**	**Técnicas**
Catarata	CT	21	ANN, CNN, SVM
Cataratas pediátricas	PCT	10	CNN, NBC, HT, SVM
Degeneración Macular	AMD	21	DCNN, OCT, ANN
Glaucoma	GL	27	DCNN, OCT, SVM
Miopía alta	HM	13	AHE, ANN, FNN
Queratocono	KT	18	SVM, RFC, SVM
Retinopatía diabética	DR	40	DCNN, OCT, CNN

AHE = Adaptative Histogram Equalization; **ANN**= Artificial Neural Networks; **CNN** = Convolutional Neural Network; **DCNN** = Deep Convolutional Neural Network; **FNN**= Feed Forward Neural Network; **HT**= Hough Transform; **PCA**= Principal Component Analysis; **RFC**= Random Forest Classifier; **SVM**= Support Vector Machine. La lista completa de artículos[1].

memorias de congresos, capítulos de libro o libros y que además fueron escritos en inglés. Posteriormente, los trabajos se clasificaron y analizaron con base en los siguientes criterios de inclusión y exclusión:

a) Trabajos que abordan el estado del arte de la detección de enfermedades mediante el análisis del ojo humano usando técnicas de aprendizaje computacional.

b) Artículos médicos relacionados con la práctica de la iridología en la detección de condiciones médicas.

c) Enfermedades abordadas por la medicina convencional y la medicina alternativa mediante el uso de aprendizaje computacional.

[1] drive.google.com/drive/folders/1wkM1ThDO8-Z6xI_iX8ohhEWWfuG1l2D5?usp=sharing

Tabla 3. Condiciones médicas identificadas en la medicina alternativa.

Enfermedad [17]	**Abreviatura**	**Artículos**	**Técnicas**
Afecciones Hepáticas	AH	1	KNN
Afecciones Pulmonares	AP	2	SVM, FCM
Alzheimer	ALZ	4	ANN, CNN, NBC, GLCM
Colesterol	COL	8	BNN, SVM, GLCM, CHT, FLBP
Desordenes Cardiacos	DC	7	PCA, SVM, CNN, ANN, BNN
Diabetes	DB	12	FNN, FFT, RFC, GLCM, ANN
Enfermedades Renales	ER	3	CNN, BNN, PCA, AHE, DCNN
Tumores cerebrales	TC	1	SVM, FTT
Trastornos Gastrointestinales	TG	2	CNN, RFC, GLMC

AHE = Adaptative Histogram Equalization; **ANN**= Artificial Neural Networks; **BNN**= Backpropagation Neural Network; **CHT**= Circle Hough Transform; **CNN** = Convolutional Neural Network; **DCNN** = Deep Convolutional Neural Network; **FCM**= Fuzzy C-Means; **FFT**= Fast Fourier Transform; **FLBP**= Fuzzy Local Binary Pattern; **FNN**= Feed Forward Neural Network; **GLCM**= Gray Level Cooccurrence Matrix; **KNN**= K-Nearest Neighbor; **NBC**= Naive Bayes classifier.

2.1. Análisis de trabajos de investigación

Medicina convencional

La Tabla 1 muestra una comparativa de las enfermedades abordadas y técnicas de aprendizaje computacional en la medicina convencional para los trabajos similares que abordan una revisión sistemática del estado del arte. Se observa que las enfermedades más abordadas son: Glaucoma Retinopatía Diabética y Degeneración Macular, mientras que para Queratocono y Catarata se reportan uno y dos trabajos respectivamente, no se identificaron trabajos relacionados que incluyeran las condiciones médicas relacionadas con: Catarata Pediátrica y Alta Miopía.

Medicina alternativa

En la revisión de trabajos para la medicina alternativa, solo fue posible identificar un trabajo que realiza una clasificación de las condiciones médicas que pueden detectarse mediante el análisis del iris del ojo usando técnicas de aprendizaje computacional, R. Esteves et al. [7] presenta un compendio de dieciséis artículos que abordan las siguientes condiciones médicas:

Diabetes Mellitus, Insuficiencia Renal Crónica, enfermedades cardíacas y signos en el corazón, enfermedades pulmonares y signos en la región pulmonar, trastornos estomacales y gastrointestinales, enfermedades hepáticas. Sin embargo, otras condiciones médicas como: Alzheimer, Colesterol y Tumores Cerebrales no fueron incluidas en el estudio.

2.2. Clasificación de trabajos analizados

Por campo de estudio. Las Tablas 2 y 3 muestran las condiciones médicas reportadas en la medicina convencional y alternativa, respectivamente, donde se emplea el aprendizaje computacional en el análisis del ojo.

Por año. La Fig. 2 muestra el número de artículos publicados por año, se observa que las condiciones médicas más abordadas son: Retinopatía Diabética (27%),

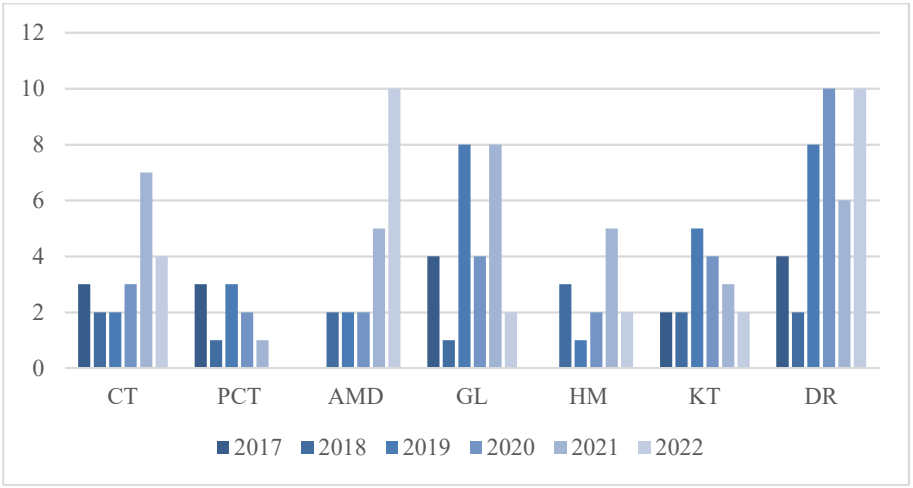


Fig. 2. Publicaciones por año en medicina convencional.

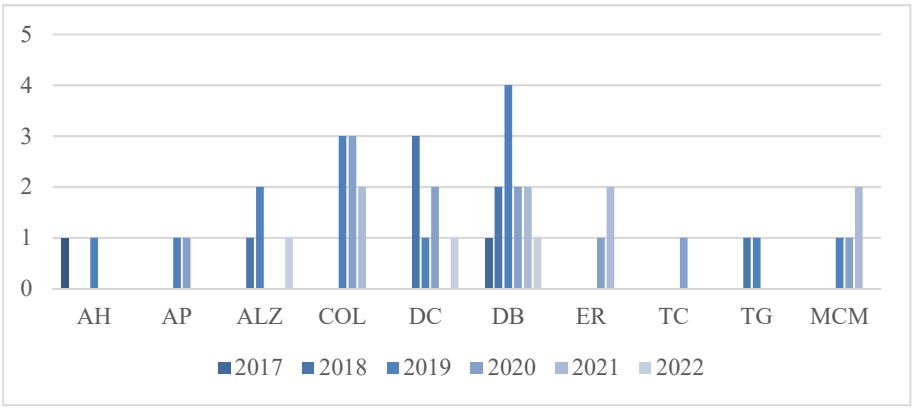


Fig. 3. Publicaciones por año en medicina alternativa.

Glaucoma (18%), y Degeneración Macular y Catarata, (14%), en conjunto suman el 73% de los artículos publicados del 2017 al 2022.

Es evidente que la tendencia va al alza, ya que en 2020 se publicaron el 18%, en 2021 el 23% y en 2022 el 20%, lo que suma el 61% del total de articulo publicados en el periodo del 2017 al 2022.

La Fig. 3 muestra la gráfica de los hallazgos relacionados con la medicina alternativa, puede observarse que la condición médica más abordada en esta área es la Diabetes reportada en el 27% de los artículos, seguida del Colesterol con 18% y los desórdenes cardiacos con el 16%, en conjunto abarcan el 61% de las condiciones médicas identificadas.

Los años con más publicaciones del periodo son 2019 32%, 2020 25%, para 2021 y lo que va del 2022 se observa un decremento en el número de publicaciones con 18% 7% respectivamente.

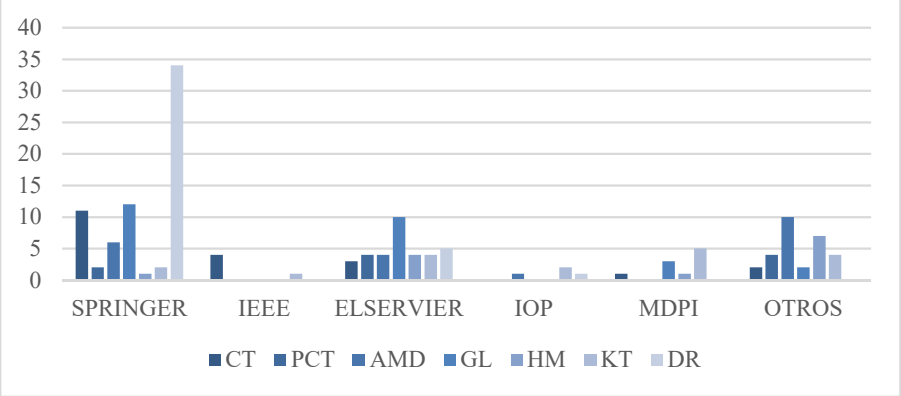


Fig. 4. Publicaciones por editorial medicina convencional.

Editorial. En la Fig. 4 se observan los trabajos publicados por editorial en el campo de la medicina convencional, se observa que la mayoría de los trabajos están concentrados en SPRINGER®, ya que esta biblioteca concentra el 45% de los artículos identificados con un total de 68, mientras que ELSEVIER® concentra el 23% de artículos con un total de 34, el 32% restante se encuentra distribuido en las demás bibliotecas de la siguiente manera IEEE® 3%, IOP® 2%, MDPI® 7%, OTROS 20%.

La Fig. 5 Muestra la distribución de artículos de la medicina alternativa en las bibliotecas digitales consultadas, el mayor número de artículos fue descargado de IEEE® con un 41%, el 25% se encontró otras bibliotecas digitales y el resto como corresponde: SPRINGER® 16%, IOP® 9%, MDPI y SCIENCEDIRECT ambas con un 5%.

Por país. En la Fig. 6 y Fig. 7 muestran la distribución por país en la medicina convencional, pude observarse que los países que encabezan la lista son India 17%, seguido de Estados Unidos y China ambos con un 16%, Reino Unido 13%, el resto de los países no supera la barrera del 10%.

En el área de la medicina alternativa, pude observarse que los países que encabezan la lista son Indonesia 43%, e India 23%, el resto de los países no supera el 10%. Es de llamar la atención que la mayoría de los trabajos publicados tienen orígenes en los continentes asiáticos y africanos, aunque también se observa una modesta participación de algunos países del continente americano.

Tras el análisis del estado del arte, se observa un uso generalizado de CNN en oftalmología, y que esta técnica se ha utilizado ampliamente para ayudar en el diagnóstico y tratamiento de diversas enfermedades oculares, incluyendo la degeneración macular relacionada con la edad, el glaucoma, la retinopatía diabética y la catarata. Estas redes neuronales se entrenan utilizando grandes conjuntos de datos de imágenes oculares y luego se utilizan para detectar patrones y características en las imágenes, lo que puede ayudar a los médicos a diagnosticar y tratar estas enfermedades de manera más efectiva.

Asimismo, se observa que los mecanismos de detección multiclase es un área poco abordada, se identificaron: 5 trabajos que realizan un tipo de detección basado en multiclase: Tres: [18, 19, 20] emplean las categorías: Temprana, Intermedia y

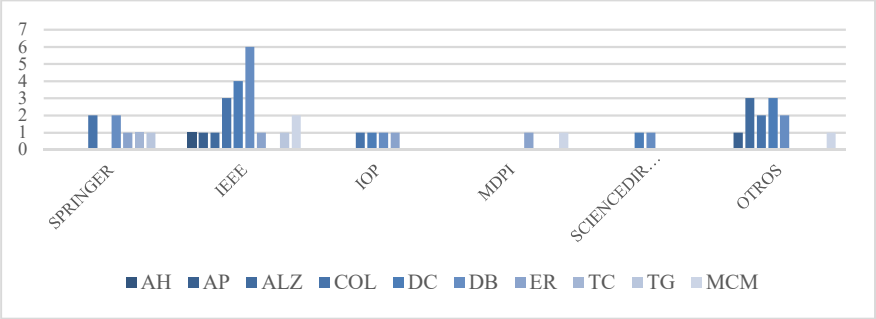


Fig. 5. Publicaciones por editorial medicina alternativa.

Avanzada, para establecer el nivel de la afección. En [21] se emplea un conjunto de 4 categorías: normal, temprano, intermedio y avanzado.

Y en la propuesta de [22] que aborda la retinopatía diabética se emplea un conjunto de 5 categorías: Sin condición médica, condición médica leve, moderada, severa y proliferativa. El resto de los trabajos solo realizan una detección binaria (si/no), esto se percibe como un área de oportunidad en el proyecto de investigación.

3. Enfoque de arquitectura

Puede ser un desafío para los médicos identificar los trastornos oculares lo suficientemente temprano utilizando imágenes de fondo de ojo. El diagnóstico manual de enfermedades oculares requiere una cantidad considerable de tiempo, es propenso a errores y en ocasiones puede ser complicado.

En este sentido, el desarrollo de un sistema automatizado de detección de enfermedades oculares con herramientas asistidas por computadora basadas en transformadores de visión para detectar diversos trastornos oculares utilizando imágenes de fondo de ojo se presenta como una solución pertinente en la detección temprana de algunas condiciones médicas.

La revisión del estado del arte demostró que tal sistema ahora es posible como consecuencia de algoritmos de aprendizaje profundo que han mejorado las capacidades de clasificación de imágenes.

Con base en lo anterior, la Fig. 8 presenta el diseño de una arquitectura propuesta para el desarrollo de un sistema automatizado para la detección temprana de características asociadas a una condición médica mediante el análisis de imágenes del fondo del ojo.

3.1. Fuente de datos

El sistema será capaz de tomar imágenes de diferentes fuentes de datos, tales como: a) Equipo oftalmológico especializado para toma de fotografías del fondo del ojo, b) Imágenes almacenadas en bases de datos o repositorios de datos validados por especialistas médicos y c) Imágenes extraídas de videos de revisiones oftalmológicas a pacientes.

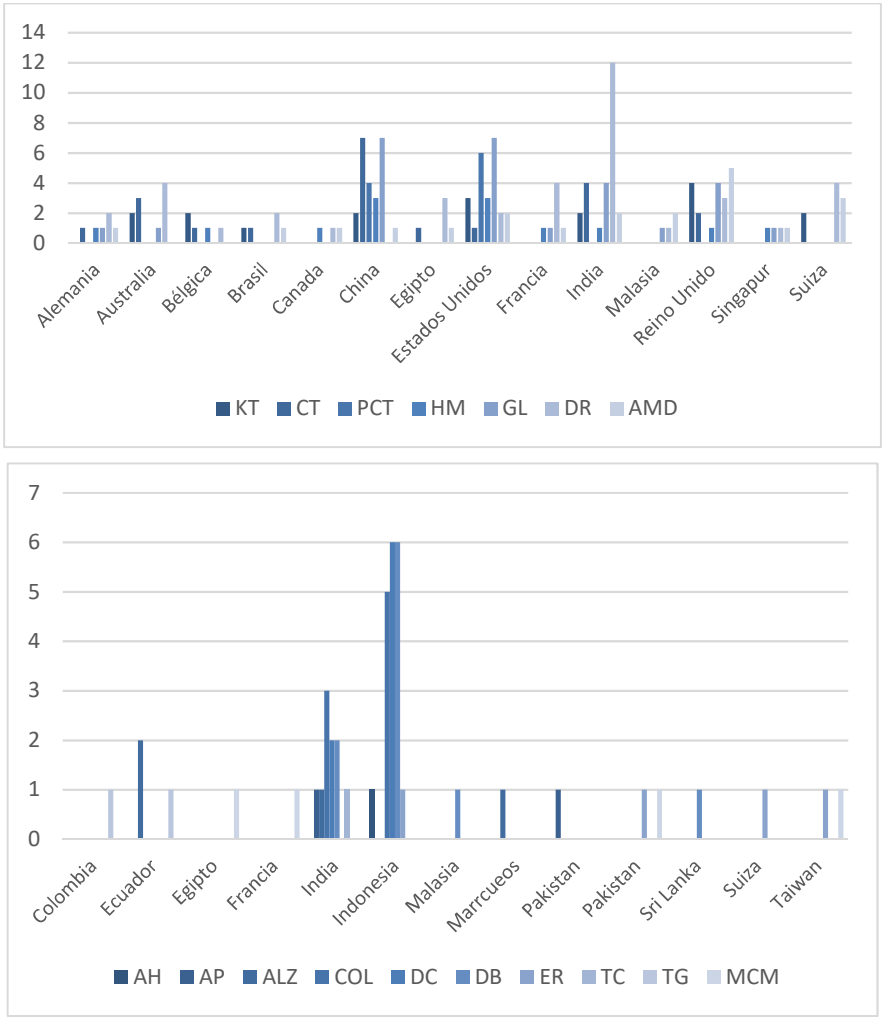


Fig. 6. artículos de la medicina convencional y alternativa por país.

3.2. Capa de presentación

Esta capa será la encargada de presentar la interfaz gráfica al usuario final, contempla cuatro componentes principales: a) Gestión de fuentes de datos: Este componente será el responsable de orquestar las conexiones con los diferentes orígenes de datos para el posterior análisis de las imágenes; b) Detección de condición médica:

Es el componente que desencadenará el proceso de análisis de la imagen y tiene comunicación directa con la capa de aprendizaje máquina; c) Presentación de resultados: Una vez culminado el proceso de análisis de la imagen, este componente recibe la información resultante de la capa de aprendizaje máquina y la muestra al usuario final, permitirá generar una copia impresa del reporte o un archivo digital en

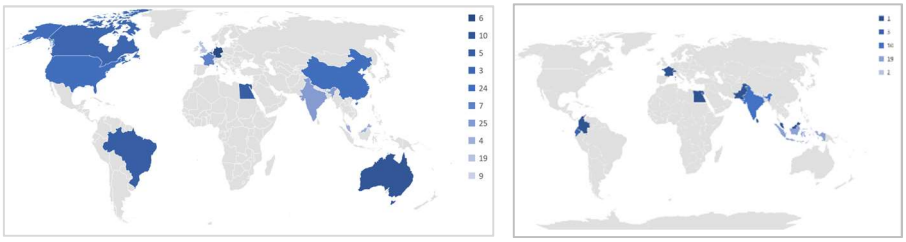


Fig. 7. Vista grafica en un mapa de la distribución de publicaciones de medicina convencional y alternativa, respectivamente por país.

formato PDF y finalmente, d) Gestión de expedientes: Se prevé que el sistema tenga la capacidad de manejar una gestión básica de expedientes con los resultados de los análisis practicados.

3.3. Capa de aprendizaje máquina

Esta capa será el núcleo de la solución propuesta, estará a cargo de la ejecución del proceso de análisis y detección de características asociadas con alguna condición médica, haciendo uso de transformadores de visión, la función de cada uno se detalla a continuación:

- **Módulo de procesamiento de imagen:** Su función será recibir la imagen de la capa de presentación y extraer de ella las características que serán clasificadas y analizadas por el módulo de aprendizaje máquina, el módulo prevé la implementación de los siguientes componentes:
  - **Procesamiento de imagen:** Será responsable de establecer las técnicas que se aplicarán sobre la imagen, en caso de ser necesario, para mejorar su calidad, tales como aumento de contraste, comprensión del rango dinámico, procesamiento de histogramas, entre otras.
  - **Segmentación de región de interés:** De acuerdo con la condición médica establecida en los parámetros de entrada, este componente realizará una identificación de la región de interés para la búsqueda de características y aplicará, de ser necesario, algunas técnicas de filtrado espacial tales como: filtros de paso alto, de Prewitt, transformaciones de intensidad, entre otros.
  - **Extracción de características:** Una vez segmentada la región de interés, este realizará un análisis de los pixeles de la imagen segmentada y generará una colección de características asociadas a la condición médica del sujeto de estudio, esta colección será la entrada para el módulo de aprendizaje máquina.
- **Módulo de aprendizaje máquina:** Este módulo será el encargado de realizar la identificación temprana de alguna condición médica en el sujeto de estudio, implementará un modelo de aprendizaje no supervisado previamente entrenado que servirá de base para el análisis y clasificación de las características extraídas por el módulo de procesamiento de imagen y con las cuales se logrará la identificación de una o varias condiciones médicas en el sujeto de estudio.

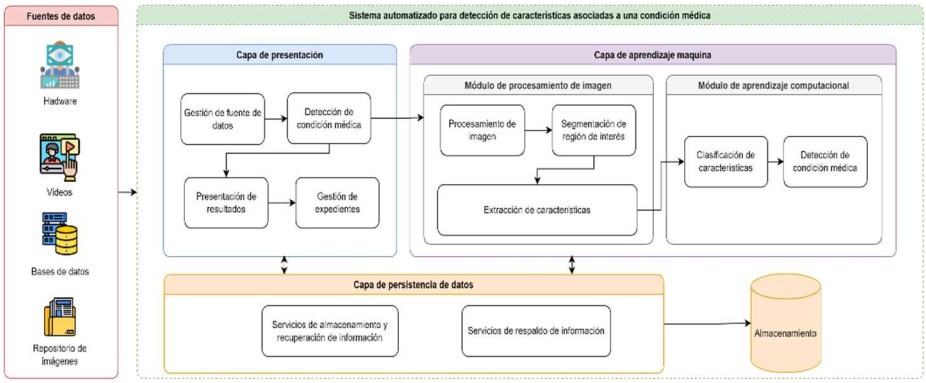


Fig. 8. Arquitectura propuesta

3.4. Capa de persistencia de datos

Esta capa implementará los servicios de recuperación de información, persistencia de datos y mecanismos de respaldo de información, a través de ella se podrán almacenar los análisis sobre los sujetos de estudio y recuperar la información de sus expedientes.

4. Discusión

Algunos trabajos refutan la validez diagnóstica de la iridología argumentando que no existen los elementos suficientes para establecer la validez de los ensayos clínicos al ser estos realizados sin un grupo de control y/o sin enmascaramiento del ensayo por parte del evaluador [8, 24, 25].

La iridología es una práctica que sostiene que se pueden diagnosticar problemas de salud y condiciones médicas a través del examen de las marcas, colores y patrones en el iris del ojo. A pesar de que cuenta con cierto seguimiento, hay varios argumentos para afirmar que la iridología no es una ciencia:

- **Falta de evidencia científica:** A diferencia de las ciencias médicas, la iridología no cuenta con estudios científicos rigurosos y ampliamente aceptados que respalden su efectividad en la detección de enfermedades o desequilibrios en el organismo.
- **Ausencia de mecanismos biológicos:** La iridología no tiene una base científica sólida en términos de mecanismos biológicos que expliquen cómo los patrones en el iris pueden estar relacionados con la salud general del individuo.
- **Falta de reproducibilidad:** Los estudios científicos se basan en la capacidad de reproducir resultados bajo condiciones controladas. La iridología no ha demostrado resultados consistentes y reproducibles en estudios controlados y ciegos.
- **Pseudociencia:** Algunos críticos argumentan que la iridología se basa en principios pseudocientíficos, ya que no sigue el método científico y no se somete a escrutinio y revisión por pares como las ciencias médicas reconocidas.

En resumen, aunque la iridología pueda tener seguidores, carece de la evidencia científica, la base biológica y la reproducibilidad necesarias para considerarse una ciencia médica legítima.

Sin embargo, en los últimos años la revisión de la literatura demuestra que la aplicación del aprendizaje computacional en la iridología como herramienta auxiliar en el diagnóstico de condiciones médicas se aborda en numerosas ocasiones sobre todo en países del continente asiático y africano [17], lo que da pie a plantear la siguiente interrogante: ¿podría el avance de la tecnología abrir las puertas a nuevas maneras de comprobar la validez diagnostica de la iridología? Será necesario realizar más investigaciones que permitan fundamentar la respuesta.

Por otro lado, en años recientes se han propuesto modelos de aprendizaje profundo basados en transformadores de visión (ViT) [23] es un modelo de aprendizaje profundo que ha demostrado ser eficaz en la clasificación de imágenes en conjuntos de datos de gran tamaño. Este modelo utiliza la arquitectura de transformador, originalmente desarrollada para el procesamiento del lenguaje natural, y la adapta para el procesamiento de imágenes.

A diferencia de los modelos de CNN más tradicionales, que procesan la imagen en bloques o parches, ViT considera la imagen completa como una secuencia de píxeles y la procesa utilizando capas de transformador. Esto permite que ViT capture relaciones globales y contextuales entre los diferentes elementos de la imagen, en lugar de centrarse únicamente en características locales.

El análisis del estado del arte realizado para este trabajo muestra que en el área de oftalmología el desarrollo de modelos computacionales basados en transformadores de visión es un área poco explorada en los últimos años, por lo que el desarrollo de un mecanismo para detección temprana de afecciones oculares haciendo uso de transformadores se presenta como una opción que empieza a cobrar relevancia.

Finalmente, en el campo de la medicina convencional, se aprecian numerosos esfuerzos de investigación basados principalmente en redes neuronales convolucionales, sin embargo, en los últimos años nuevos modelos como ViT, ha demostrado ser altamente eficaz en varios conjuntos de datos de clasificación de imágenes, incluso superando a los modelos CNN más tradicionales[26].

También se ha demostrado que ViT es eficiente en términos de uso de memoria y tiempo de entrenamiento, lo que lo hace especialmente atractivo para su uso en aplicaciones del mundo real, por lo tanto, el desarrollo de trabajos de investigación que aborden la creación de modelos computacionales bajo arquitecturas ViT contribuye al desarrollo de soluciones de asistencia médica más efectiva para su uso en la práctica clínica.

5. Conclusiones y trabajo futuro

Después del análisis de 204 artículos, se concluye que la aplicación del aprendizaje computacional en la detección temprana de enfermedades mediante el análisis ocular es un campo de estudio que ha cobrado popularidad en los últimos años en la medicina convencional y la medicina alternativa.

Sin embargo, la efectividad del uso del aprendizaje computacional en el campo de la medicina alternativa, particularmente en la iridología, ha sido puesta en duda como

método valido de diagnóstico, debido a la falta de comprobación de la validez de los ensayos clínicos reportados lo cual es un factor para tomarse en cuenta.

Por otro lado, en la medicina convencional, específicamente en el área oftalmológica se observa un gran número de esfuerzos de investigación destinados a la aplicación de técnicas basadas en el aprendizaje computacional para la detección temprana de condiciones médicas, la gran mayoría de los trabajos reportados por la literatura muestra el amplio uso de las CNN en este campo, se observa además que el uso de transformadores de visión es un área poco abordada, por lo que el desarrollo de un sistema automatizado que apoye en la detección temprana de diversas enfermedades y condiciones médicas, empleando ViT, cobra gran relevancia, por lo que los autores han decidido centrar sus esfuerzos en este campo.

Como trabajo a futuro, se revisarán las diferentes arquitecturas ViT propuestas por la literatura, se realizará una revisión sistemática de trabajos de investigación que documenten arquitecturas de modelos híbridos basados en CNN y ViT aplicados al área médica y también los dispositivos de hardware destinados a la adquisición de imágenes del ojo, poniendo especial énfasis en aquellos diseñados para el Internet de las Cosas (IoT).

Agradecimientos. Los autores de este trabajo agradecen el apoyo por parte del Consejo Nacional de Ciencia y Tecnología (CONACyT) por el soporte financiero, a la Facultad de Negocios y Tecnologías campus Ixtaczoquitlán de la Universidad Veracruzana por las facilidades prestadas para el acceso a bibliotecas digitales, a la Universidad Tecnológica de Centro de Veracruz por el tiempo prestado para el desarrollo de esta investigación y al Tecnológico Nacional de México por permitir el desarrollo de este proyecto.

Referencias

1. Alpaydın, E.: Introduction to machine learning. Fourth Edition, The Massachussetts Institute Technology (2014)
2. Sajda, P.: Machine learning for detection and diagnosis of disease. Annual Review of Biomedical Engineering, vol. 8, no. 1, pp. 537–565 (2006) doi: 10.1146/annurev.bioeng.8.061505.095802
3. Nuzzi, R., Boscia, G., Marolo, P., Ricardi, F.: The impact of artificial intelligence and deep learning in eye diseases: a review. Frontiers in Medicine, vol. 8 (2021) doi: 10.3389/fmed.2021.710329
4. Matsuba, S., Tabuchi, H., Ohsugi, H., Enno, H., Ishitobi, N., Masumoto, H., Kiuchi, Y.: Accuracy of ultra-wide-field fundus ophthalmoscopy-assisted deep learning, a machine-learning technology, for detecting age-related macular degeneration. International Ophthalmology, vol. 39, no. 6, pp. 1269–1275 (2018) doi: 10.1007/s10792-018-0940-0
5. Zollman, C., Vickers, A., Richardson, J.: ABC of complementary medicine. Wiley-BlackWell, 2nd Edition (2008)
6. Jensen, B.: Ciencia y práctica de la iridología. YIG, p. 392 (2007)
7. Esteves, R. B., Morero, J. A. P., Pereira, S. S., Mendes, K. D. S., Hegadoren, K. M., Cardoso, L.: Parameters to increase the quality of iridology studies: A scoping review. European Journal of Integrative Medicine, vol. 43, pp. 101311 (2021) doi: 10.1016/j.eujim.2021.101311
8. Ernst, E.: Iridology. Archives of Ophthalmology, vol. 118, no. 1, pp. 120 (2000) doi: 10.1001/archopht.118.1.120

9. Abbas, Q., Qureshi, I., Yan, J., Shaheed, K.: Machine learning methods for diagnosis of eye-related diseases: a systematic review study based on ophthalmic imaging modalities. Archives of Computational Methods in Engineering, vol. 29, no. 6, pp. 3861–3918 (2022) doi: 10.1007/s11831-022-09720-z
10. Vyas, A. H., Khanduja, V.: A survey on automated eye disease detection using computer vision based techniques. In: IEEE Pune Section International Conference (2021) doi: 10.11 09/punecon52575.2021.9686479
11. Dash, A., Dehury, C.: Deep learning frameworks in healthcare systems. Technical Advancements of Machine Learning in Healthcare, pp. 139–167 (2021) doi: 10.1007/978-981-33-4698-7_8
12. Malik, F. H., Batool, F., Rubab, A., Chaudhary, N. A., Khan, K. B., Qureshi, M. A.: Retinal disorder as a biomarker for detection of human diseases. In: IEEE 23rd International Multitopic Conference, pp. 1–6 (2020) doi: 10.1109/inmic50486.2020.9318059
13. Sengupta, S., Singh, A., Leopold, H. A., Gulati, T., Lakshminarayanan, V.: Ophthalmic diagnosis using deep learning with fundus images \textendash a critical review. Artificial Intelligence in Medicine, vol. 102, pp. 101758 (2020) doi: 10.1016/j.artmed.2019.101758
14. Lopes, B. T., Eliasy, A., Ambrosio, R.: Artificial intelligence in corneal diagnosis: where are we? Current Ophthalmology Reports, vol. 7, no. 3, pp. 204–211 (2019) doi: 10.1007/s40135-019-00218-9
15. Bharti, N., Gautam, G., Choudhary, K.: A review paper on eye disease detection and classification by machine learning techniques. Advances in Intelligent Systems and Computing, pp. 633–641 (2018) doi: 10.1007/978-981-13-2285-3_74
16. González-Diaz, J. E.: Lista de artículos medicina convencional (2022) drive.google.com/drive/folders/1abQ3le0n43hHoDfgh_dvwynWofmbfDoO?usp=sharing
17. González-Diaz, J. E.: Lista de artículos medicina alternativa (2022) drive.google.com/drive/folders/1wkM1ThDO8-Z6xI_iX8ohhEWWfuG1l2D5?usp=sharing
18. Fang, H., Li, F., Fu, H., Sun, X., Cao, X., Lin, F., Son, J., Kim, S., Quellec, G., Matta, S., Shankaranarayana, S. M., Chen, Y., Wang, C., Shah, N. A., Lee, C., Hsu, C., Xie, H., Lei, B., Baid, U., Innani, S., et. al.: Adam challenge: detecting age-related macular degeneration from fundus images. IEEE Transactions on Medical Imaging, vol. 41, no. 10, pp. 2828–2847 (2022) doi: 10.1109/tmi.2022.3172773
19. Perdomo, O., Andrearczyk, V., Meriaudeau, F., Müller, H., González, F. A.: Glaucoma diagnosis from eye fundus images based on deep morphometric feature estimation. In: Computational Pathology and Ophthalmic Medical Image Analysis, pp. 319–327 (2018) doi: 10.1007/978-3-030-00949-6_38
20. Yunitasari, D. A., Sigit, R., Harsono, T.: Glaucoma detection based on cup-to-disc ratio in retinal fundus image using support vector machine. In: International Electronics Symposium (2021) doi: 10.1109/ies53407.2021.9593943
21. Jadhav, A. S., Patil, P. B., Biradar, S.: Optimal feature selection-based diabetic retinopathy detection using improved rider optimization algorithm enabled with deep learning. Evolutionary Intelligence, vol. 14, no. 4, pp. 1431–1448 (2020) doi: 10.1007/s12065-020-00400-0
22. Nguyen, Q. H., Muthuraman, R., Singh, L., Sen, G., Tran, A. C., Nguyen, B. P., Chua, M.: Diabetic retinopathy detection using deep learning. In: Proceedings of the 4th International Conference on Machine Learning and Soft Computing, pp. 103–107 (2020) doi: 10.1145/3380688.3380709
23. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16×16 words: transformers for image recognition at scale (2020) doi: 10.48550/ARXIV.2010.11929
24. Ernst, E.: Iridology: a systematic review. Complementary Medicine Research, vol. 6, no. 1, pp. 7–9 (1999) doi: 10.1159/000021201

25. Stark, D. J.: Look into my eyes: iridology exposed. Medical Journal of Australia, vol. 2, no. 12, pp. 676–679 (1981) doi: 10.5694/j.1326-5377.1981.tb113049.x
26. Li, J., Chen, J., Tang, Y., Wang, C., Landman, B. A., Zhou, S. K.: Transforming medical imaging with transformers? a comparative review of key properties, current progresses, and future perspectives. Medical Image Analysis, vol. 85, pp. 102762 (2023) doi: 10.1016/ j.media.2023.102762

ISSN 1870-4069

Generador de ilustraciones para libros utilizando inteligencia artificial

Nayeli Joaquinita Meléndez-Acosta[1,2], Edmundo Bonilla-Huerta[2], José Federico Ramírez-Cruz[2], Yesenia Nohemí González-Meneses[2]

[1] Universidad del Istmo,
Campus Ixtepec,
México

[2] Tecnológico Nacional de México,
Campus Apizaco,
México

nayelim@bianni.unistmo.edu.mx, {edmundo.bh, federico.rc, yesenia.gm}@apizaco.tecnm.mx,

Resumen. En este artículo se presenta un modelo novedoso para la generación de ilustraciones para libros infantiles utilizando modelos de inteligencia artificial, se explica cada etapa del modelo propuesto, pero nos centramos en la revisión del área de investigación de los modelos existentes dentro de la generación automática de ilustraciones a partir de texto o fotografías, específicamente se realizaron pruebas utilizando DALL-E 2, una Red Adversaria Generativa (GAN) y Mini-DALL-E. Los tres modelos realizan la generación automática de ilustraciones, pero se deben considerar aspectos como la calidad de las imágenes y el tiempo de procesamiento. Una GAN requiere de mucho procesamiento, así que si se desea usar una GAN como generador de ilustraciones es necesario en nuestra propuesta pensar en otra variante que sea mucho más rápida. DALL-E 2 y Mini-DALL-E generan una gran variedad de ilustraciones con diferente calidad, estos modelos nos brindan la oportunidad de pensar en la creación de un libro con imágenes dinámicas.

Palabras clave: Generación automática de ilustraciones, red adversaria generativa (GAN), libros, ilustraciones, DALL-E 2, Mini-DALL-E.

Generator of Illustrations for Boks Using Artificial Intelligence

Abstract. In this article, a novel model for the generation of illustrations for children's books using artificial intelligence models is presented, each stage of the proposed model is explained, but we focus on reviewing the research area of existing models within the automatic generation of illustrations from text or photographs, specifically tests were performed using DALL-E 2, a Generative Adversarial Network (GAN) and Mini-DALL-E. All models made on automatic generation of illustrations, but aspects such as image quality and processing time must be considered. A GAN requires a lot of processing, so if you want to use a GAN within your model, you need to think of another variant that is much faster. DALL-E 2 and Mini-DALL-E generate a wide variety of illustrations with

different quality, these models give us the opportunity to think about creating a book with dynamic images.

Keywords: Automatic generation of illustrations, generative antagonistic network (GAN), books, illustrations, DALL-E 2, Mini-DALL-E.

1. Introducción

Dibujar ilustraciones de historietas es una tarea compleja y un proceso difícil, esta forma artística es ampliamente utilizada en varios campos, incluyendo la publicidad, el cine y la educación infantil. Los libros ilustrados para niños son una herramienta para el aprendizaje indirecto, ya que mientras miran imágenes o leen las historietas de los cuentos aprender a leer, además brindan un contenido importante que aumenta la conciencia cultural, la conciencia lingüística y la motivación [1].

Actualmente la creación de ilustraciones se basa principalmente en la implementación manual, por lo que no existen algoritmos de aprendizaje automático desarrollados para la creación de ilustraciones, además la creación manual es muy laboriosa, ya que implica una cantidad considerable de habilidades artísticas, requiere de un artista o ilustrador bien entrenado y de mucho tiempo, principalmente si se trata de un libro completo [2, 3, 4].

La Traslación de Imagen a Imagen (Image-to-Image translation I2I) basada en aprendizaje profundo ha logrado grandes resultados [4, 6]. Recientemente, los métodos de transferencia de estilo basados en aprendizaje son un parte importante y problema desafiante de la visión por computadora. I2I es un método que aprende el estilo de la imagen de referencia (imagen de estilo) y aplica el estilo a la imagen de entrada (imagen de contenido) para generar una nueva imagen que fusiona el contenido de la imagen de contenido y el estilo de la imagen de estilo.

Aunque algunos métodos existentes como las GAN han logrado resultados satisfactorios para crear ilustraciones, este modelo no necesariamente genera ilustraciones impecables [5].

Los métodos existentes suelen tener algunos problemas, entre los que se incluyen principalmente tres: (1) las imágenes generadas no tienen texturas bien definidas; (2) las imágenes generadas pierden el contenido de las imágenes originales; y (3) los parámetros de la red requieren de gran capacidad de memoria [2].

Por lo anteriormente descrito, el objetivo principal es describir un modelo para la generación automática de ilustraciones para libros infantiles. Se realiza la revisión del área de investigación de algunos modelos existentes en la generación automática de ilustraciones a partir de texto o imágenes, y también se realizar una comparativa entre modelos que generan ilustraciones de manera automática utilizando inteligencia artificial.

El resto del documento está organizado de la siguiente manera: los trabajos relacionados son descritos en la sección 2. El modelo para la generación automática de ilustraciones para libros infantiles se presenta en la sección 3. El análisis de resultados se aborda en la sección 4. Finalmente, en la sección 5 se incluyen algunas conclusiones y trabajo futuro.

2. Estado del arte

En trabajos relacionados para la generación de ilustraciones han aplicado dos enfoques principalmente, el primer enfoque es utilizar herramientas de software, pero algunas requieren un sistema de pago y la generación se realiza de manera manual.

El segundo enfoque es utilizar inteligencia artificial como procesamiento de imágenes o aprendizaje profundo, de este último específicamente las Redes Adversarias Generativas (Generative adversarial networks GAN), que permiten la generación automática a un estilo en particular, en este enfoque podemos encontrar dos categorías para la generación de ilustraciones: a partir de una imagen o a partir del lenguaje natural. Algunos trabajos basados en ambos enfoques y que son relevantes para nuestra investigación, son:

2.1. Generación de ilustraciones utilizando software

Fatimah et al. [1] realizan un estudio que tiene por objetivo conocer el uso de ToonDoo en la enseñanza de cuentos en inglés y descubrir los beneficios de esta herramienta para el desempeño de la enseñanza, además resaltan la importancia de crear comics o historietas ilustradas que benefician al estudiante, ya que el cuento puede brindar un contenido importante que aumenta la conciencia cultural, la conciencia lingüística y la motivación.

ToonDoo es un creador de cómics completamente en línea, que ofrece muchas opciones para la creación de historietas a través de una interfaz fácil de usar. Los resultados mostraron que esta herramienta al generar historias es muy útil para facilitar la imaginación de los estudiantes, promover su capacidad de hablar y producir una mejor experiencia de aprendizaje.

2.2. Generación de ilustraciones a partir de una imagen utilizando Inteligencia Artificial

Uno de los modelos generativos más utilizados y eficientes son las redes adversarias generativas (GANs), formadas por dos redes neuronales que compiten entre sí para crear una salida que se parece mucho a la entrada, así se establece un juego entre las dos redes llamadas: generadora y discriminadora, dichas redes tienen roles adversarios y cada red está representada por una función diferenciable controlada por un conjunto de parámetros.

La red generadora G intenta crear una nueva imagen basada en la imagen de entrada, mientras que la discriminadora D está entrenada para distinguir la diferencia entre imágenes reales y falsas.

El objetivo de una GAN es que la imagen creada por la red generadora se parezca tanto a la imagen real que pueda engañar al discriminador para que piense que la imagen generada es real [6]. A continuación, se muestran algunos trabajos relacionados que hace uso de una GAN:

Chen et al. [2] presentan un enfoque novedoso para transformar fotografías de escenas del mundo real en imágenes de estilo anime, para ello proponen una red

generativa adversaria ligera llamada AnimeGAN que permita lograr una transferencia de estilo rápidamente.

AnimeGAN puede ser entrenada fácilmente con dos conjuntos de datos de entrenamiento no relacionados y sus parámetros requieren una menor capacidad de memoria. Los resultados experimentales muestran que su método puede transformar rápidamente imágenes de fotos del mundo real en imágenes de estilo anime de alta calidad.

Chen et al. [7] proponen CartoonGAN, una red adversaria generativa para la estilización de ilustraciones, su método está basado en la generación de ilustraciones a partir de fotos e imágenes. Hicsonmez et al. [8] exploran las ilustraciones en libros para niños como un nuevo dominio en la traducción de imagen a imagen.

Proponen una nueva red generativa adversaria llamada GANILLA para abordar este problema y mostrar que la red resultante logra un mejor equilibrio entre estilo y contenido. También proponen una nueva métrica para la evaluación cuantitativa de la ilustración generada, donde tanto el contenido como el estilo se toman en cuenta, entonces para evaluar la preservación del contenido ellos proponen un clasificador de contenido *Content-CNN*.

Barzilay et al. [5] proponen una red adversarial generativa de estilo Multi-Ilustrador (MISS GAN), la cual tiene una arquitectura para la traducción no supervisada de imagen a ilustración, que puede generar a partir de una imagen de entrada imágenes de diferente estilo, pero conservando el contenido.

El conjunto de datos de ilustraciones está compuesto por ilustraciones de siete ilustradores diferentes, es decir contiene siete estilos diferentes. Los métodos existentes requieren entrenar a varias redes generadoras para manejar estilos diferentes, lo que limita su uso práctico. MISS GAN usa la imagen de entrada y utiliza la información de otras imágenes empleando sólo un modelo entrenado.

Shu et al. [3] proponen una arquitectura novedosa de red adversaria generativa (GAN) multiestilo, llamada MS-CartoonGAN, que puede transformar fotos en ilustraciones de múltiples estilos. Para el entrenamiento MS-CartoonGAN usa fotos e imágenes de ilustraciones de múltiples estilos.

Su arquitectura de red compartida explota las características comunes de los estilos de las ilustraciones, logrando mejores caricaturas y ser más eficiente que la caricaturización de un sólo estilo.

A través de extensos experimentos que incluye un estudio de usuario, demuestran la superioridad del método propuesto, superando a los métodos de transferencia de estilo único y de estilo múltiple de última generación. Pictonaut et al. [9] describen un método novedoso para sintetizar automáticamente tomas animadas de imágenes en movimiento.

En lugar de abordar el desafío únicamente como un problema de traducción de imágenes, se propone un enfoque híbrido que combina la estimación de la pose humana en 3D de varias personas y una GAN. Los cuadros de video se procesan con OpenPose y SMPLify-X para obtener los parámetros de la pose 3D (cuerpo, manos y expresión facial) de todos los personajes representados.

El fondo está caricaturizado con una GAN. La evaluación cualitativa muestra que el enfoque es factible, ya que en un pequeño conjunto de datos proporcionan tomas sintetizadas obtenidas de escenas de películas reales.

2.3. Generación de ilustraciones a partir de texto utilizando inteligencia artificial

Proven-Bessel et al. [10] implementan ComicGAN, un nuevo modelo para generar ilustraciones de cómics a partir de texto basado en una GAN que sintetiza cómics según los descriptores de texto.

ComicGAN tiene dos características novedosas: (1) la descripción del texto se crea a partir de etiquetas mediante permutación y aumento, y (2) la codificación de imágenes personalizadas utiliza una Red Neuronal Convolucional.

Para evaluar ComicGAN proponen dos escenarios, la generación de imágenes a partir de descripciones, y la generación de imágenes a partir de un diálogo. Por otro lado, la generación de ilustraciones a partir de descriptores proporciona cómics más claros que incluyen personajes y colores que se especifican en los descriptores.

Xu et al. [11] proponen una Red Adversaria Generativa Atencional (AttnGAN) que permite el refinamiento de varias etapas para la generación de imágenes a partir de texto. Con AttnGAN se pueden sintetizar detalles en diferentes subregiones de la imagen prestando atención a las palabras relevantes en la descripción del texto.

Reed et al. [12] desarrollan una novedosa arquitectura profunda, una red generativa adversaria para generar de manera automática imágenes realistas a partir de texto, traduciendo conceptos a personajes visuales.

En este trabajo están interesados en traducir texto en forma de descripciones escritas por humanos de una sola frase en imágenes. Por ejemplo, “este pequeño pájaro tiene un pequeño pico anaranjado y vientre blanco” o “los pétalos de las flores son rosadas”. El problema de la generación de imágenes a partir de descripciones ganó interés, pero está lejos de resolverse.

Mini Dall-E es un modelo que se puede utilizar para generar imágenes a partir de texto [13][14]. Mini DALL-E es un intento de reproducir el primer modelo para generar imágenes llamado DALL·E, pero Mini Dall-E ofrece una alternativa de código abierto, que permite a cualquier persona la capacidad de generar imágenes [14].

Bansal et al. [15] realizan un estudio para evaluar el sesgo de algunos modelos generativos de texto a imagen al introducir textos neutros ('una foto de un abogado') que carecen de cualquier señal hacia un grupo social. Consideran tres ejes de sesgo: (1) género, (2) color de la piel y (3) cultural.

Sus resultados muestran que modelos como Mini Dall-E cubren diversos grupos sociales mientras preservan la calidad de la imagen. Mini Dall-E es un transformador, una poderosa red neuronal, que se entrena a partir de una gran cantidad de datos del lenguaje natural, aprende información sobre el lenguaje en sí [13].

DALL-E 2 es un generador de imágenes sintéticas originales correspondientes a un texto de entrada creado por OpenAI [16]. Marcus et al. [17] hacen un estudio con catorce preguntas para evaluar su sentido común, razonamiento y capacidad para comprender textos complejos.

Algunas conclusiones a las que llegaron son: la calidad visual de las imágenes es muy buena, DALL-E 2 es, sin duda, extremadamente impresionante en términos de generación de imágenes, si solo se especifican dos o tres objetos, el sistema casi siempre los muestra todos, pero, por otra parte, las relaciones entre entidades son particularmente desafiantes, ya que existen fallas en DALL-E 2 al asociar correctamente propiedades especificadas con objetos como "un cubo rojo encima de un cubo azul".

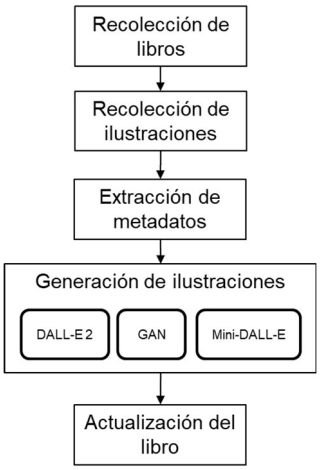


Fig. 1. Diagrama de bloques del modelo propuesto para la generación de ilustraciones para libros infantiles.

DALL-E 2 es un método zero-shot, es decir tienen la capacidad de generar imágenes basadas en descriptores de texto que no se ven durante el entrenamiento, así estos modelos pueden combinar conceptos que aprendió por separado, pero nunca vistos juntos en una imagen generada [16].

3. Modelo propuesto

La ilustración de libros es una forma de arte que se utiliza principalmente para crear imágenes y dibujos para libros. Sin embargo, el uso de ilustraciones no es sólo para embellecer los libros; son fundamentales para mejorar el contenido escrito del libro. En esta sección se muestra el diagrama general del modelo propuesto para la generación de ilustraciones.

La primera etapa es la recolección de libros en formato pdf que permiten realizar la recolección de imágenes y la generación de metadatos, los cuales son necesarios para la siguiente etapa la generación de ilustraciones usando modelos de inteligencia artificial, finalmente se debe realizar la actualización de las imágenes en el libro. El diagrama de bloques del modelo propuesto se presenta en la Fig. 1.

3.1. Recolección de libros

En esta etapa se realiza la recolección de libros ilustrados que serán utilizados para las pruebas y la recolección de ilustraciones, este tipo de libros contienen texto que se acompaña de ilustraciones que reflejan imágenes de la historia que se está narrando.

Algunos de ellos son Alicia en el País de la Maravillas, el Principito, la Vuelta al Mundo y el Mago de Oz, por ahora nos centramos en libros ilustrados de cuentos clásicos como los mencionados, la recolección se realiza de manera manual, buscando en internet los libros en formato pdf para su descarga.

Tabla 1. Metadatos de las imágenes del libro "El Principito".

#	pág	ancho	alto	formato	objeto	etiqueta
1	7	257	158	jpg	26	Boa eating a mouse cartoon
2	7	299	98	jpg	27	Boa with an elephant in the stomach cartoon
3	11	338	365	jpg	38	Little Prince cartoon
4	12	129	114	jpg	41	White lamb jumping cartoon
5	12	135	127	jpg	42	White lamb cartoon
6	12	135	116	jpg	43	White lamb cartoon
7	12	146	62	jpg	44	White box with three holes cartoon
8	15	262	357	jpg	51	A boy walked on the planet earth cartoon
⋮	⋮	⋮	⋮	⋮	⋮	⋮
40	82	278	272	jpg	222	A boy sitting on the sand cartoon
41	84	476	660	jpg	228	A boy gently falling into the sand cartoon

3.2. Recolección de ilustraciones

Los libros en pdf recolectados en la etapa anterior son usados para realizar la recolección de imágenes, para ello primero usamos un manipulador de archivos pdf llamado *PDFtf*, está herramienta descomprime el archivo y hace más fácil su lectura y manipulación.

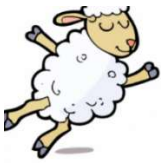
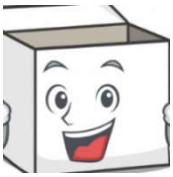
La extracción de ilustraciones se realiza de manera semiautomática haciendo uso de la herramienta *pdfimages* (parte de poopler-utils), la cual permite extraer los archivos de todas las imágenes en sus formatos originales, creando la base de datos de ilustraciones.

3.3. Extracción de metadatos

En esta etapa se realiza la construcción de una tabla con los metadatos de las imágenes, esta construcción se realiza de manera semiautomática utilizando la herramienta *pdfimages,* la cual también nos proporciona una lista de todas las imágenes del pdf con sus metadatos.

Una vez que se extraen los metadatos del pdf se crea una tabla, esta tabla está compuesta por siete metadatos: (1) número de imagen, (2) página de localización de la imagen, (3) las dos dimensiones (ancho, alto) de la imagen, (4) formato de la imagen, (5) número de objeto (localización de la imagen en el pdf) y (5) la etiqueta de la descripción de la imagen.

La Tabla 1 muestra la tabla construida con los metadatos que describen el contenido de imágenes del libro "El Principito".

Etiqueta	Little Prince cartoon	White lamb jumping cartoon	White box with three holes cartoon	A boy walked on the planet earth cartoon
Imagen Generada 1				
Imagen Generada 2				
Imagen Generada 3				

3.4. Generación de ilustraciones

La generación automática de imágenes es una tarea compleja y en la que muchos investigadores aún están trabajando. En esta etapa se realiza la revisión de tres modelos para la generación de ilustraciones, los cuales utilizan inteligencia artificial: DALL-E, una GAN y Mini-DALL-E.

Generación de imágenes utilizando DALL-E 2. DALL-E 2 puede crear imágenes realistas de una descripción en lenguaje natural, primero se realiza una consulta a los metadatos creados en la etapa anterior para obtener los parámetros necesarios para poder ejecutar DALL-E 2.

Los parámetros utilizados para generar la imagen son: las dos dimensiones (ancho, alto) de la imagen y la etiqueta de la descripción de la imagen, una vez creadas las nuevas imágenes se salvan en el banco de ilustraciones.

Generación de ilustraciones utilizando una GAN. Una GAN es un modelo de aprendizaje profundo que pretende construir una réplica $x = g(y)$ para generar las muestras deseadas de x a partir de la variable y, su enfoque es encontrar el equilibrio entre un generador y un discriminador.

Para probar la generación de ilustraciones usando una GAN, tomamos como referencia la GAN de [8], la cual ya contiene 10 modelos pre-entrenado de 10 artistas diferentes.

Finalmente, **generación de imágenes utilizando Mini-DALL-E.** Este proceso se realiza de manera semejante a como lo hicimos utilizando DALL-E 2, realizamos la consulta a los metadatos para obtener los parámetros necesarios (las dimensiones ancho y alto; y la etiqueta de descripción de la imagen).

3.5. Actualización del libro

Finalmente, en esta última etapa las ilustraciones son actualizadas por las nuevas ilustraciones generadas utilizando los modelos generadores de ilustraciones revisados en la etapa anterior.

Para realizar la actualización de la ilustración es necesario consultar la tabla de metadatos para identificar por cada imagen: el número de objeto (localización de la imagen dentro del pdf) y las dimensiones (ancho, alto) de la imagen.

La identificación del número de objeto en el pdf que representa a la imagen es muy importante para localizar la posición de la imagen dentro del archivo pdf. Así se genera una nueva versión del libro que contiene imágenes generadas ya sea por DALL-E 2, la GAN o Mini-DALL-E. Resumiendo, después de revisar cada una de las etapas de la generación de ilustraciones a continuación se presenta el algoritmo de todo el proceso:

Algoritmo 1. Generación de ilustraciones

#	Algoritmo
1	Recolección de libros en internet (archivos en PDF)
2	Descomprimir el archivo PDF usando la herramienta *PDFtf*
3	Extraer las ilustraciones originales del libro usando la herramienta *pdfimages*
4	Extraer los metadatos de las ilustraciones contenidas en el libro usando la herramienta *pdfimages*
5	Crear la tabla de metadatos usando los datos del paso anterior
6	Generar las imágenes usando un modelo de inteligencia artificial
7	Abrir el archivo PDF descomprimido a actualizar
8	Identificar la imagen a actualizar a través del número de objeto guardado en la tabla de metadatos
9	Abrir el archivo PDF y actualizar la ilustración

4. Análisis de resultados

En esta sección se muestra nuestros experimentos preliminares sobre la generación automática de ilustraciones, la implementación se realiza en Python y los modelos pre entrenados de la GAN usan PyTorch.

Para realizar los experimentos se ha utilizado el libro “El principito”. Las imágenes generadas son 256 x 256 píxeles, así que se debe realizar un redimensionamiento al tamaño de la imagen original. A continuación, se muestran algunos resultados de la generación de ilustraciones utilizando cada uno de los modelos generadores que son revisados en este trabajo.

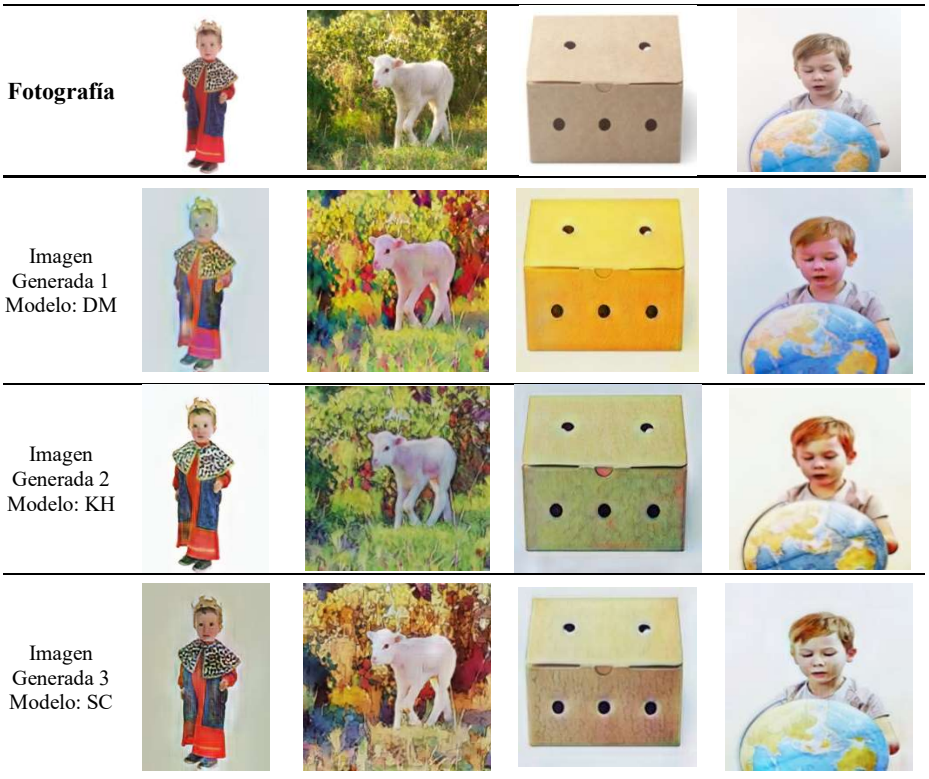


Fig. 3. Ejemplos de imágenes generadas por la GAN utilizada en [8] usando tres modelos pre-entrenados: DM, KH y SC.

Para mostrar la generación automática de ilustraciones se han seleccionado las imágenes 3, 5, 6, y 7 de la tabla de metadatos (Tabla 1).

La Fig. 2 muestra los resultados de generar tres ilustraciones por cada imagen utilizando DALL-E-2, que hace uso de la etiqueta de descripción de cada imagen, cabe mencionar que las etiquetas están en inglés debido a que la generación de imágenes es mucho mejor, los resultados muestran que la calidad visual de las imágenes generadas es muy buena, ya que los objetos que componen a la ilustración están bien definidos [18].

La generación de ilustraciones usando algunos de los modelos en [8] que utilizan una GAN se muestran en la Fig. 3. Estos modelos no necesitan imágenes relacionadas, sino solo dos conjuntos de imágenes diferentes, uno para el dominio origen y otro para el dominio destino.

Así esta GAN utiliza fotografías como dominio de origen e ilustraciones como dominio de destino. Para probar estos modelos se han utilizado fotografías que representan a las imágenes del libro ilustrado, debido a que las imágenes del libro ya son ilustraciones, se han seleccionado los tres modelos: DM (David MCKee), KH (Kevin Henkes) y SC (Stephen Cartwright) que a nuestro criterio muestran los mejores resultados y podemos ver que algunas imágenes generadas tienen algunos pequeños defectos o detalles en la generación de bordes [5, 7], incluso a veces los resultados

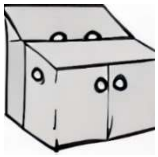
Etiqueta	Little Prince cartoon	White lamb jumping cartoon	White box with three holes cartoon	A boy walked on the planet earth
Imagen Generada 1				
Imagen Generada 2				
Imagen Generada 3				

Fig. 4. Ejemplo de imágenes generadas por Mini-DALL-E.

aparentan sólo cambiar la tonalidad de los colores y la textura, y hay ocasiones en las que se observa que el estilo ha cambiado con éxito.

Continuando con el último modelo generador de ilustraciones, la Fig. 4 muestra las ilustraciones generadas por Mini-DALL-E, podemos ver que las ilustraciones generadas son de buena calidad [15], pero a veces los bordes de los elementos que componen a la ilustración no están bien definidos, principalmente en las manos, los ojos, la textura del cabello y el rostro, pero en particular los rostros no son simétricos y las expresiones no son naturales. Concluyendo que las caras y las personas no se generadas correctamente [14].

Los tres modelos revisados en este trabajo realizan la generación automática de ilustraciones, pero hay ventajas y desventajas que se deben considerar, de esta forma DALL-E-2 genera ilustraciones con elementos bien definidos, pero la etiqueta debe ser lo más detallada posible.

Por otro lado, la GAN genera la ilustración; en ocasiones parece que no realiza ningún cambio, solo modifica colores y texturas; los resultados de las ilustraciones se obtienen en un tiempo aproximado de 1 a 2 minutos, dependiendo del número de imágenes. Finalmente, los elementos de las ilustraciones generadas por Mini-DALL-E no están bien definidos y pueden presentar ilustraciones con elementos borrosos. También es importante mencionar que la generación de imágenes es mucho mejor en el caso de DALL-E -2 y Mini-DALL-E con las etiquetas en inglés [14].

Puede ser difícil decidir qué modelo podría ser el más adecuado. La Tabla 2 proporciona una comparación de los modelos mencionados anteriormente en términos del método utilizado, el tamaño del base de datos, las capacidades del modelo y su

Tabla 2. Resumen de los modelos generativos

Modelo	Método	Tamaño de la base de datos	Código abierto	Capacidades
DALL-E 2	Zero-Shot	650M	Parcialmente	Texto a Imagen
GAN	No supervisado	Varias	Parcialmente	Imagen a Imagen
Mini DALL-E	Transformador	3M, 12M y 15M	SI	Texto a Imagen

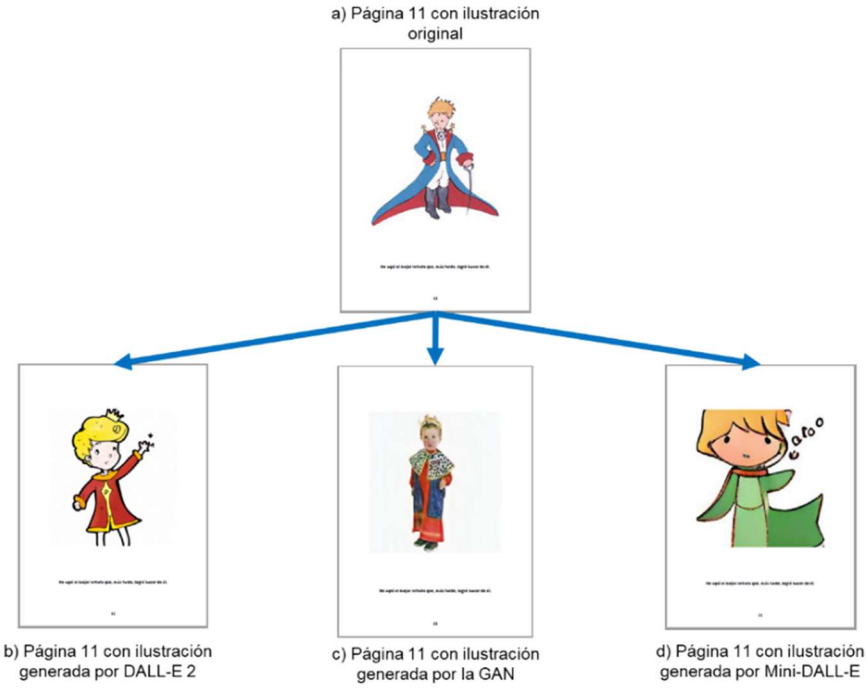


Fig. 5. Actualización de una ilustración en la página 11 del libro “El Principito”.

accesibilidad, Además, se puede acceder a estas herramientas a través de Python, pero DALL-E 2 sólo permite generar un número limitado de imágenes.

Finalmente, después de la generación de las ilustraciones se realiza la actualización de ilustraciones en el libro, generando una nueva versión del libro. La actualización de ilustraciones se realiza a través de la identificación del objeto del número de objeto de la imagen a actualizar, luego se reemplaza la imagen con la nueva imagen generada.

La Fig. 5 muestra la actualización de una ilustración por una ilustración generada por DALL-E -2 (ver Fig. 5a), la GAN (ver Fig. 5b) y Mini-DALL-E (ver Fig. 5d) de la página 11 del libro “El Principito”. La Fig. 6 muestra la actualización de las ilustraciones de la página 12 del libro “El Principito”, se han considerado una ilustración generada por DALL-E -2, dos generadas por la GAN y una generada por Mini-DALL-E.

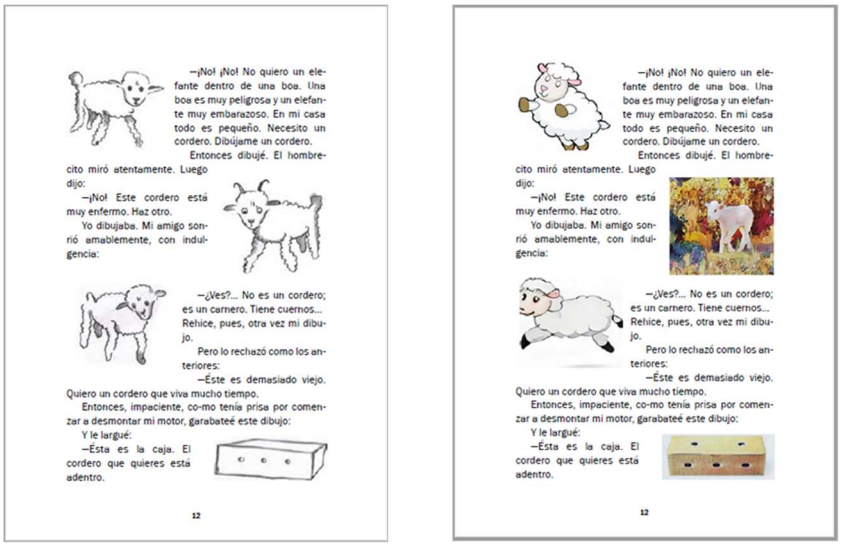


Fig. 6. Actualización de las ilustraciones en la página 12 del libro “El Principito”.

Es importante mencionar que es muy conveniente tener un banco de imágenes para generar un libro dinámico en el que de manera aleatoria cada vez que un niño lea un cuento pueda observar una gran variedad de ilustraciones diferentes.

5. Conclusiones y trabajo a futuro

La generación automática de ilustraciones es una tarea muy difícil, que se puede realizar a través del lenguaje natural o utilizando una imagen de referencia, por lo que hemos revisado algunos modelos existentes dentro de la generación automática de ilustraciones, para poder seleccionar un modelo que mejor se adapte como generador de ilustraciones para libros.

Los resultados experimentales nos indican que utilizar una GAN requiere de mucho procesamiento, ya que tarda demasiado en el proceso de entrenamiento, de esta forma si se pretende utilizar una GAN es necesario pensar en otra variante, realizando modificaciones a la arquitectura para que sea mucho más rápida o explorar la generación automática de ilustraciones utilizando sólo procesamiento digital de imágenes.

Por otro lado, la revisión de estos modelos nos hace considera la generación automática de ilustraciones utilizando lenguaje natural, ya que DALL-E 2 o Mini-DALL-E pueden generar una gran cantidad y variedad de ilustraciones, así estos modelos nos brindan la oportunidad de pensar en la creación de un libro con imágenes dinámicas, es decir, el usuario podría elegir que ilustración ver de un conjunto en un libro determinado utilizando realidad aumentada.

Es importante mencionar que es pertinente investigar la motivación de lectura de los niños y realizar muestreos en escuelas con el apoyo de pedagogos y psicólogos para ver las reacciones de los niños al leer y ver un libro con ilustraciones dinámicas. Incluso también se está considerando permitir que el niño haga sus propios dibujos y usando estas herramientas generar la ilustración.

Además, también como trabajo futuro se pretenden realizar pruebas utilizando otro tipo de libros a ilustrar, por ejemplo, considerar en generar ilustraciones a color para libros con ilustraciones en blanco y negro. Por último, se piensa en la implementación de una función discriminadora que permita seleccionar la mejor imagen generada por alguno de los modelos generadores automáticos de ilustraciones revisados en este trabajo.

Referencias

1. Fatimah, A. S., Santiana, S., Saputra, Y.: Digital comic: An innovation of using toondoo as media technology for teaching english short story. English Review: Journal of English Education, vol. 7, no. 2, pp. 101–108 (2019) doi: 10.25134/erjee.v7i2.1526
2. Chen, J., Liu, G., Chen, X.: AnimeGAN: A novel lightweight GAN for photo animation. Communications in Computer and Information Science, vol. 1205, pp. 242–256 (2020) doi: 10.1007/978-981-15-5577-0_18
3. Shu, Y., Yi, R., Xia, M., Ye, Z., Zhao, W., Chen, Y., Lai, Y., Liu, Y.: GAN-based multi-style photo cartoonization. IEEE Transactions on Visualization and Computer Graphics, vol. 28, no. 10, pp. 3376–3390 (2022) doi: 10.1109/tvcg.2021.3067201
4. Yi, Z., Zhang, H., Tan, P., Gong, M.: DualGAN: Unsupervised dual learning for image-to-image translation. In: IEEE International Conference on Computer Vision, pp. 2868–2876 (2017) doi: 10.1109/iccv.2017.310
5. Barzilay, N., Shalev, T. B., Giryes, R.: MISS GAN: A multi-illustrator style generative adversarial network for image to illustration translation. Pattern Recognition Letters, vol. 151, pp. 140–147 (2021) doi: 10.1016/j.patrec.2021.08.006
6. Pang, Y., Lin, J., Qin, T., Chen, Z.: Image-to-image translation: Methods and applications. IEEE Transactions on Multimedia, vol. 24, pp. 3859–3881 (2022) doi: 10.1109/tmm.2021.3109419
7. Chen, Y., Lai, Y., Liu, Y.: CartoonGAN: Generative adversarial networks for photo cartoonization. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 9465–9474 (2018) doi: 10.1109/cvpr.2018.00986
8. Hicsonmez, S., Samet, N., Akbas, E., Duygulu, P.: GANILLA: Generative adversarial networks for image to illustration translation. Image and Vision Computing, vol. 95 (2020) doi: 10.1016/j.imavis.2020.103886
9. Tous, R.: Pictonaut: Movie cartoonization using 3D human pose estimation and GANs. Multimedia Tools and Applications, vol. 82, no. 14, pp. 21101–21115 (2023) doi: 10.1007/s11042-023-14556-1
10. Proven-Bessel, B., Zhao, Z., Chen, L.: ComicGAN: Text-to-comic generative adversarial network (2021) doi: 10.48550/ARXIV.2109.09120
11. Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X., He, X.: AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1316–1324 (2018) doi: 10.1109/cvpr.2018.00143

12. Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., Lee, H.: Generative adversarial text to image synthesis. In: Proceedings of The 33rd International Conference on Machine Learning, vol. 48, pp. 1060–1069 (2016) doi: 10.48550/arXiv.1605.05396
13. Zhang, M., Zhang, Z. L. X.: Transformers, Dall-E mini, and next level text to video. pp. 1–8 (2023) https://www.researchgate.net/publication/369299281
14. Dayma, B., Pcuenca, Saifullah, K., Ghosh, R., Abraham, T., Abid, A., Patil, S., Khắc, P. L.: Borisdayma/dalle-mini: Initial release (2021) doi: 10.5281/ZENODO.5146400
15. Bansal, H., Yin, D., Monajatipoor, M., Chang, K.: How well can text-to-image generative models understand ethical natural language interventions? In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pp. 1358–1370 (2022) doi: 10.48550/arXiv.2210.15230
16. Maerten, A., Soydaner, D.: From paintbrush to pixel: A review of deep neural networks in AI-generated art (2023) doi: 10.48550/arXiv.2302.10913
17. Marcus, G., Davis, E., Aaronson, S.: A very preliminary analysis of DALL-E 2 (2022) doi: 10.48550/arXiv.2204.13807

ISSN 1870-4069

Validación de la escala de valencia para repositorios de imágenes emocionales en poblaciones de adultos jóvenes mexicanos

Derick A. Lagunes-Ramírez[1], Gabriel González-Serna[1], Nimrod González-Franco[1], Dante Mújica-Vargas[1], María-Yasmín Hernández-Pérez[1], José-Alejandro Reyes-Ortiz[2], Leonor Rivera-Rivera[3]

[1] Tecnológico Nacional de México,
Centro Nacional de Investigación y Desarrollo Tecnológico,
Departamento de Ciencias Computacionales,
México

[2] Universidad Autónoma Metropolitana,
Departamento de Sistemas,
Azcapotzalco,
México

[3] Instituto Nacional de Salud Pública,
Centro de Investigación en Salud Poblacional,
México

{d18ce078, gabriel.gs, nimrod.gf, dante.mv,
yasmin.hp}@cenidet.tecnm.mx, jaro@azc.uam.mx, lrivera@insp.mx

Resumen. En varios campos de la ciencia se busca una línea base emocional como referencia para comparar patrones de comportamiento o recepción de sentimientos en las investigaciones. En estos casos se hace uso de repositorios de contenido emocional, los cuales han sido validados y liberados para su uso. Sin embargo, la recepción de emociones es subjetiva. En este estudio, analizamos la evaluación de imágenes de dos repositorios de imágenes emocionales (IAPS y OASIS). Para ello, realizamos una comparación de la evaluación de valencia para una muestra de las imágenes positivas, negativas y neutrales de los repositorios en una población de adultos jóvenes mexicanos. La comparación de los resultados demuestra similitudes en los estudios. Sin embargo, se observaron sesgos en cuanto a la recepción de algunos estímulos negativos y neutrales. Se concluye que la valencia de la mayoría de los estímulos se alinea con los estudios originales al mismo tiempo que los sesgos observados pueden deberse a las diferencias del contexto temporal, de edad y cultural entre los estudios comparados. Finalmente, se recomienda considerar todos estos factores en la selección de imágenes y repositorios para futuros trabajos.

Palabras clave: Aplicaciones, computo afectivo, repositorios emocionales.

Valence Scale Validation for Emotional Image Repositories in Mexican Young Adult Populations

Abstract. In various science fields, an emotional baseline is sought as a reference to compare behavioral or reception patterns of feelings in applied research. In those cases, emotional content repositories are used, which have been validated

and released for use. However, the reception and interpretation of emotions is subjective. In this study, we analyzed the evaluation of images from two emotional image repositories (IAPS and OASIS). To do this, we performed a comparison of the valence evaluation for a stimuli sample of positive, negative, and neutral images from the repositories in a population of young Mexican adults. The comparison results demonstrate similarities in the studies. However, biases were observed regarding the reception of some negative and neutral stimuli. It is concluded that the valence of most of the stimuli is in line with the original studies while the observed biases may be due to differences in the temporal, age and cultural context between the compared studies. Finally, it is recommended to consider all these factors in the selection of images and repositories for future work.

Keywords: Applications, affective computing, emotional repositories.

1. Introducción

Un repositorio se refiere a colecciones administradas de datos o conjuntos de datos. Los datos que componen al repositorio generalmente comparten una cierta temática y consistencia, ya que el objetivo de los repositorios es presentar la información para su explotación en investigaciones o sistemas. En el contexto de la investigación del cómputo afectivo y la interacción humano-computadora (HCI), los repositorios de multimedia emocional (vídeos, sonidos e imágenes) son utilizados como base para experimentos de inducción emocional y comportamiento humano.

Cada uno de los elementos de los repositorios emocionales, al cual le llamamos estímulo, produce una reacción en un sujeto de pruebas. Para evaluar y clasificar la reacción del estímulo, normalmente se utiliza el modelo circunflejo[1], donde las emociones se distribuyen en dos dimensiones: 1) valencia y 2) excitación (también llamado activación).

La valencia, en un estado positivo, se refiere al nivel de atractivo o bondad de un evento, objeto o situación, mientras que en un nivel negativo se refiere a la aversión o maldad del mismo[2]. Por otro lado, la excitación, representa el nivel de impacto causado por una emoción en una persona. Un valor positivo se puede definir como la intensidad con la que se experimenta una emoción como el placer[3], sin embargo, la excitación positiva también se puede ligar con un alto nivel de estrés.

El nivel de valencia puede ir de negativo (experiencia de calma) a positivo (experiencia de estrés o felicidad) [4]. El modelo circunflejo se puede representar en un plano bidimensional, donde la excitación representa el eje vertical y la valencia del eje horizontal. Con este modelo se ubicaría una emoción completamente neutra en el punto de origen del plano bidimensional.

Esta clasificación emocional permite a los investigadores clasificar y elegir estímulos con niveles emocionales adecuados (normalmente divididos en negativos, positivos y neutros) para los experimentos. Además, existe otra dimensión al evaluar una emoción, esta es llamada “dominio” y es una medida que representa que tanto el control tiene una persona sobre el sentimiento que le provoca un estímulo[5].

El objetivo de los repositorios emocionales es estandarizar las experiencias y respuestas emocionales, esto al formar un marco de referencia (validado) con el cual se puede comparar y contrastar el comportamiento emocional. Gracias a los repositorios,

se puede crear una línea base que se puede utilizar universalmente. El uso de los mismos datos permite que diferentes investigadores prueben, repliquen y comparen resultados, lo cual ayuda a desarrollar una comprensión más sólida del tema de investigación.

Aunque los repositorios de estímulos emocionales estén relacionados entre ellos, cada publicación de un repositorio se realiza con un objetivo particular en mente, por ejemplo: La inducción emocional en niños o personas con enfermedades mentales y traumas o la clasificación de estados emocionales con inteligencia artificial.

Además, que también se consideran para un contexto específico, tomando en cuenta diferencias culturales y grupos de personas. Algunos de los conjuntos de imágenes emocionales encontrados en la literatura son:

- El conjunto de expresiones faciales NimStim [6].
- Taiwanese Female Expression Image Database (TFEID)[7].
- Japanese Female Facial Expression (JAFFE) [8].
- Pictures of Facial Affect (PFA) [9].
- Karolinska Directed Emotional Faces (KDEF) [10].
- Chinese Facial Affective Picture System (CFAPS) [11].
- El conjunto NIMH con rostros emocionales de niños (NIMH-ChEFS) [12].
- Tsinghua Psychological Image System (ThuPIS) [13].
- Radboud Faces Database (RaFD) [14].
- International Affective Picture System (IAPS) [5].
- Open Affective Standardized Image Set (OASIS) [15].

Para este trabajo de evaluación de valencia, solo se consideran dos de los repositorios de imágenes emocionales más utilizados: 1) IAPS y 2) OASIS. Se reportan los resultados de la evaluación para los dos repositorios de imágenes emocionales, la comparación entre estudios y se discutirá sobre el uso de ambos para experimentos de cómputo afectivo, IHC y IA.

1.1 IAPS

Para la creación de este repositorio, en el año 2008, Lang et al. recopilaron la evaluación de 3 escalas emocionales de un conjunto de imágenes: 1) valencia, 2) excitación y 3) dominio. Los datos fueron obtenidos con la escala de autoinforme Self-Assessment Manikin (SAM) [2] (ver Figura 1).

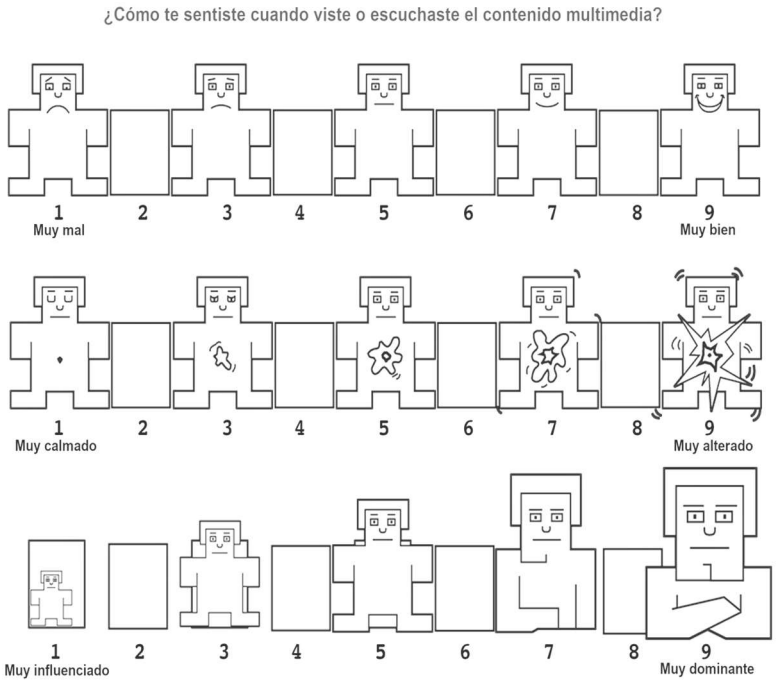


Fig. 1. Adaptación al español del instrumento de autoinforme Self-Assessment Manikin (SAM), [2].

La escala SAM, muestra con figuras, una escala tipo Likert que representan las 3 dimensiones emocionales. De esta manera:

- La escala SAM para valencia muestra un personaje con el ceño fruncido para la respuesta más negativa y un personaje sonriente para la respuesta más positiva.
- El nivel de excitación está representado por un impacto o explosión cada vez más grande en el pecho.
- La escala de dominio muestra un personaje que crece en tamaño (y presenta una "mirada cada vez más agresiva en sus cejas y brazos").

Las imágenes de IAPS se mostraron a los participantes durante 26 segundos cada una, en conjuntos de 12, hasta un total de 60 imágenes en cada sesión. Actualmente, el repositorio cuenta con un poco más de 1000 imágenes catalogadas en función de la media y desviación estándar para cada evaluación, además de una clasificación por separado entre hombres y mujeres.

Originalmente, este repositorio de imágenes fue probado en países de habla inglesa, pasando por una adaptación al contexto español en España en 2013 [16] y posteriormente fue validado en México en 2015 [17].

IAPS, incluye imágenes a color que representan una amplia gama de eventos en la experiencia humana: personas, casas, objetos del hogar, funerales, contaminación,

Fig. 2. Algunas de las imágenes emocionales de OASIS.

suciedad, paisajes, eventos deportivos, guerras, desastres naturales, tratamientos médicos, enfermedades, animales, insectos, eventos familiares, niños, etc.

Sin embargo, las imágenes no pueden ser visualizadas públicamente, para el acceso a ellas, solo los investigadores académicos pueden solicitarlas para "ser utilizadas en proyectos de investigación básicos y de salud". El repositorio IAPS se utiliza en el laboratorio de Sistemas Híbridos Inteligentes del TecNM/CENIDET para diferentes pruebas e investigaciones relacionadas con el cómputo afectivo.

1.2 OASIS

El conjunto abierto y estandarizado de imágenes afectivas (OASIS) cuenta con más de 900 estímulos clasificados en cuatro clases: 1) Personas, 2) animales, 3) objetos y 4) escenarios. En el caso de este estudio, las imágenes se recopilaron de fuentes en línea y las evaluaciones de valencia/excitación se obtuvieron en un estudio en línea donde se pidió a los participantes que indicasen el nivel emocional intrínseco a cada imagen con una escala Likert de 7 puntos.

Los resultados del estudio cubrieron gran parte del espacio circunflejo de las emociones y fueron altamente confiables y consistentes en todos los grupos. A diferencia de IAPS, OASIS clasifica sus estímulos en 4 clases; sus datos fueron recopilados en 2015 y, por lo tanto, OASIS presenta imágenes y evaluaciones más actuales.

Este repositorio permite el uso gratuito de las imágenes en estudios de investigación en línea y fuera de línea, ya que no están sujetas a las restricciones de confidencialidad que manejan algunos otros repositorios. Las imágenes de OASIS, junto con las clasificaciones normativas de valencia y excitación, están disponibles para descargarse y usarse. La Figura 2 muestra algunos ejemplos de los estímulos emocionales del repositorio OASIS.

Tabla 1. Comparación de características IAPS y OASIS.

	IAPS	**OASIS**
Objetivo	Evaluar un conjunto de imágenes emocionales, accesibles internacionalmente.	Crear un conjunto de estímulos estandarizados de acceso abierto.
Año	2008	2016
Origen	Universidad de Florida	Universidad de Harvard
Población de prueba	n = 100	n= 822
Rango de edad	Sin información	18-74 años (σ = 11.91)
Distribución de sexo	50 hombres 50 mujeres	420 mujeres 398 hombres
Tipo de muestra	Estudiantes universitarios y niños (7-14 años)	Trabajadores de estados unidos
Idioma	Inglés	Inglés
Datos	1196 imágenes En 20 conjuntos de 60 imágenes c/u	900 imágenes En 4 conjuntos con 225 c/u
Tipo de evaluación	Escala Likert de 9 puntos (SAM)	Escala Likert de 7 puntos
Valencia	$\bar{x}$ = 5.03 σ = 1.58	$\bar{x}$ = 4.33 σ = 1.09
Excitación / Activación	$\bar{x}$ = 4.82 σ = 2.16	$\bar{x}$ = 3.67 σ = 1.68
Dominio	$\bar{x} \approx$ 5.15 $\sigma \approx$ 2.08	N/A
Disponibilidad	Disponible con consentimiento	Disponible para descarga

1.3 Comparativa entre IAPS y OASIS

En términos generales ambos repositorios son parecidos, sin embargo, los estímulos, metodología de evaluación y contexto de las poblaciones de ambos estudios difieren. Una diferencia importante es que la población experimental de IAPS fue menor y más específica (estudiantes de universidad y niños), esto le ha atribuido cierto criticismo sobre su nivel de "internacionalidad".

Por otro lado, la población de OASIS fue conformada por personas de distintas edades, educación, trabajos, ideas y nivel de riqueza. A continuación, la Tabla 1 muestra una comparativa entre las características generales de los dos repositorios.

2. Metodología

2.1 Materiales

Para este estudio, se utilizaron los 2 repositorios emocionales: IAPS y OASIS filtrando cualquier imagen que incluyera contenido clasificado como "fuerte", esto quiere decir, que incluyera representaciones de: violencia, desnudos y/o muerte.

A continuación, se describe el proceso de selección de una muestra de estímulos para el estudio además de instrumentos de autoinforme para la obtención de información de los participantes.

2.1.1 Selección de estímulos IAPS

Ya que los estímulos IAPS no se encuentran clasificados por categorías, se tomaron imágenes que generalmente incluyeran escenas donde se muestran personas. Del subconjunto de imágenes de personas IAPS se seleccionaron las primeras 15 imágenes con un valor más alto para cada nivel de valencia (positiva, neutral y negativa), obteniendo una muestra de 45 estímulos emocionales IAPS.

2.1.2 Selección de estímulos OASIS

Para la selección de estímulos OASIS, de las cuatro clases de imágenes que ofrece OASIS, se realizó un filtrado por su nivel de valencia, clasificando cada clase en estímulos negativos, positivos y neutrales. Posteriormente, se realizó un análisis de cada uno de los niveles para cada clase, donde se encontró que:

- La valencia positiva se atribuía más a los estímulos de personas y animales.
- La valencia neutral se atribuía más a los estímulos de objeto y escena.
- La valencia negativa se atribuía más a los estímulos de personas y objetos.

De esta manera, se seleccionaron los 10 estímulos más prominentes para cada nivel de valencia (positivo, neutral y negativo) en los pares de clasificación OASIS ($10 \times 3 \times 2 = 60$ imágenes emocionales). Posteriormente, se realizó una prueba AB, la cual consistió decidir entre qué clase de estímulo emocional inspira cierto nivel de valencia.

En ese contexto, mediante una encuesta en línea donde se presentan pares de estímulos con niveles de valencia similares, se les preguntó a 20 personas jóvenes de universidad y posgrado cuál de esas imágenes le inspira más sentimientos positivos, neutrales o negativos.

Finalmente, se seleccionaron 15 imágenes de cada una de las clases que tuvieron más votos en la prueba AB: Animales para valencia positiva, personas para valencia negativa y objetos para valencia neutra, resultando en una muestra de 45 estímulos emocionales de OASIS.

2.1.3 Instrumentos

Para recopilar la información de evaluación de los participantes, se utilizó Google Forms para crear dos cuestionarios de cada evaluación de repositorio. Estos cuestionarios se componían de: Un mensaje de agradecimiento por participar, información general sobre el motivo de la investigación, información sobre qué datos se recopilan, instrucciones sobre cómo evaluar los estímulos y un aviso de privacidad de datos personales.

Para ilustrar el procedimiento de evaluación a los participantes, se utilizó el modelo SAM que representa la escala Likert de 9 puntos para valencia con imágenes de maniquís. Las indicaciones fueron:

"Por favor utiliza como referencia la siguiente escala visual del 1 al 9 para representar la emoción que te inspiran las imágenes, esta puede ir desde 1 (máxima tristeza) hasta 9 (máxima alegría), siendo 5 un sentimiento neutral".

Finalmente, los datos de cada participante fueron guardados en una hoja de cálculo que fue utilizada para analizar los resultados. Debido a que el estudio original de OASIS utiliza una escala Likert de 7 puntos y la evaluación se realizó con una escala de 9 puntos, se utilizó una transformación lineal de los resultados (1) para representar los datos en un rango de 0 a 1:

$$y = \frac{x - m}{a}. \tag{1}$$

2.2 Participantes

Para ambas pruebas, los participantes eran estudiantes del nivel de educación superior o posgrado. Las edades de los participantes no fueron especificadas, se preguntó sobre la pertenencia a los siguientes grupos de edad: “entre 15 y 18 años” (15.9%), “entre 19 y 27 años” (44.9%) y “más de 27 años” (39.1%).

El total de personas que participaron en la evaluación de IAPS fue de 69, 40.60% mujeres y 59.40% hombres. Mientras que 71 personas participaron en la evaluación de OASIS con 49.29% mujeres y 50.71% hombres.

2.3 Evaluación de la valencia

Se compartió el cuestionario entre distintos grupos de jóvenes mediante redes sociales y correos electrónicos institucionales (TecNM campus Cuautla y TecNM campus Cenidet). No se fijó un límite de tiempo para contestar y los participantes podían utilizar una computadora o teléfono inteligente, con conexión a Internet, para contestarla.

Luego de que cada participante evaluara los estímulos, los datos se guardaron en una hoja de cálculo. Posteriormente, se evaluó la consistencia interna de los instrumentos mediante el uso del Alfa de Cronbach. Finalmente, los datos de la evaluación de la valencia se analizaron y se compararon con los resultados de los estudios originales de Kurdi et al. y Lang et al.

3. Resultados

3.1 Validación de los instrumentos con Alfa de Cronbach

El coeficiente de Alfa de Cronbach aplicado a los 45 ítems del instrumento de evaluación IAPS, se calculó a través del software SPSS y su resultado es de 0.79, el que según la interpretación de Celina Oviedo y Campo Arias [18] tiene una “confiabilidad aceptable”, porque se encuentra en el rango 0.70-0.90.

Además, Nunnally y Bernstein [19] recomiendan un valor de al menos 0.70 para etapas tempranas de investigación hasta 0.80 para investigación aplicada. Por lo tanto, la consistencia interna del instrumento IAPS para esta investigación se considera aceptable para su aplicación.

Tabla 2. Evaluación y comparación de IAPS.

	IAPS	Lang et al. (2008)	Madera-Carrillo et al. (2015)	Este trabajo
Clase	Id	Valencia	Valencia	Valencia
Positivos	4612	6.82	7.6	6.88
	2341	7.38	7.63	6.86
	2340	8.03	7.63	7.35
	2347	7.83	7.65	7.22
	8170	7.63	7.66	6.74
	2151	7.32	7.68	7.01
	2332	7.64	7.71	7.57
	2091	7.68	7.73	7.58
	8190	8.1	7.74	7.22
	2165	7.63	7.74	6.83
	2655	6.88	7.79	7.41
	2155	6.78	7.83	6.52
	2057	7.81	7.89	6.75
	2540	7.63	8	6.49
	8496	7.58	8.11	7.51
Negativos	2688	2.73	2.1	2.59
	2703	1.91	2.18	1.96
	2095	1.79	2.77	1.67
	2456	2.84	3.71	2.04
	6212	2.19	2.04	1.96
	2345.1	2.26	2.67	3.96
	3160	2.63	3.08	4.68
	3180	1.92	2.8	5.42
	6311	2.58	2.61	4.70
	9041	2.98	2.86	5.33
	9330	2.89	2.53	5.25
	9332	2.25	2.87	5.07
	9920	2.5	2.82	4.68
	9341	3.38	2.35	5.20
	9342	2.85	2.41	5.42
Neutrales	2220	5.03	4.7	5.16
	2305	5.41	4.92	5.38
	2372	5.48	5.08	5.32
	2385	5.2	5	5.01
	2393	4.87	5.06	3.30
	2396	4.91	4.92	1.61
	2397	4.98	5.04	2.59
	2410	4.62	4.95	1.78
	2495	5.22	5.04	2.14
	2575	5.46	5.06	1.84
	2595	4.88	4.95	2.75
	2745.1	5.31	5	2.23
	2850	5.22	5.07	2.03
	2104	4.42	4.84	2.10
	9002	3.39	4.92	2.43

En el caso del instrumento de evaluación OASIS, el resultado del Alfa de Cronbach fue de 0.93, esto puede interpretarse como una consistencia interna "buena" para investigación aplicada. Sin embargo, un valor de consistencia interna en el rango de 0.91-1.00 indica la existencia de redundancia o duplicación en el instrumento[18].

3.2 Resultados de la evaluación de valencia

La Tabla 2 muestra los resultados de la evaluación para la muestra de estímulos en el repositorio de imágenes emocionales IAPS. Se observa una comparación de la media de la valencia (en escala SAM de 9 puntos) en los estudios de Lang et al. (2008), un trabajo similar donde se probó todo el conjunto de estímulos IAPS en población mexicana (Madera-Carrillo et al., 2015) y este trabajo.

Los estímulos IAPS se identifican con un número de 4 dígitos. Por otro lado, la Tabla 3 muestra los resultados para el repositorio OASIS. Donde se observa una comparativa entre la media de la valencia para los estímulos en el trabajo de Kurdi et al. (2016) y este trabajo. En este caso, las columnas "Valencia" índica el valor de la escala convertido en un rango de 0 a 1 utilizando la fórmula (1). Los estímulos OASIS se identifican por los nombres de cada imagen.

Se realizó una comparación entre los estudios originales y el actual utilizando pruebas t para dos muestras con varianzas iguales, donde la comparación de la evaluación de los estímulos IAPS en este estudio ($\bar{x} = 4.70$, $\sigma = 2.10$) con el de Lang et al. (2008) ($\bar{x} = 5$, $\sigma = 2.12$), demuestra que no existen diferencias significativas en la evaluación de la valencia realizada por los participantes, $t(88) = 0.67$, $p = 0.51$.

Además, la comparación entre este estudio y el de Madera-Carrillo et al. (2015) ($\bar{x} = 5.12$, $\sigma = 2.13$), también resulta en la falta de diferencias significativas para los estudios con los estímulos IAPS, $t(88) = 0.95$, $p = 0.34$.

Mientras que, en el caso de OASIS, la comparación del presente estudio ($\bar{x} = 0.54$, $\sigma = 0.16$) con el de Kurdi et al. (2016) ($\bar{x} = 0.53$, $\sigma = 0.19$), igualmente demuestra que no existen diferencias significativas en la evaluación de la valencia realizada por los participantes, $t(88) = -0.16$, $p = 0.88$.

Se observa que para la evaluación IAPS, la valencia para imágenes positivas se encuentra en un rango superior a 6, lo cual indica una buena relación con el ejemplo emocional de la escala SAM de un maniquí alegre, al mismo tiempo se observan valores similares entre los estudios. En el caso de algunos de los estímulos negativos y neutrales, los participantes evaluaron imágenes originalmente negativas como neutrales y viceversa (valores sombreados en gris).

En contraste, los resultados de la evaluación en OASIS se muestran bastante cercanos a los valores del estudio original. Sin embargo, se pueden observar algunas diferencias en valores de evaluación para algunos estímulos negativos y positivos.

4. Discusión

Como se ha mencionado en secciones anteriores, algunos de los estímulos del repositorio IAPS, en comparación con OASIS, fueron interpretados de forma diferente por la muestra de este estudio. Esto puede deberse a distintos factores sociodemográficos de la muestra.

Tabla 3. Evaluación y comparación de OASIS.

	OASIS	Kurdi et al. (2016)	Este trabajo
Clase	Identificador	Valencia	Valencia
Positivos	Bear 1	0.69	0.69
	Bear 3	0.74	0.73
	Bird 1	0.82	0.65
	Bird 2	0.75	0.77
	Bird 3	0.84	0.65
	Bird 4	0.76	0.74
	Cat 1	0.74	0.64
	Cat 2	0.75	0.64
	Cat 3	0.83	0.71
	Cat 4	0.81	0.77
	Cat 5	0.87	0.74
	Cat 6	0.62	0.78
	Cat 8	0.63	0.76
	Cat 9	0.72	0.79
	Cat 10	0.74	0.69
Negativos	Angry face 1	0.40	0.39
	Angry face 2	0.35	0.42
	Angry face 3	0.25	0.39
	Angry face 4	0.25	0.43
	Angry face 5	0.31	0.39
	Angry pose 1	0.30	0.43
	Angry pose 2	0.30	0.32
	Baby 7	0.33	0.38
	Bored pose 2	0.36	0.36
	Bored pose 3	0.35	0.38
	Bored pose 4	0.35	0.33
	Child labor 2	0.30	0.33
	Depressed pose 1	0.28	0.27
	Depressed pose 2	0.41	0.32
	Depressed pose 3	0.28	0.25
Neutrales	Alcohol 2	0.54	0.53
	Alcohol 3	0.53	0.51
	Alcohol 7	0.58	0.53
	Bark 1	0.51	0.52
	Bark 2	0.54	0.57
	Bed 1	0.50	0.62
	Bottle 1	0.53	0.62
	Bricks 1	0.53	0.55
	Car 2	0.49	0.48
	Cardboard 1	0.50	0.49
	Cardboard 2	0.50	0.48
	Cardboard 3	0.46	0.45
	Cold 1	0.50	0.52
	Cotton swabs 1	0.52	0.52
	Cotton swabs 2	0.51	0.51

Por ejemplo, la diferencia de edades, contexto y culturas entre grupos. Por ejemplo, a comparación de los estudios de 2008 y 2015, los participantes de este estudio (2022), pertenecen a una generación con mayor acceso a las tecnologías y, por lo tanto, los participantes se han enfrentado a más cantidad de estímulos y contenidos en los medios electrónicos.

La población de Lang et al. consistió en grupos de personas pertenecientes a la cultura de Estados Unidos y con un rango de edad joven (niños y estudiantes de universidad). Mientras que la muestra de este estudio contenía grupos de jóvenes y adultos (estudiantes de posgrado) pertenecientes a la cultura mexicana.

Por otro lado, aunque el estudio de Madera-Carrillo también manejo una población similar a la de este estudio, se observan las diferencias en la percepción de los estímulos. Las imágenes negativas que fueron interpretadas como neutrales muestran personas con rasgos de tristeza moderada, lesiones y lugares contaminados.

Mientras que las imágenes neutrales que fueron percibidas negativamente incluyen personas con rasgos de vejez y aburrimiento en distintos escenarios. En el caso de los estímulos OASIS los participantes interpretaron los estímulos de manera similar al estudio original. Sin embargo, pueden notarse algunas diferencias.

Por ejemplo, la valencia para estímulos positivos de animales fue más baja, lo cual podría indicar la costumbre de observar medios de animales en las redes sociales. En el caso de los estímulos negativos, las imágenes que presentan personas con rasgos de ira fueron evaluadas como un poco más positivas que el estudio original, esto podría deberse a que, en las imágenes, la emoción de ira en los rostros de las personas llega a ser un tanto exagerada y detectada como falsa.

También se debe mencionar que las referencias a repositorios de imágenes clásicas y conceptos de la evaluación emocional son de referencias antiguas, sin embargo, hoy en día continúan siendo la base de muchos trabajos de investigación en el campo de las emociones.

5. Conclusiones

El presente estudio muestra los resultados de una muestra de 69 personas en la evaluación IAPS y 71 en OASIS. Los resultados se alinean con los estudios originales, sin embargo, para la muestra de participantes IAPS, se observa un sesgo en los estímulos de tipo negativo y neutrales en donde los participantes no percibieron la etiqueta establecida del estímulo, esto podría deberse a distintos factores de edad, cultura e interpretación por lo que se recomendaría seleccionar otros de los estímulos IAPS en poblaciones similares a las de este estudio para lograr el efecto deseado del estímulo.

Así mismo, como trabajo futuro se debería buscar diferencias de la evaluación de las imágenes en poblaciones de diversos contextos para complementar la selección objetiva de los mismos. En el caso de OASIS, se concluye que los estímulos logran inducir emociones similares a las del estudio original sin mayor problema. Además, que estas imágenes contienen un rango mayor de clases que podrían ser utilices en estudios donde la misma muestra de participantes sea enfrentada a distintos tipos de estímulos positivos, negativos y neutrales.

Se debe considerar la interpretación subjetiva de los estímulos para ser utilizados en investigaciones, los factores sociales y demográficos que componen a la muestra de participantes siempre podrían generar sesgos en la recepción de los estímulos. Además, la experiencia personal del participante puede generar que ciertos estímulos causen diferentes emociones, por ejemplo, la imagen de una familia podría causar alegría en una persona y tristeza en otra.

Por lo tanto, se recomienda que los estudios que hagan uso de repositorios de imágenes emocionales, definan correctamente criterios de inclusión y exclusión para la muestra experimental, así mismo la selección de estímulos emocionales apropiados para el tipo de experimento a realizar y el contexto de la muestra. Estos repositorios de imágenes emocionales no solo son utilizados en ramas de la computación afectiva y la interacción humano-computadora (HCI), también son base de estudios en psicología, medicina y comportamiento humano.

Por ejemplo, existen estudios donde se diferencia el comportamiento ocular de personas con depresión al observar distintos estímulos emocionales utilizando aprendizaje máquina. Por lo tanto, este estudio puede ayudar a futuras investigaciones a seleccionar un repositorio de imágenes emocionales.

Referencias

1. Russell, J. A.: A circumplex model of affect. Journal of Personality and Social Psychology, vol. 39, no. 6, pp. 1161–1178 (1980) doi: 10.1037/h0077714
2. Bradley, M. M., Lang, P. J.: Measuring emotion: The self-assessment manikin and the semantic differential. Journal of Behavior Therapy and Experimental Psychiatry, vol. 25, no. 1, pp. 49–59 (1994) doi: 10.1016/0005-7916(94)90063-9
3. Galentino, A., Bonini, N., Savadori, L.: Positive arousal increases individuals' preferences for risk. Frontiers in Psychology, vol. 8 (2017) doi: 10.3389/fpsyg.2017.02142
4. Frijda, N. H.: The emotions. Cambridge University Press (1986)
5. Bradley, M. M., Lang, P. J.: International affective picture system. Encyclopedia of Personality and Individual Differences, pp. 1–4 (2017) doi: 10.1007/978-3-319-28099-8_42-1
6. Tottenham, N., Tanaka, J. W., Leon, A. C., McCarry, T., Nurse, M., Hare, T. A., Marcus, D. J., Westerlund, A., Casey, B., Nelson, C.: The NimStim set of facial expressions: Judgments from untrained research participants. Psychiatry Research, vol. 168, no. 3, pp. 242–249 (2009) doi: 10.1016/j.psychres.2008.05.006
7. Chen, L. F., Yen, Y. S.: Taiwanese facial expression image database. Brain Mapping Laboratory, Institute of Brain Science, National Yang-Ming University, Taipei (2007)
8. Lyons, M., Kamachi, M., Gyoba, J.: The japanese female facial expression (JAFFE) dataset (1998) doi: 10.5281/ZENODO.3451524
9. Ekman, P., Friesen, W. V.: Pictures of facial affect. Consulting Psychologists Press, Palo Alto, CA (1976)
10. Lundqvist, D., Flykt, A., Öhman, A.: Karolinska directed emotional faces. APA PsycTests (1998) doi: 10.1037/t27732-000
11. Gong, X., Huang, Y. X., Wang, Y., Luo, Y. J.: Revision of the chinese facial affective picture system. Chinese Mental Health Journal, vol. 25, no. 1, pp. 40–46 (2011)
12. Egger, H. L., Pine, D. S., Nelson, E., Leibenluft, E., Ernst, M., Towbin, K. E., Angold, A.: The NIMH child emotional faces picture set (NIMH-ChEFS): A new set of children's facial emotion stimuli. International Journal of Methods in Psychiatric Research, vol. 20, no. 3, pp. 145–156 (2011) doi: 10.1002/mpr.343

13. Bao, S., Ma, H., Li, W.: ThuPIS: A new affective image system for psychological analysis. In: IEEE International Symposium on Bioelectronics and Bioinformatics, pp. 1–4 (2014) doi: 10.1109/isbb.2014.6820908
14. Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D. H. J., Hawk, S. T., Knippenberg, A. V.: Presentation and validation of the radboud faces database. Cognition and Emotion, vol. 24, no. 8, pp. 1377–1388 (2010) doi: 10.1080/02699930903485076
15. Kurdi, B., Lozano, S., Banaji, M. R.: Introducing the open affective standardized image set (OASIS). Behavior Research Methods, vol. 49, no. 2, pp. 457–470 (2016) doi: 10.3758/s13428-016-0715-3
16. Moltó, J., Segarra, P., López, R., Esteller, Fonfría, A., Pastor, M. C., Poy, R.: Adaptación española del "International Affective Picture System" (IAPS). Tercera Parte, Anales de Psicología, vol. 29, no. 3 (2013) doi: 10.6018/analesps.29.3.153591
17. Madera-Carrillo, H., Zarabozo, D., Ruiz-Díaz, M., Saez-de-Nanclares, P. B.: El sistema internacional de imágenes afectivas (IAPS) en población mexicana. Autoevaluación con Maniquíes y Etiquetas (2015) doi: 10.13140/RG.2.1.3220.3683
18. Oviedo, H. C., Campo-Arias, A.: Aproximación al uso del coeficiente alfa de Cronbach. Revista Colombiana de Psiquiatría, vol. 34, no. 4, pp. 572–580 (2005)
19. Nunnally, J., Bernstein, I. H.: Psychometric theory. McGraw-Hill Companies Incorporated (1994)

ISSN 1870-4069

Impulsando los rostros del futuro: Evaluación comparativa de tecnologías de captura de movimiento facial para humanos digitales

Sharon Ramírez Lechuga[1], Carlos Vilchis[1],
Miguel Gonzalez Mendoza[1], Armando Rodríguez Mendoza[2],
Carmina Pérez Guerrero[2]

[1] Tecnológico de Monterrey,
Escuela de Ingeniería y Ciencias,
México

[2] Eugenia Virtual Humans S.A. de C.V.,
Laboratorio de Investigación,
México

{A01379035,carlos.vilchis, mgonza}@tec.mx,
{armando,carmina}@eugenia.tech

Resumen. El creciente universo de creadores de contenido virtuales, avatares del metaverso, y humanos digitales en general, ha creado una oportunidad para integrar soluciones de captura facial en un amplio panorama de nuevas aplicaciones para la industria de creación de contenido. Junto con este crecimiento se ha incrementado la demanda por humanos digitales que generen empatía y cuenten con un mejor desempeño en sus expresiones faciales. Es por esto que en el presente artículo se exploran las principales codificaciones faciales empleadas para la captura de movimiento facial y las diversas soluciones existentes para dar vida a humanos digitales. Adicionalmente, se presenta un experimento realizado con un humano digital dentro de un ambiente de realidad virtual para medir el vínculo empático creado a partir de algunas tecnologías recientes de captura facial, Faceware, Live Link UE, y Avatary. Los resultados exploran la percepción de determinadas expresiones emocionales, la respuesta empática, y la semblanza de familiaridad que reflejan las soluciones disponibles. Finalmente, se discute la necesidad de alternativas nuevas y accesibles con una codificación más expresiva, como medio para abrir el panorama a un amplio campo de investigación que busca mejorar la captura facial.

Palabras clave: Realidad virtual, captura de movimiento facial, interfaces humano-computadora, humanos digitales, codificación facial.

Driving the Faces of the Future: Benchmarking Facial Motion Capture Technologies for Digital Humans

Abstract. The growing universe of virtual content creators, metaverse avatars, and digital humans in general has created an opportunity to integrate facial capture solutions into a broad landscape of new applications for the content

creation industry. Along with this growth, the demand has increased for digital humans who generate empathy and have a better performance in their facial expressions. That is why this article explores the main facial codifications used for facial movement capture and the various existing solutions to bring digital humans to life. Additionally, an experiment carried out with a digital human within a virtual reality environment is presented to measure the empathic link created from some recent facial capture technologies, Faceware, Live Link UE, and Avatary. The results explore the perception of certain emotional expressions, the empathic response, and the semblance of familiarity that reflect the available solutions. Finally, the need for new and accessible alternatives with a more expressive coding is discussed, as a means to open the panorama to a wide field of research that seeks to improve facial capture.

Keywords: Virtual reality, facial motion capture, human-computer interfaces, digital humans, facial codification.

1. Introducción

Desde hace varios años, los avatares forman parte de un área de investigación experimental que busca explorar las interfaces humano-computadora [7]. Ahora, la tendencia de humanos digitales tiene expectativas importantes para los próximos 5 a 8 años del mercado de consumo [5]. Las posibles aplicaciones incluyen servicio al cliente, asistentes conversacionales, y soporte técnico virtual [6].

Recientemente, soluciones como Metahumans de Epic Games [11], han creado nuevas oportunidades, ya que pone a disposición del público una amplia gama de humanos digitales realistas y gratuitos, listos para ser utilizados dentro de procesos profesionales. Esto convierte a los humanos digitales realistas en herramientas sencillas y asequibles, que con tecnologías como la captura de movimiento y los gráficos 3D en tiempo real, se logran resultados de interacción mejorados [16].

Todos estos avances tienen la capacidad de cambiar la percepción de un ser humano digital gracias a la construcción de un vínculo emocional [24]. Este vínculo es necesario para impulsar el realismo, a fin de superar el Valle Inquietante [22], un término creado para describir el punto donde la respuesta emocional a representaciones humanas que tienen una apariencia y comportamiento similar al de un ser humano, causan una reacción negativa de extrañeza e inquietud.

Se han empleado humanos digitales para determinar la respuesta empática, la aceptabilidad, y la calidad de la interacción entre las computadoras y los humanos [27, 21]. Utilizando metodologías de seguimiento modernas, los resultados han mejorado [16] en comparación con experimentos realizados unos años atrás [1].

Entonces, la captura facial y su rendimiento es mejor ahora, pero ¿cómo mejoran las tecnologías de seguimiento facial disponibles, la respuesta empática de los humanos digitales democratizados de última generación? Para dar respuesta a esta pregunta y para aprovechar las interfaces empáticas y realistas modernas, este documento propone un experimento de percepción que involucra la realidad virtual (RV).

Tabla 1. Comparación de las principales soluciones analizadas para la captura de movimiento facial en tiempo real usadas en humanos digitales.

Solución de Captura Facial	Número de Blendshapes	Capacidad de Tiempo Real	Calibración Específica al Sujeto	Basado en Inteligencia Artificial	Codificación Facial	Inversión Aproximada
Faceware Studio	59	✓	×	×	FAPs	$2,340 US / Year
Live Link Face UE	51	✓	×	×	FAPs	Gratuito
Avatary	Ilimitadas	✓	✓	✓	FAPs & FACS	$2,388 US / Year

Los experimentos se basan en la interacción con un MetaHumano [11], impulsado con un conjunto contemporáneo de sistemas de captura facial: Faceware Studio de Faceware Tech, Live Link Face UE de Epic Games, y Avatary de Facegood. Todas son soluciones de vanguardia disponibles para la investigación y la creación de contenido, sin embargo, según nuestros conocimientos, este es el primer trabajo que los compara entre sí.

Por lo que los resultados obtenidos son importantes para medir el funcionamiento de estas tecnologías de seguimiento facial para ofrecer un rendimiento realista y de calidad. El resto del documento está estructurado de la siguiente manera: la sección 2 introduce el concepto de codificación facial utilizado para el seguimiento facial. La sección 3 presenta las soluciones de seguimiento facial contemporáneas.

La sección 4 detalla los métodos utilizados, los datos demográficos de los sujetos y el análisis estadístico aplicado a los resultados. La sección 5 muestra los resultados de la investigación. Finalmente, la sección 6 ofrece una discusión basada en los hallazgos, las áreas de oportunidad para futuras investigaciones y resume el trabajo de investigación presentado en este documento.

2. Codificación facial para la captura de movimiento

Las herramientas de seguimiento facial deconstruyen los rostros humanos para replicar su función, y la industria respalda este proceso con metodologías y estándares llamados Codificación facial. Hay dos corrientes principales, el Sistema de Codificación de Acción Facial (FACS por sus siglas en inglés) y los Parámetros de Animación Facial (FAPs por sus siglas en inglés).

La metodología FACS fue propuesta por Paul Ekman en los años 70s, para entender cómo las emociones y las expresiones faciales se relacionan con los huesos y músculos de nuestro rostro [10]. El modelo clasifica las expresiones facial por medio de etiquetas numeradas con diferentes niveles de intensidad (A, B, C, D o E), denominados Unidades de Acción. Estas unidades, con ciertas configuraciones, pueden representar emociones específicas denominadas FACS emocionales (emFACS) [13].

Hasta el día de hoy, esta metodología se utiliza como una de las formas más fiables de comprender las expresiones humanas. Más tarde, en los años 90, Moving Pictures Experts Group (MPEG-4) creó un estándar internacional para representar el habla y los gestos humanos en la animación, y un componente de ese estándar es el modelo de FAPs [23], que describe los movimientos faciales a medida que la unidad cambia desde una cara neutra.

Este modelo ha sido el estándar más común utilizado en escenarios de animación tradicional durante casi dos décadas debido a la simplicidad de su implementación en dibujos animados y modelos 3D [4]. Sin embargo, a medida que el realismo se convirtió en una prioridad, la industria comenzó a alcanzar los límites de esta codificación facial.

El modelo FACS se ha utilizado para experimentos con humanos digitales desde 1995 [19, 26], con investigaciones que continúan hasta el siglo XXI [20], evocando emociones de manera confiable en los rostros de los avatares.

Además, el modelo emFACS tiene un conjunto específico de combinaciones de movimientos musculares creadas para determinar las expresiones más comunes junto con la emoción humana correspondiente.

El modelo emFACS original incluye emociones como tristeza, neutral, disgusto, enojo, felicidad, sorpresa, y miedo. A base de este estándar, es posible verificar la efectividad de las expresiones faciales humanas digitales.

3. Estado del arte en soluciones de captura facial

Las herramientas de captura facial se hicieron populares a mediados de la década del 2010 [3], tras la mejora del procesamiento de video en tiempo real para hacer seguimiento en vivo. Las primeras opciones se basaban en Head-Mounted Cameras (HMC), que solo graban vídeo para ser procesado en postproducción, como Vicon Cara o primeras versiones de Faceware con cámaras GoPro.

Más tarde, las herramientas de seguimiento en vivo basadas en el reconocimiento facial estuvieron disponibles con las cámaras de los teléfonos inteligentes, como la cámara TrueDepth incluida en los iPhone y utilizada por Live Link Face UE. Hoy en día, las opciones más populares y disponibles para el seguimiento facial son Faceware [2], Live Link Face UE [14] y Avatary [12].

Faceware y Live Live Link Face UE utilizan el método de 51 o más blendshapes porque se ha convertido en el estándar de conjuntos de expresiones en la animación facial debido a su similitud con la codificación FAP. Solo Avatary de Facegood tiene la opción de impulsar conjuntos de expresiones personalizados más grandes, como lo requiere el modelo FACS. Cada una de estas soluciones están compiladas en la tabla 1, teniendo en cuenta características específicas que se compararon para los fines de esta investigación.

4. Configuración experimental

El análisis para evaluar la percepción de muestras emotivas visuales en la interacción humano-computadora, por lo general involucra la medición del Efecto del Valle Inquietante, que se puede realizar a través de la evaluación de las dimensiones de la personalidad con la encuesta Big–Five [8] como en Hyde et al. [18], a través de la percepción de las dimensiones afectivas definidas por Ho y MacDorma [17] o a través de ensayos interactivos como la modificación al experimento del Mago de Oz, aplicado con humanos digitales por Seymour, Riemer y Kay [24], donde un humano digital se expone a sujetos humanos, mitigando el efecto del Valle Inquietante con interactividad.

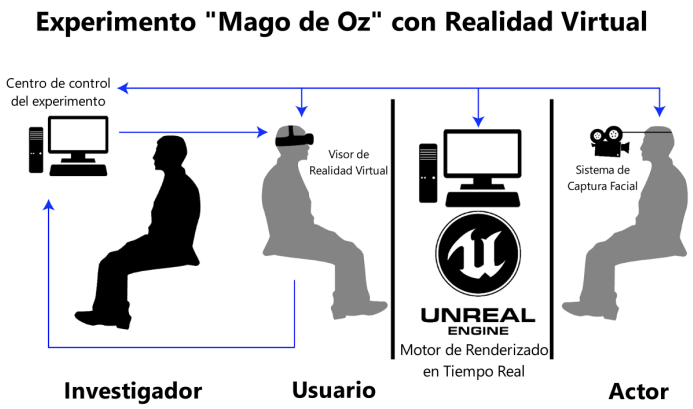


Fig. 1. El diseño del Experimento Mago de Oz utilizado en este documento, basado en el trabajo de Seymour [24] para integrar la interactividad en tiempo real como una variable clave, que se complementa con el uso de un visor de RV en el presente trabajo.

Para lograr esto se necesita un conjunto específico de elementos tecnológicos y la definición de un sistema de evaluación que permita explorar cómo se percibe a un Humano Digital, junto con la efectividad de los sistemas de seguimiento facial. Estos requisitos llevaron a la creación de pruebas específicas que mezclan aspectos psicológicos, de familiaridad e interactividad.

El experimento propuesto en esta investigación agrega una nueva capa de interactividad al experimento modificado del ”Mago de Oz”, al involucrar el uso de la realidad virtual, como se ilustra en la Figura 1. El estudio se realiza entre varios sujetos para evaluar diferentes tecnologías de captura facial a través de la interacción con un Metahumano [11] en un entorno de realidad virtual.

El diseño del estudio propuesto expone a cada sujeto humano a una sola tecnología de captura facial y una sola exposición interactiva, evitando cualquier sesgo debido a experiencias previas o expectativas emocionales. El experimento cuenta con dos salas diferentes que permiten un espacio suficiente para la experiencia del usuario y el rendimiento de los sistemas de captura facial. Las salas son la sala experimental y la sala de control.

La sala experimental es donde el usuario está expuesto a un humano digital a través de un visor de RV. Para el experimento descrito en este documento, los visores contienen animaciones de emFAC pre-grabadas con los sistemas de captura facial obtenidas de un actor, que luego se transmiten a través de un motor de 3D en tiempo real, en este caso, Unreal Engine 5.0.2.

La sala de control, por otro lado, contiene la estación de trabajo donde las expresiones grabadas a partir de los diferentes sistemas de seguimiento facial se transmiten en diferentes sesiones y en orden aleatorio. Un Metahumano [11] descargado con la más alta calidad, se carga en Unreal Engine y se procesa en una computadora con un procesador Intel i7-8700K de 12 núcleos, 40 GB DDR4 de RAM y una tarjeta GPU RTX 3070 con 12 GB GDDR6X.

La computadora entrega la representación final en tiempo real de las secuencias animadas a la sala experimental. El sistema de comunicación conecta ambas estaciones de trabajo en una dirección de transmisión unidireccional.

4.1. Diseño de experimentos

El experimento expone a un sujeto humano con un visor de RV, el usuario, a un entorno virtual con un humano digital y le asigna la tarea de resolver una encuesta dentro del ambiente de RV. El usuario de RV se encuentra aislado dentro de la experiencia para no sentir presión ni estímulos externos. El experimento toma el tiempo necesario para cubrir un conjunto aleatorio de emFACS creados para construir un puente empático entre los sujetos, además de una actuación adicional de 60 segundos, donde el humano digital habla sobre su vida, sus cosas favoritas, pensamientos, entre otras cosas, para crear una experiencia empática con el sujeto.

En la Figura 2 se puede observar un ejemplo del emFACS de enojo capturado por las diferentes soluciones de seguimiento facial exploradas en este experimento.Las preguntas de la encuesta se dividen en dos grupos. El primer grupo de preguntas está relacionado con la identificación del conjunto de emFACS específicos representados en el humano digital (tristeza, neutral, disgusto, enojo, felicidad, sorpresa, y miedo), que consiste en la presentación de una expresión emFACS seleccionada aleatoriamente, y el sujeto tiene la tarea de seleccionar el emFACS percibido.

El segundo grupo de preguntas está relacionado con la encuesta Big–Five de empatía y familiaridad [18], que consta de 5 preguntas en las que el usuario comparte su opinión sobre cuán confiable, amigable, familiar, atractivo y realista parece el humano digital, con opciones para cada cualidad presentadas en una Escala Likert de 1 a 7 para su análisis, donde 1 significa fuerte desacuerdo, 7 significa fuerte acuerdo y 4 es una respuesta neutral.

4.2. Participantes

El experimento se realizó con un grupo de 42 personas que fueron seleccionadas al azar de la comunidad que deambula cerca del laboratorio de ingeniería. Los sujetos fueron expuestos a solo una tecnología de seguimiento facial cada uno, lo que resultó en 3 grupos de 14 sujetos, un grupo por tecnología. Se les pidió que interactuaran con el experimento durante cinco minutos sin información previa sobre la experiencia interactiva.

Cuando se les presentó el experimento, se preguntó a los sujetos sobre su género, edad y si estaban familiarizados con humanos digitales o RV. La distribución del grupo fue un 58,1 % de hombres y un 41,9 % de mujeres. La edad media del grupo es de 25.7 años con una desviación estándar de 6.9.

Los sujetos que conocían el concepto de humanos digitales antes del experimento representan el 45,2 %, y los que no conocían el concepto de humanos digitales representan el 54,8 %. Los sujetos que estaban familiarizados con la RV antes del experimento representan el 71 %, mientras que los que no estaban familiarizados con la RV representan el 29 %.

Fig. 2. Una muestra del emFACS de enojo. De izquierda a derecha se muestra una captura del video del actor, la expresión por Live Link Face UE, la expresión por Faceware y la expresión por Avatary.

4.3. Análisis estadístico

Para validar los resultados del experimento presentado, se emplean diferentes métodos de prueba de hipótesis para la identificación emFACS y la encuesta empática y de familiaridad Big–Five con un valor de nivel de significacia de $0{,}05$. Se tomó en cuenta que cada sujeto fue expuesto aleatoriamente a una sola tecnología y experimentó una sola ejecución del experimento, así como el hecho que los emFACS se presentaron en un orden aleatorio por tecnología. Más información sobre los enfoques estadísticos utilizados se detalla en las siguientes subsecciones.

Identificación emFACS Para las pruebas de percepción resultantes, el enfoque propuesto es una prueba U de Mann-Whitney para analizar la diferencia entre los porcentajes obtenidos a partir de una matriz de confusión de los resultados. También se propone una prueba de bondad de ajuste Chi-Cuadrado de Pearson, con las observaciones percibidas y las observaciones esperadas donde todos los emFACS serían correctamente identificados.

La prueba U de Mann-Whitney se puede utilizar para comprobar si dos muestras independientes tienen una diferencia estadísticamente significativa. Esta prueba también se considera el equivalente no paramétrico de la prueba t de independencia de dos muestras. Los supuestos para la prueba U de Mann-Whitney incluyen muestras aleatorias e independientes, así como un tamaño de muestra pequeño con menos de 30 muestras.

Dado que las muestras se obtienen por tecnología, donde un sujeto se expone una vez a una sola tecnología, con un orden aleatorio de emFACS por ejecución, la suposición de muestras aleatorias e independientes se aplica al experimento. Dado que el grupo de muestra por tecnología es de 14 sujetos, esta condición también se aplica al experimento.

Para esta prueba, la hipótesis nula supone que ninguno de los modelos comparados funciona mejor que el otro, y la hipótesis alternativa supone que los rendimientos de los modelos comparados no son iguales. El valor crítico U en el nivel de significancia $0{,}05$ es 8.

La prueba de chi-cuadrado de Pearson es una prueba estadística para datos categóricos. Se puede usar para probar la bondad de ajuste, la independencia o la homogeneidad. La prueba de bondad de ajuste chi-cuadrado se puede usar cuando se trata de una variable categórica. Le permite probar si la distribución de frecuencias de la variable categórica es significativamente diferente de las expectativas de proporciones iguales.

Para esta prueba, la hipótesis nula asume que los emFACS percibidos obtenidos a partir de tecnologías de seguimiento facial están en proporciones iguales a los emFACS observados, y la hipótesis alternativa asume que los emFACS percibidos a partir de tecnologías de seguimiento facial están en diferentes proporciones a los emFACS observados.

Cuestionario de empatía y familiaridad Big–Five. Los resultados obtenidos a partir de los resultados de la escala de Likert se comparan a través del análisis estadístico mediante la prueba de hipótesis. El enfoque propuesto es una prueba U de Mann-Whitney, que se ha descrito en la subsección anterior. Dado que las muestras se obtienen por tecnología, donde un sujeto se expone una vez a una sola tecnología, con un rendimiento de 60 segundos, la suposición de muestras aleatorias e independientes es adecuada.

Finalmente, dado que el tamaño de muestra por tecnología es de 14, inferior a 30, se puede considerar un tamaño de muestra pequeño. Para esta prueba, la hipótesis nula asume que las dos tecnologías comparadas tienen respuestas empáticas y similares, y la hipótesis alternativa asume que existe una diferencia estadísticamente significativa en las respuestas empáticas y de familiaridad entre las dos tecnologías comparadas. El valor crítico U en el nivel de significancia $0{,}05$ es 7.

5. Resultados

5.1. Identificación de emFACS

El experimento de identificación emFACS consistió en el uso de 7 conjuntos diferentes de videos en orden aleatorio de humanos digitales que expresan emociones. Se encargó a un grupo de sujetos que observaran y reconocieran los emFACS que podrían ser tristeza, neutral, disgusto, enojo, felicidad, sorpresa, o miedo.

Matriz de confusión. Los resultados del reconocimiento general se pueden observar en la matriz de confusión ilustrada en la Tabla 2. Dado que en otros casos de investigación de reconocimiento de emociones [15, 25, 9], una precisión considerada fiables en escenarios aleatorios podría ir del 60 %-86 %, el umbral utilizado para evaluar los resultados de estos experimentos como confiables van desde el 60 % en adelante.

Con eso en consideración, se puede observar que la mayoría de los sujetos pueden reconocer de manera confiable ciertas expresiones a través de las tres tecnologías, como Neutral (Faceware: 100 %, Avatary: 100 %, Live Link Face UE: 100 %), Felicidad (Faceware: 71,43 %, Avatary: 92,86 %, Live Link Face UE: 78,57 %), y Sorpresa (Faceware: 78,57 %, Avatary: 92,86 %, Live Link Face UE: 92,86), con Avatary y Live Link Face UE generalmente mostrando mejor desempeño que Faceware.

Tabla 2. Porcentaje de precisión de las emociones reconocidas por el grupo de prueba. FW representa Faceware; AV representa Avatary; LLF representa Live Link Face UE.

	emFACS Reconocidas																					
	Tristeza			**Neutral**			**Disgusto**			**Enojo**			**Felicidad**			**Sorpresa**			**Miedo**		
	FW	**AV**	**LLF**	**FW**	**AV**	**LLF**	**FW**	**AV**	**LLF**	**FW**	**AV**	**LLF**	**FW**	**AV**	**LLF**	**FW**	**AV**	**LLF**	**FW**	**AV**	**LLF**
Tristeza	7.14	64.28	28.57	0	0	0	0	21.43	0	0	7.14	0	0	0	14.29	35.71	0	50	57.14	7.14	7.14
Neutral	0	0	0	100	100	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Disgusto	7.14	78.57	0	0	0	0	64.29	7.14	50	21.43	0	14.29	7.14	0	21.42	0	14.29	0	0	0	14.29
Enojo	0	0	0	0	0	0	71.43	21.43	42.86	14.29	78.57	42.86	7.14	0	14.29	0	0	0	7.14	0	0
Felicidad	0	0	0	0	0	7.14	7.14	0	7.14	0	0	7.14	71.43	92.86	78.57	21.43	7.14	0	0	0	0
Sorpresa	0	0	0	0	0	0	0	7.14	0	0	0	0	0	0	0	78.57	92.86	92.86	21.43	0	7.14
Miedo	7.14	7.14	0	0	0	7.14	42.86	0	28.57	7.14	0	14.29	0	0	0	14.29	92.86	42.86	28.57	0	7.14

En Tristeza, solo Avatary obtuvo un resultado fiable del 64,28 %, mientras que con Faceware se confundía mayoritariamente con Miedo (57,14 %) y Sorpresa (35,71 %), y con Live Link Face UE se confundiá mayoritariamente con Sorpresa (50 %). Para Disgusto, solo Faceware obtuvo un resultado fiable del 64,29 %, mientras que con Avatary se confundió mayoritariamente con Tristeza (78,57 %), y con Live Link Face UE se confundió un poco con Felicidad (21,42 %).

Para Enojo, solo Avatary obtuvo un resultado fiable del 78,57 %, mientras que Faceware y Live Link Face UE se confundieron mayoritariamente con Disgusto (71,43 % y 42,86 % respectivamente).

Finalmente, para Miedo, las tres tecnologías funcionaron mal (Faceware: 28,57 %, Avatary: 0 %, Live Link Face UE: 7,14 %), con Faceware se confundió mayoritariamente con Disgusto (42,86 %), con Avatary fue completamente confundido con Sorpresa (92,86 %), y con Live Link Face UE se confundió mayoritariamente con Sorpresa también (42,86 %).

Prueba U de Mann-Whitney. Con base en los valores obtenidos de la prueba U de Mann-Whitney sobre los resultados de percepción emFACS, se encontró que entre Link Face UE y Faceware, el valor U es de 22 y el valor p es de $0{,}7949$; entre Link Face UE y Avatary, el valor U es $22,5$ y el valor p es $0,8493$; entre Avatary y Faceware el valor U es 21 y el valor p es $0{,}7039$.

Dado que el valor p de todos los sistemas de seguimiento facial es mayor que nuestro umbral de significancia asumido ($\alpha = 0{,}05$), y todos los valores U son mayores que el valor crítico en ese nivel de importancia ($U = 8$), no se puede rechazar la hipótesis nula y se concluye que no hay pruebas suficientes para afirmar que existe una diferencia estadísticamente significativa entre cualquiera de las tecnologías de seguimiento facial.

Sin embargo, según los valores p y dado que cada tecnología comparte el mismo tamaño de muestra, el orden de las tecnologías desde el valor p más pequeño hasta el valor p más grande es Avatary y Faceware, Link Face UE y Faceware, y Link Face UE y Avatary.

Prueba de Chi-Cuadrado. Con base en los valores obtenidos de la prueba Chi-Cuadrado en los resultados de percepción de emFACS, se encontró que para Faceware, el valor p es $0{,}00153$ y el estadístico de prueba X^2 es $21{,}44$; para Link Face UE, el valor p es $0{,}00201$ y la estadística de prueba X^2 es $20{,}78$; para Facegood, el valor p es $0{,}00435$ y la estadística de prueba X^2 es $18{,}89$.

Dado que el valor p de todos los sistemas de seguimiento facial es más pequeño que nuestro umbral de importancia asumido ($\alpha = 0{,}05$), y todas las estadísticas de prueba X^2 no están en la región de aceptación de $95\,\%$, rechazamos nuestra hipótesis nula y asumimos que existe una diferencia estadísticamente significativa entre el emFACS percibidas de las tecnologías de seguimiento facial y las observaciones esperadas de emFACS. Sin embargo, según los valores p y dado que cada tecnología comparte el mismo tamaño de muestra, el orden de las tecnologías del valor p más pequeño al valor p más grande es Faceware, Link Face UE y Avatary.

5.2. Encuesta empática y de familiaridad big-five

La encuesta Big–Five consistió en 5 preguntas en las que los sujetos compartían su percepción de cuán confiable, amigable, familiar, atractivo y realista parecía el ser humano digital durante una actuación adicional de 60 segundos. Las respuestas se representaron en una escala de Likert de 1 a 7, donde 1 significa fuerte desacuerdo, 7 significa fuerte acuerdo y 4 es una respuesta neutral.

Escala Likert. Los resultados se resumen en la Fig. 3. Se puede observar que en la categoría de Realista, Faceware y Avatary presentan un mayor porcentaje que Live Link Face UE, donde Faceware es el que presenta mayor porcentaje por una pequeña diferencia.

En la categoría Atractivo, Avatary presenta un porcentaje visiblemente menor en comparación con las otras dos tecnologías, sin embargo, en la categoría Familiar se presenta el comportamiento contrario. Finalmente, Live Link Face UE tiene una clara ventaja sobre las otras tecnologías en las categorías de Amigable y Confiable.

Prueba U de Mann-Whitney. Con base en los valores obtenidos de la Prueba U de Mann-Whitney en los resultados de la Encuesta Big–Five, se encontró que entre Faceware y Link Face UE, el valor U es 8 y el valor p es $0{,}13362$; entre Avatary y Link Face UE, el valor U es 6 y el valor p es $0{,}02574$; entre Avatary y Faceware, el valor U es 12 y el valor p es $0{,}27572$.

Solo el valor p de la comparación entre Avatary y Live Link Face UE es menor que nuestro umbral de importancia asumido ($\alpha = 0{,}05$) y el valor U es menor que el valor crítico en ese nivel de importancia ($U = 7$), por lo que es el único caso en el que rechazamos nuestra hipótesis nula y asumimos que existe una diferencia estadísticamente significativa en las respuestas empáticas y de similitud entre Avatary y Link Face UE.

Sin embargo, en base a los otros valores p y dado que comparten el mismo tamaño de muestra, el orden del par de tecnologías desde el valor p más pequeño hasta el valor p más grande es Faceware y Link Face UE, seguidos por Facegood y Faceware.

6. Discusión y conclusiones

Este documento explica las diferencias entre las codificaciones faciales, presenta una comparación de algunas soluciones comerciales para el seguimiento facial en humanos digitales, expone un diseño experimentos empáticos que usan RV y compara Faceware, Live Link Face UE y Avatary, sistemas contemoráneos de captura facial.

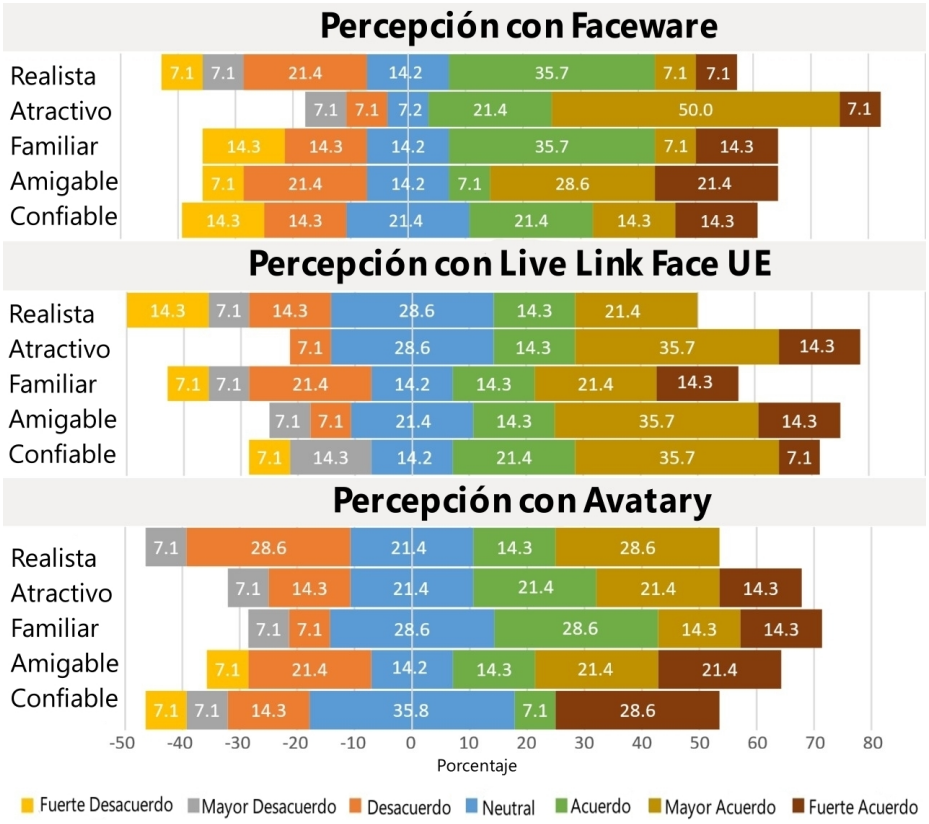


Fig. 3. Visualización de los resultados de la encuesta Big–Five representados en una escala Likert.

El objetivo de los experimentos es evaluar las capacidades de representación de emociones faciales de estas tecnologías a través de una prueba de percepción emFACS, y hasta qué punto las expresiones mostradas con estas tecnologías se alejan del Valle Inquietante, a través de la encuesta Big-Five.

De igual manera, durante cada paso del experimento se evaluaron las limitaciones y ventajas de las opciones disponibles. Las tres tecnologías muestran una representación confiable de emFACS Neutral, Felicidad y Sorpresa, con Avatary y Live Link Face UE, generalmente funcionando mejor que Faceware. Con respecto a los emFACS de Tristeza, Disgusto y Enojo, generalmente una tecnología muestra mejor desempeño que otra, sin embargo, ninguna de las soluciones pudo representar de manera confiable el emFACS de Miedo, lo que muestra un área de oportunidad enfocada en la representación realista del Miedo.

El análisis estadístico de los resultados no pudo encontrar una diferencia estadísticamente significativa entre las tecnologías o que alguna de ellas mostrara similitudes estadísticamente significativas con la percepción esperada de emFACS, por lo que una investigación adicional debe incluir un grupo más grande de sujetos para mostrar potenciales diferencias y similitudes estadísticamente significativas.

En comparación con otras tecnologías, Faceware presenta un realismo y atractivo superiores, Avatary presenta una familiaridad superior y Live Link Face UE presenta una amigabilidad y confiabilidad superiores. El análisis estadístico de los resultados de la encuesta Big–Five solo encontró una diferencia estadísticamente significativa entre Avatary y Live Link Face UE. por lo tanto, una investigación adicional debe incluir un grupo más grande de sujetos para mostrar potenciales diferencias estadísticamente significativas entre Faceware y Link Face UE o Avatary y Faceware.

Dado que la codificación FAP se utiliza en la mayoría de las soluciones existentes, esta investigación demuestra que las soluciones de última generación tienen un área de oportunidad relacionada con la investigación con otras codificaciones faciales que pueden aprovechar aún más las expresiones.

Además, Live Link Face UE representa una solución sencilla, asequible y democratizada para la investigación y el desarrollo con resultados estandarizados similares a las costosas herramientas de seguimiento facial. Las ventajas de usar opciones democratizadas pueden abrir nuevas direcciones para la investigación y la innovación en el campo, pero aún existe la necesidad de mejorar la respuesta empática y el desempeño de las expresiones faciales.

Estas necesidades pueden conducir a una investigación predominante sobre opciones democráticas de codificaciones más expresivas, soluciones novedosas para realizar la captura de movimiento de personas, y bases de datos con humanos digitales. Finalmente, el enfoque de evaluación presentado, podría usarse para evaluar futuras soluciones de seguimiento facial en términos de percepción.

Agradecimientos. Este trabajo fue apoyado a través de una beca para Carlos Vilchis por parte del Consejo Nacional de Ciencia y Tecnología de México (CONACYT). Este trabajo también fue apoyado por el programa Epic MegaGrants bajo el nombre de Grant FACS DEEP LEARNING TOOL.

Referencias

1. Amini, R., Lisetti, C., Ruiz, G.: Hapfacs 3.0: Facs-based facial expression generator for 3d speaking virtual characters. IEEE Transactions on Affective Computing, vol. 6, no. 4, pp. 348–360 (2015) doi: 10.1109/TAFFC.2015.2432794
2. Bausch, P.: Faceware website. Faceware, (2021)
3. Bennett, G., Kruse, J.: Teaching visual storytelling for virtual production pipelines incorporating motion capture and visual effects. In: Special Interest Group on Computer Graphics and Interactive Techniques Asia Symposium on Education (2015) doi: 10.1145/2818498.2818516
4. Briggs, C.: An essential introduction to maya character rigging. Chemical Rubber Company Press (2021)
5. Burke, B., Davis, M., Dawson, P.: Hype cycle for emerging technologies. Gartner Research (2021)
6. Caballar, R. D.: Are digital humans the next step in human-computer interaction? Spectrum IEEE (2021)
7. Cassell, J.: Embodied conversational interface agents. Communications of the Association for Computing Machinery, vol. 43, no. 4, pp. 70–78 (2000)

8. Costa, P., McCrae, R.: A five-factor theory of personality. The Five-Factor Model of Personality: Theoretical Perspectives, vol. 2, pp. 51–87 (1999)
9. Dzedzickis, A., Kaklauskas, A., Bucinskas, V.: Human emotion recognition: Review of sensors and methods. Sensors, vol. 20, no. 3, pp. 592 (2020)
10. Ekman, P., Rosenberg, E. L.: What the face reveals: Basic and applied studies of spontaneous expression using the facial action coding system, Oxford University Press (2005)
11. Epic Games: Epic games metahuman creator. Metahuman Unreal Engine (2021)
12. FACEGOOD Co. Ltd.: Avatary by facegood (2021)
13. Friesen, W. V., Ekman, P.: Emfacs-7: Emotional facial action coding system. University of California at San Francisco, vol. 2, no. 36, pp. 1 (1983)
14. Games, I. E.: Live link face for UE (2021)
15. Gavrilescu, M.: Recognizing emotions from videos by studying facial expressions, body postures and hand gestures. In: 23rd Telecommunications Forum Telfor, pp. 720–723 (2015) doi: 10.1109/TELFOR.2015.7377568
16. Higgins, D., Egan, D., Fribourg, R., Cowan, B., McDonnell, R.: Ascending from the valley: Can state-of-the-art photorealism avoid the uncanny? In: Association for Computing Machinery Symposium on Applied Perception 2021 (2021) doi: 10.1145/3474451.3476242
17. Ho, C. C., MacDorman, K. F.: Measuring the uncanny valley effect. International Journal of Social Robotics, vol. 9, pp. 129–139 (2017) doi: 10.1007/s12369-016-0380-9
18. Hyde, J., Carter, E. J., Kiesler, S., Hodgins, J. K.: Using an interactive avatar's facial expressiveness to increase persuasiveness and socialness. In: Proceedings of the 33rd Annual Association for Computing Machinery Conference on Human Factors in Computing Systems, pp. 1719–1728 (2015)
19. Ko, H., Kim, J. H., Kim, J.: Searching for facial expression by genetic algorithm. In: Virtual Environments '95, pp. 87–98 (1995) doi: 10.1007/978-3-7091-9433-1_8
20. Malatesta, L., Raouzaiou, A., Karpouzis, K., Kollias, S.: MPEG-4 facial expression synthesis. Personal and Ubiquitous Computing, vol. 13, pp. 77–83 (2009) doi: 10.1007/s00779-007-0164-1
21. McDonnell, R., Breidt, M., Bulthoff, H.: Render me real?: Investigating the effect of render style on the perception of animated virtual humans. Association for Computing Machinery, vol. 31, pp. 1–91 (2012)
22. Mori, M., MacDorman, K. F., Kageki, N.: The uncanny valley [from the field]. IEEE Robotics & Automation Magazine, vol. 19, no. 2, pp. 98–100 (2012)
23. Pandzic, I. S., Forchheimer, R.: MPEG-4 facial animation: The standard, implementation and applications. John Wiley and Sons, Inc (2003)
24. Seymour, M., Riemer, K., Kay, J.: Interactive realistic digital avatars-revisiting the uncanny valley. In: Proceedings of the 50th Hawaii International Conference on System Sciences, pp. 547–556 (2017)
25. Shan, C., Gong, S., McOwan, P. W.: Beyond facial expressions: Learning human emotion from body gestures. In: Proceedings of the British Machine Vision Conference, pp. 43–44 (2007) doi: 10.5244/C.21.43
26. Terzopoulos, D.: Modeling living systems for computer vision. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence, vol. 1, pp. 1003–1013 (1995)
27. Zibrek, K., Martin, S., McDonnell, R.: Is photorealism important for perception of expressive virtual humans in virtual reality? Association for Computing Machinery, vol. 16, no. 3 (2019) doi: 10.1145/3349609

ISSN 1870-4069

Segmentando imágenes gastrointestinales usando ensamble ponderado U-NET++ y 2D-HMM

Jairo Enrique Ramírez Sánchez[1], Pedro Martínez Barrón[2], Hannia Medina Aguilar[2], Romeo Sánchez Nigenda[2]

[1] Tecnológico de Monterrey, México

[2] Universidad Autónoma de Nuevo León, México

{alfonso.martinezbrrn,hannia.medinaglr, romeo.sanchezng}@uanl.edu.mx, A01750443@tec.mx

Resumen. Uno de los tratamientos más utilizados para el cáncer del tracto gastrointestinal (GI) es la radioterapia que requiere de la segmentación manual de los órganos afectados para el suministro de radiación sin afectar células sanas. Para automatizar este proceso, se han utilizado técnicas de aprendizaje profundo, especialmente variantes de U-Net. Sin embargo, la segmentación de los órganos del tracto GI permanece desafiante por la alta capacidad que tienen de deformarse por el movimiento corporal y la función respiratoria. Este trabajo propone una metodología que desarrolla un ensamble ponderado integrando modelos de U-Net++ y Modelos Ocultos de Markov (2D-HMM) para una segmentación semántica del estómago y los intestinos. Los experimentos reportan una precisión de 0.811 del coeficiente de Dice usando Leave-One-Out Cross-Validation, otorgando robustez a los resultados.

Palabras clave: Segmentación de imágenes, arquitectura U-NET, aprendizaje máquina, modelos ocultos de Markov.

Gastrointestinal Image Segmentation Using a Weighted U-NET++ and 2D-HMM Ensemble

Abstract. One of the most widely used treatments for cancer of the gastrointestinal tract is radiotherapy, which requires manual segmentation of the affected organs to deliver radiation without affecting healthy cells. To automate this process, deep learning techniques have been used, especially variants of U-Net. However, the segmentation of the GI tract organs remains challenging, due to their high capacity to deform due to body movement and respiratory function. This work proposes a methodology that develops a weighted ensemble integrating U-Net++ models and Hidden Markov Models (2D-HMM) for semantic segmentation of the stomach and intestines. From our experiments, we obtained a precision of 0.811 for Dice coefficient by means of Leave-One-Out Cross-Validation, which provides robustness to the results.

Keywords: Image segmentation, U-NET architecture, machine learning, hidden Markov models.

1. Introducción

A nivel mundial, en el 2018, alrededor de 4.8 millones de personas fueron diagnosticadas con cáncer del tracto gastrointestinal, representando el $26\,\%$ de la incidencia mundial de cáncer. Las proyecciones basadas en las tendencias actuales predicen un incremento del $58\,\%$ a 7.5 millones para 2040 [1]. La mitad de estos pacientes son elegibles para radioterapia [13].

Durante la radioterapia, un acelerador lineal médico (LINAC) administra altas dosis de radiación a las células cancerosas para matarlas, y con cierta probabilidad daña al mismo tiempo a las células sanas cercanas. El daño a las células sanas provoca efectos secundarios asociados con este tratamiento, como pérdida de audición, vómitos, y cansancio extremo, entre otros efectos [18]. Para disminuir los daños colaterales, los oncólogos tratan de dirigir los rayos X hacia los tumores evitando los órganos en riesgo.

El Acelerador Lineal Guiado por Imágenes de Resonancia Magnética (MR-Linac), permite observar tumores y órganos en tiempo real para ajustar la dirección de la radiación; sin embargo, los oncólogos deben segmentar manualmente los órganos, extendiendo las sesiones de tratamiento hasta en una hora, tiempo en el que el paciente debe permanecer inmóvil.

En los últimos años, técnicas de Inteligencia Artificial como las redes neuronales convolucionales han sido capaces de realizar auto-segmentación en casos de tumores cerebrales [6], cáncer de cuello [11] y cáncer de próstata [9, 8], reduciendo a la mitad el tiempo de las sesiones de tratamiento [3]; no obstante, hay pocos avances en la segmentación de los órganos del tracto gastrointestinal (GI), principalmente porque los órganos abdominales están rodeados de tejido blando y pueden variar en forma y ubicación a lo largo del día debido a los movimientos digestivos y respiratorios [10].

En este trabajo proponemos una metodología basada en aprendizaje profundo para el pre-procesamiento y segmentación de imágenes de resonancia magnética del tracto digestivo. La arquitectura de nuestro enfoque es un ensamble ponderado basado en modelos de U-Net y Modelos Ocultos de Markov en dos dimensiones (2D-HMM) que realiza una segmentación semántica del estómago y los intestinos delgado y grueso.

La metodología propuesta tiene el potencial de ayudar a implementar tratamientos más efectivos y eficientes para los pacientes al acelerar el proceso de segmentación.La metodología presentada se evaluó sobre el conjunto de imágenes de la UW-Madison Carbone Center, proporcionadas de manera pública en la plataforma Kaggle como parte de la competencia UW-Madison GI Tract Image Segmentation[3] sin poner en riesgo el tiempo de ejecución y los requerimientos de espacio en memoria del proceso de segmentación.

El presente trabajo está organizado de la siguiente manera. En la siguiente sección presentamos revisión de la literatura. La sección 3 describe la metodología propuesta ilustrando las diferentes etapas del proceso. La sección 4 discute los resultados obtenidos de los modelos generados. En la última sección presentamos nuestras conclusiones.

[3] kaggle.com/competitions/uw-madison-gi-tract-image-segmentation

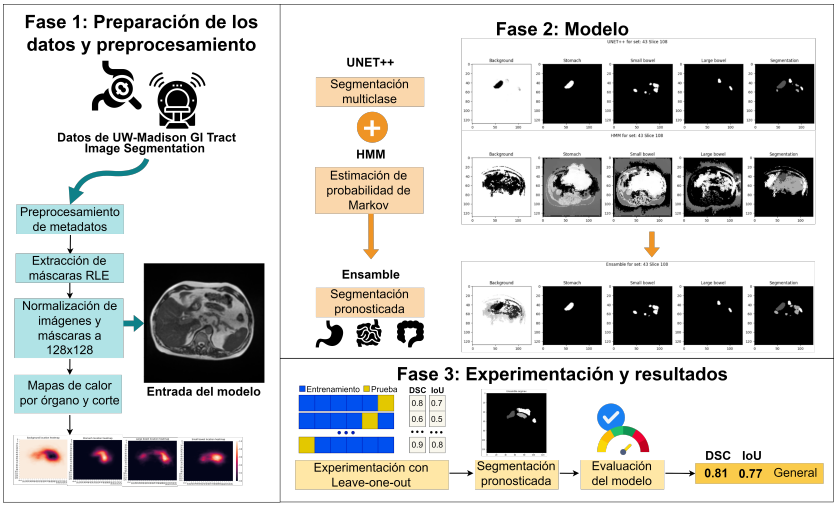


Fig. 1. Metodología para el Diseño y Validación de Modelos de Segmentación.

2. Trabajo relacionado

Estudios recientes en el área médica han empleado técnicas de aprendizaje profundo de Inteligencia Artificial para asistir en la segmentación automática de imágenes médicas en procesos de diagnóstico y tratamiento [16]. En particular, se han considerado variantes de arquitecturas U-Net para la segmentación de imágenes biomédicas.

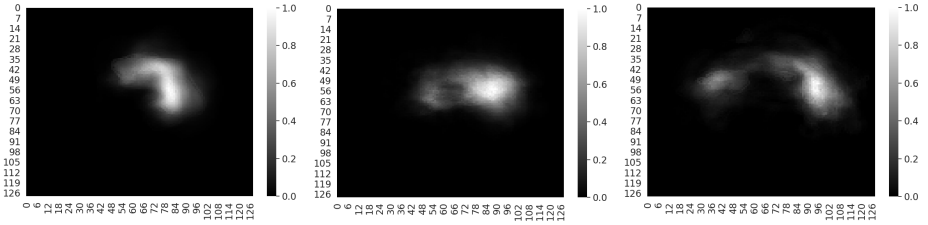
En este tipo de arquitecturas se consideran modelos de aprendizaje profundo con buen rendimiento en la auto-segmentación de imágenes debido a que tienen la capacidad de combinar simultáneamente información de alto y bajo nivel para extraer características complejas.

Sin embargo, la segmentación de los órganos del tracto GI sigue siendo una tarea desafiante [7], ya que estos órganos tienen una alta capacidad de deformarse por el movimiento corporal y la función respiratoria del individuo.

Debido a lo anterior, existen pocos estudios sobre el uso exitoso y amplio de MR-Linac para casos de cáncer de estómago [21], y sobre la aplicación de arquitecturas U-Net para éste tipo de imágenes, la mayoría de los estudios se basan en modelos complejos como 3D U-Net.

En 2020, [12] propusieron U-Net para segmentar el hígado, el estómago, el duodeno y el riñón en imágenes de tomografía computarizada (TC) basadas en parches 3D. Sus resultados fueron prometedores para el estómago, alcanzando un 0.813 para el Coeficiente de Similitud de Dice (DSC), pero menos significativos para el duodeno donde obtuvieron 0.595.

[19] propuso un enfoque similar para segmentar los órganos del tracto GI en el 2022. En un reporte preliminar compara el rendimiento de diferentes codificadores para una arquitectura U-Net clásica, siendo el codificador Resnet34 el que reporta mejores resultados.



(a) Estómago (b) Intestino Delgado (c) Intestino Grueso

Fig. 2. Mapas de calor para cada órgano.

(a) Estómago (b) Intestino Delgado (c) Intestino Grueso

Fig. 3. Matrices de transición.

El trabajo por [5] presenta una U-Net y métodos de Redes Neuronales Convolucionales por Regiones (Mask R-CNNs) para realizar la segmentación de los órganos del tracto GI en el mismo conjunto de datos de UW-Madison que utilizamos en este trabajo. Los autores reportan que su modelo Mask R-CNN obtuvo una puntuación del DSC de 0.73 en sus datos de validación.

Otros trabajos utilizan transformadores de visión (Vision Transformers) para segmentar, de igual manera, las imágenes de UW-Madison [15]. La solución que se propone es híbrida ya que utilizan en su modelo una arquitectura LeViT como el codificador, y una U-Net++ como el decodificador. Los resultados de su modelo alcanzan una puntuación de 0.79 para DSC y de 0.72 para IoU.

En [7], se describe un método de refinamiento automático de contornos (ACR) basado en mapas de probabilidad para corregir contornos auto-segmentados en radioterapia guiada por resonancia magnética. La auto-segmentación fue generada por una arquitectura CNN profunda en 3D (una 3D-ResUNet modificada), el DSC cambió de 0.44 a 0.56, de 0.33 a 0.55, y de 0.34 a 0.54, en el estómago, intestino delgado e intestino grueso, respectivamente.

[17] desarrollaron un método basado en Modelos Ocultos de Markov (HMM) llamado Complete Enumeration Propagation para la segmentación de imágenes multi-clase, en el cual los estados ocultos de un modelo Markoviano representan la verdadera segmentación de la imagen.

[2] utilizaron modelos de Markov bidimensionales (2D-HMM) para la segmentación efectiva de radiografías, imágenes multiespectrales y sintéticas.A pesar del potencial de HMMs, no existen estudios comprensivos de su aplicación en la segmentación de imágenes de resonancias magnéticas.

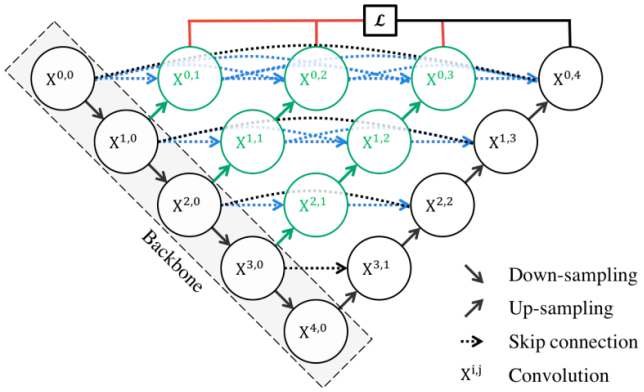


Fig. 4. Arquitectura original de U-Net++ de [22].

Existen trabajos recientes en la literatura que combinan el uso de operadores convolucionales con HMMs adaptativos para segmentar imágenes cerebrales [14, 20]. Sin embargo, hasta donde conocemos, no existe un método que incorpore HMMs en la segmentación de imágenes del tracto GI como proponemos en este trabajo.

En resumen, los enfoques de aprendizaje profundo, especialmente variantes de U-Net, son los métodos más explorados en la literatura para analizar imágenes biomédicas [16]. La aplicación de estos métodos para segmentar imágenes del tracto gastrointestinal sigue siendo un desafío y una área de investigación abierta.

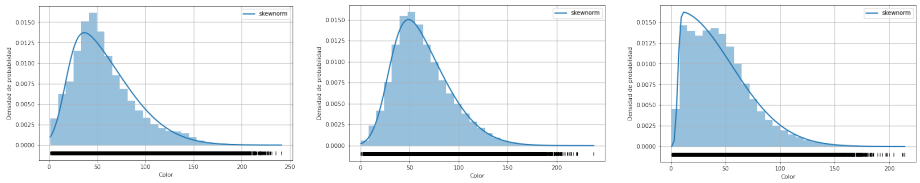
3. Metodología

En esta sección presentamos la metodología propuesta la cual consta de tres fases. La primera fase incluye pre-procesamiento de las imágenes del conjunto de datos (3.1), la segunda es el diseño y construcción de los modelos de segmentación (3.2), y por último, la fase de validación de los modelos a través de experimentación y análisis de resultados (4). En la Fig.1 se pueden observar las etapas generales de la metodología propuesta.

3.1. Pre-procesamiento de los datos

Como puede observarse en el diagrama de la Fig.1, la primera fase de la metodología consiste en la preparación de los datos. El conjunto de datos, utilizado en esta investigación es público y fue proporcionado por el UW-Madison Carbone Cancer Center.

El repositorio de datos consiste de 272 conjuntos de resonancias magnéticas en formato PNG en escala de grises de 16 bits de 85 pacientes con cáncer durante su tratamiento por radiación. Cada escaneo tiene 144 cortes, lo que da un total de 39,168 imágenes. Las anotaciones de entrenamiento son las máscaras codificadas por RLE (Run-Length Encoding) para la segmentación de tres órganos del tracto GI: el estómago, el intestino grueso y el intestino delgado.



(a) Estómago (b) Intestino Delgado (c) Intestino Grueso

Fig. 5. Distribuciones de probabilidad en función del color del píxel.

Algorithm 1 2D-HMM Algoritmo de Segmentación.

Calcular P_e, P_O y P_T
$\omega[128][128] : float$ ▷ Imagen a procesar
$M[128][128][4] : float$ ▷ Matriz de segmentación predicha
$S \leftarrow [0, 1, 2, 3]$ ▷ Clasificaciones de los estados
for $i, j \in \omega$ **do**
  for $l \in S$ **do**
    $O_{i,j} = \omega[i][j]$
    $P_{i,j} = P_O(l \| O_{i,j}) \cdot P_e(l \| i, j)$
    temp[4][4]: float
      for $m, n \in S$ **do**
        temp[m][n] = $P_{i,j} \cdot P_T(l \| m, n)$
      $M[i][j][l] = max(temp)$
return $\arg\max_l M$ ▷ Segmentación final

Las imágenes son de diferentes dimensiones por lo que fue necesario redimensionarlas a un tamaño estandarizado. En consecuencia, se decidió normalizar todas las imágenes y sus respectivas máscaras codificadas en RLE a un tamaño de 128 x 128 px. Para visualizar el patrón en la distribución de los órganos en la muestra, trazamos el mapa de calor para cada órgano que se presenta en Fig.2.

3.2. Diseño y construcción de los modelos de segmentación

Como se mencionó anteriormente, nuestra propuesta metodológica incluye la creación de 2 modelos para la segmentación de los órganos, y un ensamble que integra ambos modelos. El primer modelo se basa en una arquitectura de tipo U-Net++, y el segundo modelo está basado en HMMs bidimensionales (2D-HMM). A continuación se describen los procesos individuales para la construcción y entrenamiento de ambos modelos, así como el proceso de su integración para el ensamble.

Modelo U-Net++. Dicha arquitectura fue diseñada para resolver limitaciones del modelo U-Net base en la segmentación de imágenes médicas [22] al incluir una serie de conexiones adicionales a la U-Net original para la recuperación efectiva de los detalles de granularidad fina de los objetos, incluyendo supervisión profunda que permite establecer diferentes configuraciones de sus parámetros.

Tabla 1. Valores de Ponderación α de ensamble 2D-HMM . U-Net++ evaluados.

Métrica		**Ponderación** α				
		0.05	0.10	0.20	0.30	0.40
Dice	General	0.811	0.799	0.788	0.771	0.742
	Estómago	0.888	0.885	0.872	0.844	0.749
	Intestino Delgado	0.812	0.791	0.759	0.701	0.601
	Intestino Grueso	0.817	0.814	0.804	0.786	0.747
IoU		0.777	0.770	0.748	0.709	0.628

Tabla 2. Resultados de la experimentación con los modelos propuestos.

Métrica		**Modelos**			
		2D-HMM . U-Net++	2D-HMM . U-Net	U-Net++	U-Net
Dice	General	0.811 (32 %)	0.723 (34 %)	0.610	0.538
	Estómago	0.888 (10 %)	0.803 (26 %)	0.808	0.635
	Intestino Delgado	0.812 (38 %)	0.711 (29 %)	0.585	0.548
	Intestino Grueso	0.817 (5.6 %)	0.774 (43 %)	0.773	0.538
IoU		0.777 (18 %)	0.696 (36 %)	0.657	0.511

Las conexiones adicionales de la U-Net++ siguen una regla piramidal, donde la forma de **U** se llena con bloques convolucionales, cada uno de los cuales consta de un cierto número de capas que varía según los nodos de la red. El diagrama original de la U-Net++ tomado de [22] se muestra en la figura 4.

En este trabajo la red se implementó en Python 3.8 siguiendo la versión propuesta en [22]. Los hiper-parámetros del modelo fueron ajustados con la búsqueda por cuadrícula propia de la API de Keras, seleccionando relu como función de activación en las capas ocultas, 0.1 como dropout rate, $5x10^{-4}$ como learning rate durante 50 épocas y Adam como optimizador.

Finalmente, se utilizó sigmoid como función de activación en la última capa en lugar de softmax con la finalidad de asignar probabilidades a cada clase en lugar de distribuirlas. Para la determinación de los hiper-parámetros se realizó una partición del 80 % del total de imágenes para entrenamiento y 20 % para validación.

Como función de pérdida se optimizó el coeficiente de DICE por órgano, integrándose por medio de una suma ponderada debido al desbalanceo de clases. Sea $y \in \mathbb{R}^{128x128x4}$ la matriz de segmentación real, $\hat{y} \in \mathbb{R}^{128x128x4}$ la segmentación predicha por la red, sea $S \in \{$estómago, intestino delgado, intestino grueso, fondo$\}$ el conjunto de estados de la clasificación. Con lo cual $\hat{y}_l \in \mathbb{R}^{128x128}$ hace referencia a la segmentación correspondiente al órgano $l \in S$. Finalmente, sea α_l la frecuencia inversa de la clase del órgano l. La ecuación 1 muestra matemáticamente el proceso:

$$\mathcal{L}(y, \hat{y}) = \sum_{l \in S} \alpha_l \frac{2|y_l \cap \hat{y}_l|}{|y_l| + |\hat{y}_l|}. \tag{1}$$

Modelo bidimensional de Markov (2D-HMM). Los Modelos Ocultos de Markov son una técnica estadística que permite crear un modelo con eventos observados y ocultos como factores causales en un modelo probabilístico.

Tabla 3. Comparación con Modelos Recientes de Segmentación del Tracto GI Estos modelos utilizan datos diferentes a los usados por este trabajo en su evaluación.

Métrica		Modelos en la Literatura						
		2D-HMM U-Net U-Net++	U-Net [5]	Mask R-CNN[5]	Resnet34 [19]	LeViT384-UNet++ [15]	3D-ResUnet [7]*	3D U-Net [12]*
Dice	General	0.811	0.51	0.73			0.79	
	Estómago	0.888				0.813	0.77	0.813
	Int Delgado	0.812					0.75	
	Int Grueso	0.817					0.76	
IoU		0.777			0.852		0.728	

Un HMM consta de dos procesos estocásticos, un proceso invisible de estados ocultos y un proceso visible de símbolos observables, donde los estados ocultos forman una cadena de Markov y la distribución de probabilidad del símbolo observado depende de los estados subyacentes.

En el caso de segmentación de imágenes, la intuición es que los píxeles en una imagen presentan dependencia de aquellos aledaños; es decir, comparten características en común como color y ubicación espacial. Debido a esto es posible tratar a esta dependencia como un Markov Random Field, el cual relaciona dos probabilidades principales: de transición (P_T) y de observación (P_O). La intuición indica que los píxeles (i,j) de una imagen se relacionan con sus vecinos.

La probabilidad de transición P_T indica que el estado s al que pertenece un píxel, expresado como $s_{i,j}$, está relacionado con el estado del píxel lateral izquierdo $s_{i-1,j}$ y el superior $s_{i,j-1}$. Es decir, $P_T = P(s_{i,j} = l \| s_{i-1,j} = n, s_{i,j-1} = m)$. Sea S el conjunto de estados en los que un píxel puede ser clasificado y sea Ω el conjunto total de imágenes. El cálculo de P_T se expresa en la ecuación 2, siendo $I(\cdot)$ la función contadora que retorna 1 si se cumple la condición y 0 en caso contrario:

$$P_T(l\|n,m) = \frac{1}{\|\Omega\|} \sum_{\omega \in \Omega} \frac{\sum_{i,j} I(s_{i,j} = l, s_{i-1,j} = n, s_{i,j-1} = m)}{\sum_{i,j} I(s_{i,j} = l)} \; \forall l, m, n \in S. \quad (2)$$

Las probabilidades de transición calculadas se muestran en las matrices de la Fig.3. Por ejemplo, podemos observar en la Fig.**??** que si el estado del píxel actual correspondiera al estómago existiría una probabilidad del 4.5 % que el píxel superior fuese el estómago y el izquierdo el fondo de la imagen; mientras que existe una probabilidad de un 87 % que ambos correspondan al mismo órgano.

Por otro lado, la probabilidad de que los píxeles aledaños correspondan a los otros órganos es prácticamente nula. La probabilidad de observación es calculada con base al color de cada píxel ($O_{i,j}$) medido entre 0 y 255. Se realizó un ajuste de una función de distribución de probabilidad skewnorm a los colores de cada uno de los órganos.

Sea $P_O(s_{i,j} = l \| O_{i,j})$ la función que toma como entradas el color del píxel i, j y regresa la probabilidad de que pertenezca al estado $l \in S$. Dichas funciones se muestran gráficamente en la figura 5. Como se puede observar, realizar una estimación por máxima verosimilitud sería poco precisa debido a que las probabilidades de observación para el estómago y el intestino delgado son similares.

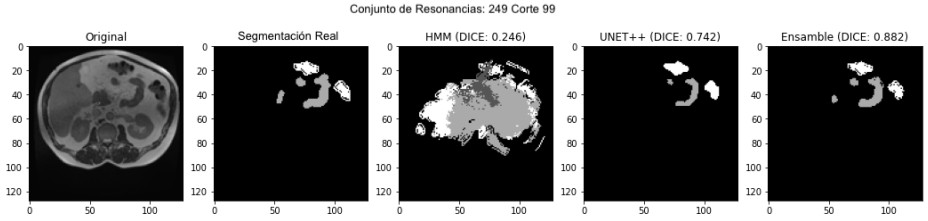


Fig. 6. Ejemplo de segmentación en conjunto 249 para un determinado corte.

Es necesario integrar más información. En el trabajo actual, el estado de un píxel además de estar condicionado por las probabilidades anteriormente descritas, también lo está por la posición espacial en la imagen; es decir, existen zonas de alta probabilidad en las que un órgano puede aparecer como se muestra en los mapas de calor de la Fig.2. Así, la probabilidad P_e de que un píxel (i,j) pertenezca a un estado $s_{i,j} = l$ se calcula integrando las probabilidades antes descritas como se muestra en la ecuación 3:

$$P_e(l \| i, j) = \frac{1}{\|\Omega\|} \sum_{\omega \in \Omega} I(s_{i,j} = l). \tag{3}$$

En consecuencia, el cálculo final de la probabilidad de que el estado del píxel (i, j) sea l se muestra en la ecuación 4. Naturalmente, P_e y P_T son calculadas con anterioridad durante la fase de entreno y almacenadas para su consulta. En el caso de P_O se guardan los parámetros de las distribuciones y se calcula su valor:

$$P(s_{i,j} = l) = P_T(l||n, m) \cdot P_O(l||O_{i,j}) \cdot P_e(l||i, j). \tag{4}$$

El cálculo propuesto toma en cuenta el sentido espacial, observacional y de transición. La forma de incorporar los cálculos para la segmentación de una nueva imagen se muestra en el algoritmo 1.

En el presente trabajo, la multiplicación de probabilidades fue sustituida por la suma logarítmica de la probabilidad para evitar un problema de desbordamiento negativo o underflow. Es importante destacar que el propósito del 2D-HMM no es la segmentación por si misma, si no el cálculo de probabilidades eficientes para mejorar el rendimiento de U-Net++. Por este motivo se omite incluir una adaptación del algoritmo de Viterbi en 2D.

Ensamble. Con el propósito de integrar la información de los modelos U-Net++ y 2D-HMM, se propone una capa de ensamble ponderada la cual utiliza las probabilidades otorgadas por ambos modelos para potenciar la clasificación. Sean $H, U \in \mathbb{R}^{128x128x4}$ las matrices de probabilidades calculadas por 2D-HMM y U-Net++ respectivamente. La integración ponderada se muestra por la ecuación 5:

$$E(H, U) = \alpha U + (1 - \alpha)H. \tag{5}$$

El valor de α es seleccionado del espacio de búsqueda $\{0.05, 0.10, 0.20, 0.30, 0.40\}$ que se determinó experimentalmente.

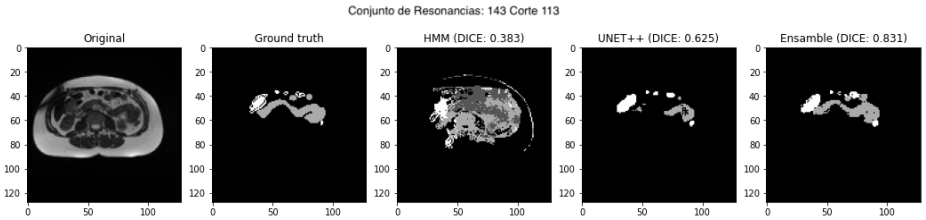


Fig. 7. Ejemplo de segmentación en conjunto 143 para un determinado corte.

Fig. 8. Ejemplo de segmentación 3D para conjunto 249.

Los resultados de las métricas para cada valor de α se muestran en la siguiente sección. En la segmentación final se aplica máxima verosimilitud sobre el ensamble para obtener el estado l de mayor probabilidad en cada píxel en la matriz.

4. Resultados y análisis

La experimentación se llevó a cabo en la plataforma Colab de Google utilizando una Colabnotebook con Intel(R) Xeon(R) CPU @ 2.20GHz de 6 núcleos, GPU NVIDIA A100-SXM y 12 GB de memoria RAM. Para la fase de entrenamiento y validación de los modelos se siguió el método de Leave-One-Out Cross-Validation que es uno de los métodos recomendados en ciencias biomédicas para mejorar la tasa predictiva de modelos para estudios clínicos [4].El método consiste en realizar las pruebas del modelo sobre un conjunto de imágenes ω y entrenar tanto los parámetros del 2D-HMM como la U-Net++ con $\Omega \setminus \omega$.

Con esto, el entrenamiento consiste de 271 conjuntos de 39,024 imágenes en total y el de prueba de 1 conjunto de 144 imágenes. Una de las primeras tareas de la fase de evaluación fue ajustar el parámetro de ponderación α del ensamble propuesto, los resultados se muestran en la tabla 1. El caso $\alpha = 0$ haría referencia a la U-Net++ y $\alpha = 1$ al 2D-HMM. Podemos observar que los mejores resultados para el ensamble en las métricas de evaluación se obtienen cuando $\alpha = 0.05$.

Podemos observar en la Tabla2 los resultados de todos los modelos propuestos. Es de resaltar que los modelos U-Net, que incorporan la información del proceso de Markov, reportan mejores resultados en ambas métricas de evaluación satisfaciendo nuestra intuición para su integración.

Fig. 9. Ejemplo de segmentación 3D para conjunto 143.

Por ejemplo, en el Dice general para el ensamble de la U-Net++ se tiene un porcentaje de mejora del 32 %, mientras que el ensamble para la U-Net obtiene una mejora del 34 % con respecto a sus modelos individuales.

En el caso del IoU el porcentaje de mejora es del 18 % para el ensamble de la U-Net++ y del 36 % para la U-Net. Finalmente, la Tabla3 compara los resultados del modelo propuesto 2D-HMM U-Net++ con trabajos recientes en la literatura en el problema de segmentación de imágenes biomédicas del tracto GI. Como lo mencionamos en la sección 1, existen pocos avances en la segmentación de órganos GI debido a su estructura fisiológica.

Podemos observar que nuestro enfoque supera a la mayoría de los trabajos en las métricas de evaluación, a excepción del modelo Resnet34 que solo reporta resultados respecto al IoU; sin embargo, este trabajo siguió una metodología tradicional de particiones 80 - 20 para evaluar, lo cual puede hacer que el resultado sea altamente dependiente de la partición utilizada.

En las figuras 6 - 8 y 7 - 9 se muestran ejemplos visuales 2D y 3D de la segmentación para un corte específico de un conjunto de resonancias. En estos ejemplos el ensamble permite potenciar las predicciones de la U-Net++ hasta en un 19 %. Por ejemplo, la predicción de la U-Net++, ilustrada por la cuarta imagen de la Fig.7, pierde múltiples detalles de los órganos si se compara con la imagen verdadera.

Sin embargo, el ensamble ponderado es capaz de restaurar estos detalles, como puede observarse en la última columna de la misma figura. En general, podemos observar como el ensamble propuesto incrementa notablemente la calidad de la segmentación.

En resumen, aunque el modelo U-Net++ ha demostrado ser una arquitectura muy eficiente para la segmentación de órganos, como lo demuestra la revisión de la literatura, ésta presenta deficiencias para segmentar ciertas secciones del tracto GI al contener dos o más clases de órganos con una alta probabilidad.

Este trabajo integra las probabilidades del modelo oculto de Markov para discernir mejor aquellos casos donde el modelo base no logra segmentar correctamente, ya que toma en cuenta las probabilidades espaciales y de transición, constituyendo la principal diferencia del trabajo propuesto respecto al trabajo relacionado.

5. Conclusiones

La segmentación de órganos para el tratamiento del cáncer del tracto gastrointestinal es una labor de alta importancia que necesita precisión y rapidez. Es vital contar con algoritmos que puedan auxiliar en la automatización del proceso de segmentación, como apoyo para los especialistas médicos, con el fin de disminuir daños colaterales a células sanas sin incrementar los tiempos de tratamiento.

Sin embargo, la segmentación de los órganos del tracto GI sigue siendo una tarea desafiante, debido a las deformaciones que sufren por el movimiento corporal y la función respiratoria. En este trabajo se propone una metodología, basada en aprendizaje profundo, que desarrolla un ensamble ponderado integrando modelos de U-Net++ y 2D-HMM para una segmentación semántica del estómago y los intestinos. A pesar que el 2D-HMM por sí mismo no otorga una segmentación con gran precisión, permite potenciar las predicciones de la U-Net++ hasta en un 32 % en el Dice general, y un 18 % en el IoU.

La precisión final de 0.811, obtenida por el ensamble en el Dice general, es mejor a los resultados reportados en la literatura. Además, al utilizar el método de evaluación Leave-One-Out, la métrica proporcionada cuenta con un alto nivel de confiabilidad sobre el dataset utilizado. La arquitectura propuesta tiene el potencial de ayudar a implementar tratamientos más efectivos y eficientes para los pacientes con cáncer al acelerar el proceso de segmentación y minimizar los riesgos.

Parte del trabajo futuro consistirá en la integración de técnicas de refinamiento automático de contornos, la cuales creemos podrían mejorar la calidad otorgada por las probabilidades espaciales y de transición del ensamble propuesto. Aunado a lo anterior, se replicará la metodología a otros datasets para determinar su generalización.

Referencias

1. Arnold, M., Abnet, C. C., Neale, R. E., Vignat, J., Giovannucci, E. L., McGlynn, K. A., Bray, F.: Global burden of 5 major types of gastrointestinal cancer. Gastroenterology, vol. 159, no. 1, pp. 335–349 (2020) doi: 10.1053/j.gastro.2020.02.068
2. Baumgartner, J., Flesia, A., Gimenez, J., Pucheta, J.: A new approach to image segmentation with two-dimensional hidden Markov models. In: Congress on Computational Intelligence and 11th Brazilian Congress on Computational Intelligence, pp. 213–222 (2013) doi: 10.1109/BRICS-CCI-CBIC.2013.43
3. Bertelsen, A., Schytte, T., Moller, P.: First clinical experiences with a high field 1.5 T Mr linac. Acta Oncologica, vol. 58, no. 10, pp. 1352–1357 (2019) doi: 10.1080/0284186X.2019.1627417
4. Chicco, D., Jurman, G.: The ABC recommendations for validation of supervised machine learning results in biomedical sciences. Frontiers in Big Data, vol. 5 (2022) doi: 10.3389/fdata.2022.979465
5. Chou, A., Li, W., Roman, E.: Gi tract image segmentation with U-Net and mask R-CNN. Technical Report 164, Stanford University (2022)
6. Despotovic, I., Goossens, B., Philips, W.: MRI segmentation of the human brain: Challenges, methods, and applications. Computational and Mathematical Methods in Medicine, vol. 2015, pp. 1–23 (2015) doi: 10.1155/2015/450341

7. Ding, J., Zhang, Y., Amjad, A., Xu, J., Thill, D., Li, X. A.: Automatic contour refinement for deep learning auto-segmentation of complex organs in MRI-guided adaptive radiation therapy. Advances in Radiation Oncology, vol. 7, no. 5, pp. 100986 (2022) doi: 10.1016/j.adro.2022.100968
8. Eppenhof, K., Maspero, M., Savenije, M., de Boer, J., van der Voort van Zyp, J., Raaymakers, B., Raaijmakers, A., Veta, M., van den Berg, C., Pluim, J.: Fast contour propagation for MR-guided prostate radiotherapy using convolutional neural networks. Medical Physics, vol. 47, no. 3, pp. 1238–1248 (2020) doi: 10.1002/mp.13994
9. Fransson, S., Tilly, D., Strand, R.: Patient specific deep learning based segmentation for magnetic resonance guided prostate radiotherapy. Physics and Imaging in Radiation Oncology, vol. 23, pp. 38–42 (2022) doi: 10.1016/j.phro.2022.06.001
10. Johansson, A., Balter, J. M., Cao, Y.: Gastrointestinal 4D MRI with respiratory motion correction. Medical Physics, vol. 48, no. 5, pp. 2521–2527 (2021) doi: 10.1002/mp.14786
11. Kawahara, D., Tsuneda, M., Ozawa, S., Okamoto, H., Nakamura, M., Nishio, T., Nagata, Y.: Deep learning-based auto segmentation using generative adversarial network on magnetic resonance images obtained for head and neck cancer patients. Journal of Applied Clinical Medical Physics, vol. 23, no. 5, pp. e13579 (2022) doi: 10.1002/acm2.13579
12. Kim, H., Jung, J., Kim, J., B., C., Kwak, J., Yun, J., Lee, S., Lee, J., Yoon, S.: Abdominal multi-organ auto-segmentation using 3D-patch-based deep convolutional neural network. Scientific Reports, vol. 10, pp. 6204 (2020) doi: 10.1038/s41598-020-63285-0
13. Lee, S. L., Li, Y., Meudt, J. J., Strang, J., Hebel, D., Alfson, A., Olson, S. J., Kruser, T. R., Smilowitz, J. B., Borchert, K., Loritz, B., Bayouth, J., Bassetti, M.: UW-Madison GI tract image segmentation (2022)
14. Li, G., Sun, J., Song, Y.: Segmentation of medical images with a combination of convolutional operators and adaptive hidden Markov model. In: IEEE 5th International Conference on Computer and Communications, pp. 1782–1786 (2019) doi: 10.1109/ICCC47050.2019.9064034
15. Nemani, P., Vollala, S.: Medical image segmentation using LeViT-UNet++: A case study on GI tract data. In: 26th International Computer Science and Engineering Conference (2022) doi: 10.1109/ICSEC56337.2022.10049343
16. Punn, N. S., Agarwal, S.: Modality specific U-Net variants for biomedical image segmentation: A survey. Artificial Intelligence Review, vol. 55, no. 7, pp. 5845–5889 (2022) doi: 10.1007/s10462-022-10152-1
17. Pyun, K., Lim, J., Won, C. S., Gray, R. M.: Image segmentation using hidden Markov Gauss mixture models. IEEE Transactions on Image Processing, vol. 16, no. 7, pp. 1902–1911 (2007) doi: 10.1109/TIP.2007.899612
18. Schaue, D., McBride, W. H.: Opportunities and challenges of radiotherapy for treating cancer. Nature Reviews Clinical Oncology, vol. 12, no. 9, pp. 527–540 (2015) doi: 10.1038/nrclinonc.2015.120
19. Sharma, M.: Automated GI tract segmentation using deep learning (2022) doi: 10.48550/ARXIV.2206.11048
20. Song, Y., Adobah, B., Qu, J., Liu, C.: Segmentation of ordinary images and medical images with an adaptive hidden Markov model and viterbi algorithm. Current Signal Transduction Therapy, vol. 15, no. 2, pp. 109–123 (2020) doi: 10.2174/1574362413666181109113834
21. Song, Y., Li, Z., Wang, H., Zhang, Y., Yue, J.: MR-LINAC-guided adaptive radiotherapy for gastric MALT: Two case reports and a literature review. Radiation, vol. 2, no. 3, pp. 259–267 (2022) doi: 10.3390/radiation2030019
22. Zhou, Z., Rahman-Siddiquee, M. M., Tajbakhsh, N., Liang, J.: UNet++: A nested U-Net architecture for medical image segmentation. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, pp. 3–11 (2018)

ISSN 1870-4069

Detección de enfermedades en cultivos de yuca a través de CNNs

David Hiram Vázquez Santana

Instituto Politécnico Nacional,
Centro de Investigación en Computación,
México

dvazquezs2019@cic.ipn.mx

Resumen. Gracias a su gran aportación de carbohidratos, la yuca es de vital importancia para la seguridad alimentaria y ha sido clasificado como el cuarto cultivo básico más importante para casi mil millones de personas. Se estima que la producción mundial asciende a más de 302 millones de toneladas anuales, siendo África el mayor productor con una cuota cercana al 63 % de la producción mundial. Además, es una fuente importante de carbohidratos para más de 200 millones de personas en el África subsahariana y en los últimos años su uso se ha extendido a la producción de pienso y otros fines industriales. La yuca, como cualquier otro cultivo, no está a salvo de sufrir enfermedades y debido a su relevancia, la detección de estas es de gran importancia para mejorar el rendimiento de las plantaciones. Los métodos más usados en la actualidad son intensivos y requieren grandes cantidades de dinero, tiempo y mano de obra especializada, por lo que en este trabajo se propone un método barato y que no requiere de especialistas agrícolas para la detección de cuatro de las enfermedades más comunes en esta planta mediante la clasificación de imágenes a través de tres algoritmos de clasificación y utilizando un banco de datos con imágenes obtenidas por agricultores y etiquetadas por expertos agrónomos, alcanzando valores F1 de 63.60, 73.97, 70.82, 85.77 y 76.22 para las clases CBB, CBSD, CGM, CMD y sana respectivamente.

Palabras clave: Yuca, deep learning, visión artificial, detección de enfermedades en cultivos.

Cassava Crop Diseases Recognition Using CNNs

Abstract. Thanks to its high carbohydrate content, cassava is vital for food security and has been ranked as the fourth most important staple crop for almost one billion people. World production is estimated at more than 302 million tons per year, being Africa the largest producer with an estimated 63% quota of the world's production. In addition, it is an important source of carbohydrates for more than 200 million people in sub-Saharan Africa and in recent years its use has been extended to animal feed production and other industrial purposes. Cassava, like any other crop, is not immune to diseases and due to its relevance, the control of these diseases is of great importance to improve the productivity of

the plantations. Currently, the most used methods are intensive and require large amounts of money, time and expert work force. Hence, in this work we propose a cheap method that does not require agro-specialists for the detection of four of the most common diseases in this plant by classifying images through three algorithms and using a dataset with images obtained by farmers and labeled by agro-specialists, reaching F1 values of 63.60, 73.97, 70.82, 85.77 and 76.22 for CBB, CBSD, CGM, CMD and healthy classes, respectively.

Keywords: Cassava, deep learning, machine vision, crop disease detection.

1. Introducción

Gracias a la facilidad de la yuca para adaptarse a diferentes tipos de suelos y climas, esta es cultivada en varios continentes, incluyendo África, América y Asia. El cultivo es de especial importancia en África, donde es una fuente importante de carbohidratos para más de 200 millones de personas en la región subsahariana.

Además, el continente es responsable de alrededor del 63 % de la producción mundial [13]. La yuca no solo es importante para garantizar la seguridad alimentaria. También es importante para la industria, se utiliza principalmente en la producción de papel, pienso y almidón [2, 19].

Además, gracias a su alta aportación de carbohidratos, este cultivo ha sido clasificado como el cuarto más importante para casi mil millones de personas [9]. Debido a su alto valor nutricional y económico, la producción de yuca se ha convertido en un tema de interés para los países productores.

Es por eso que el rendimiento de los cultivos es de suma importancia. Existen diversos factores que afectan el rendimiento de los cultivos. Estos pueden agruparse en tres grandes grupos: el primero está compuesto por los factores meteorológicos, como heladas, lluvias torrenciales, sequias, entre otros factores; el segundo grupo está compuesto por los factores relacionados al manejo de los cultivos y del suelo.

Finalmente, el tercer grupo está relacionado con el manejo de enfermedades. No es común que se realice el diagnóstico de enfermedades foliares debido a que, en la mayoría de los casos, es necesario realizarlo de forma manual por expertos agrónomos, lo que lo convierte en un método de diagnóstico costoso e ineficiente, por lo que se vuelve necesario desarrollar métodos más precisos para reducir los costos asociados a la detección de enfermedades y lograr que los agricultores estén dispuestos a implementarlos en sus plantaciones.

En la actualidad, existen diversos trabajos que utilizan una combinación de técnicas de visión artificial, procesamiento de imágenes y Machine Learning (ML). Por ejemplo, Meunkaewjinda A. et al. [10] propusieron un sistema capaz de identificar si una vid tiene sarna, roya o si se encuentra libre de enfermedades a través de la segmentación de las hojas y su posterior clasificación a través de una máquina de soporte vectorial (SVM).

Tal como lo muestra la revisión del estado del arte realizada por Lavika y Jyoti [6], las SVM son una de las técnicas más ampliamente utilizadas en los métodos de detección de enfermedades en plantas para llevar a cabo la fase de clasificación y algunas veces se combina con otras técnicas ML.

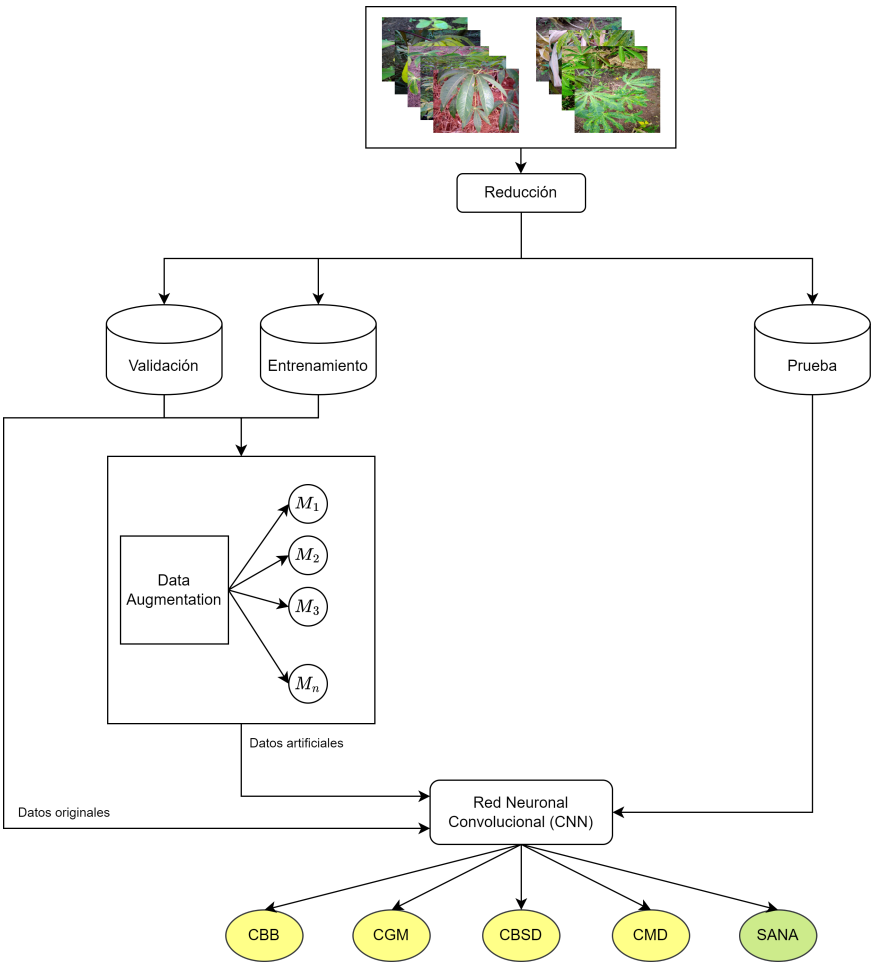


Fig. 1. Diagrama general del método propuesto.

Tal es el caso del trabajo presentado por Santosh y Manish [8], quienes desarrollaron un método de detección de enfermedades basado en la segmentación de hojas a través de clustering y su clasificación a usando un sistema hibrido compuesto por Random Forest (RF) y una SVM.

Uno de los principales retos al utilizar técnicas ML es la selección y extracción de características. De acuerdo con una revisión sistemática referente a la detección de enfermedades en plantas, las características que más frecuentemente se extraen son la forma, la textura y el color y las especies más estudiadas son el maíz, las papas, la soya y el algodón [6].

Esta investigación se enfoca en mejorar el rendimiento de los cultivos a través de la detección de enfermedades foliares utilizando técnicas de Deep Learning (DL), como las redes neuronales convolucionales (CNNs), las cuales permiten la detección y extracción automática de características importantes.

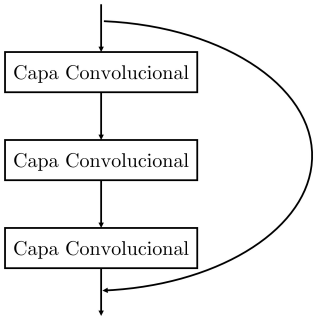


Fig. 2. Bloque residual.

Algunas de las enfermedades que más afectan la producción son cassava green mite (CGM), cassava bacterial blight (CBB), cassava brown streak disease (CBSD), cassava mosaic disease (CMD), cassava American latent leaf disease (CALD), cassava brown streak Uganda disease (CBSUD) y cassava Colombian symptomless disease (CCSD) [14].

2. Materiales y métodos

En esta sección se presenta la estructura general del modelo propuesto, los detalles de las CNNs seleccionadas y el banco de datos utilizado. El modelo propuesto utiliza una red neuronal convolucional como núcleo de funcionamiento. Se decidió no realizar preprocesamiento a las imágenes y dejar que la CNN se encargue de la extracción de características principales. En la Figura 1, se muestra el diagrama general del modelo presentado.

El banco de datos fue dividido en tres conjuntos: entrenamiento, validación y prueba. Posteriormente, con el objetivo de mejorar la generalización obtenida por la CNN, se crean instancias artificiales a partir de instancias reales. Después, la CNN es entrenada utilizando los patrones originales y sintéticos. Finalmente, el conjunto de prueba es presentado al clasificador y se mide el desempeño obtenido por el modelo.

2.1. Arquitecturas

Las redes neuronales convolucionales son un tipo particular de redes neuronales artificiales y su funcionamiento se basa en una operación conocida como convolución, la cual permite reducir el número de parámetros. Las CNNs combinan el poder de un extractor automático de rasgos y la capacidad de clasificación de un perceptrón multicapa [1].

La imagen de entrada es considerada como una matriz de tamaño MxN y se representa como $W_{m,n}$. Posteriormente, se aplica la convolución a la entrada utilizando un kernel $k_{p,q}$ de tamaño PxQ. A continuación, se describen las arquitecturas seleccionadas.

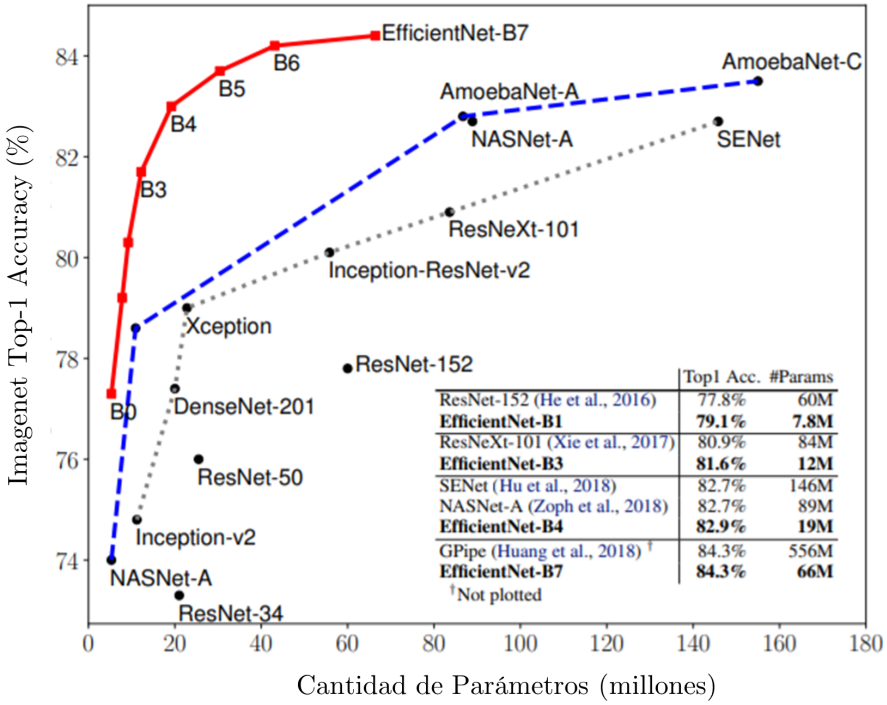


Fig. 3. Comparación de EfficientNets con respecto a otros modelos.

ResNet-50: ResNet-50 es un modelo que pertenece a la familia de redes ResNet [7], que es el nombre corto de red residual (residual network). Esta familia fue creada para solucionar el problema del gradiente de desaparición, el cual impide que las redes profundas actualicen sus pesos [4].

Esta CNN, como todas las de su familia, está formada por bloques residuales, los cuales se pueden entender como atajos para saltar 2 o 3 capas y evitar que estas se entrenen. ResNet-50 está formada por 15 bloques residuales como el que se muestra en la figura 2.

En cuestión de capas, está compuesta por 50 capas: 1 capa de entrada, 46 capas convolucionales, 2 capas de pooling y una capa densa con 1,000 neuronas a la salida. Esta es una de las redes neuronales convolucionales más ampliamente usada [5, 12, 16], razón por la cual se decidió implementarla en el modelo presentado en el presente artículo.

EfficientNet-B4: Esta arquitectura pertenece a la familia de redes neuronales convolucionales EfficientNets, la cual fue desarrollada por el equipo Google Brain y presentada en el artículo EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks [17]. La arquitectura de estas redes se basa en el escalamiento controlado a través de un método conocido como compound scaling y funciona escalando tres atributos al mismo tiempo usando un coeficiente (ϕ): profundidad, ancho y resolución. En la ecuación 1 se muestra este método:

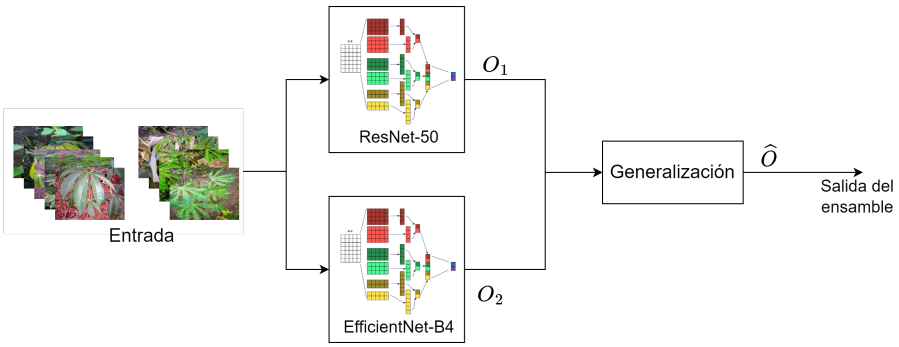


Fig. 4. Ensamble Stacking utilizando ResNet-50 y EfficientNet-B4.

(a) CBB (b) CBSD (c) CGM (d) CMD (e) Sana

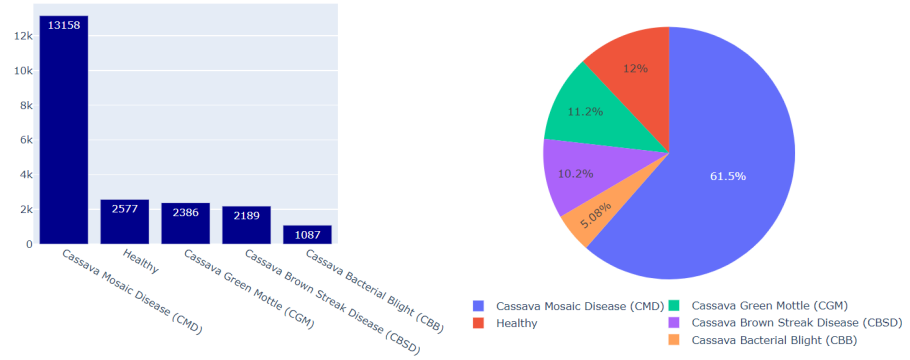
Fig. 5. Hojas de yuca con CBB, CBSD, CGM, CMD y sin enfermedades.

$$\begin{aligned} \text{profundidad} : d &= \alpha^{\phi}, \\ \text{ancho} : w &= \beta^{\phi}, \\ \text{resolución} : r &= \gamma^{\phi}, \\ \alpha \cdot \beta^2 \cdot \gamma^2 &\approx 2, \\ \alpha \geq 1, \beta \geq 1, \gamma &\geq 1. \end{aligned} \tag{1}$$

Se seleccionó esta arquitectura debido a que su tamaño, tomando como referencia la cantidad de parámetros, es pequeño en comparación con otros modelos más profundos. Además, como se muestra en la figura 3 [17], esté modelo presenta una notable mejoría en términos de accuracy alcanzado en el ImageNet comparado con los modelos B0, B1, B2 y B3.

Ensamble de redes neuronales convolucionales: Los métodos de ensambles han tomado relevancia en los últimos años, especialmente en la clasificación de patrones [3, 15, 18]. El objetivo de estos modelos es obtener un mejor rendimiento que los algoritmos con los que está compuesto, ya que permiten eliminar los errores no correlacionados de los clasificadores individuales por medio de votación.

Existen diversas estrategias para la creación de ensambles, las más utilizadas son Bagging, Boosting y Stacking. En este trabajo se presenta un ensamble utilizando la técnica Stacking y los modelos ResNet-50 y EfficientNet-B4, tal como se muestra en la figura 4.



(a) Cardinalidad de clases (b) Porcentaje de patrones por clase

Fig. 6. Distribución de clases en el banco de datos.

2.2. Banco de datos

Para entrenar los modelos seleccionados, se utilizó un el Cassava Leaf dataset con 21,397 imágenes capturadas por agricultores de plantaciones en Uganda y etiquetadas por expertos del Instituto Nacional de Investigación de Recursos Agrícolas (NaCRRI) en colaboración con el laboratorio de IA de la Universidad de Makerere en Kampala [11]. 1,087 imágenes corresponden a cassava bacterial blight (CBB), 13,158 a cassava mosaic disease (CMD), 2,189 a cassava brown streak disease (CBSD), 2,386 a cassava green spot (CGM) y 2,577 a hojas sanas.

En la figura 5, se muestran cinco imágenes extraídas del banco de datos, cuatro de hojas de yuca con alguna enfermedad y una que no presenta ninguna enfermedad. En la figura 6, se muestra la distribución de las 21,397 imágenes en las 5 clases dentro del banco de datos. Se observa que, con 13,158 imágenes la clase 3 (CMD) es la clase mayoritaria, mientras que la clase 0 (CBB) con 1,087 imágenes es la clase minoritaria. Como se muestra en la ecuación 2, el banco de datos esta extremadamente desbalanceado, por lo que no es posible utilizar accuracy como métrica de desempeño:

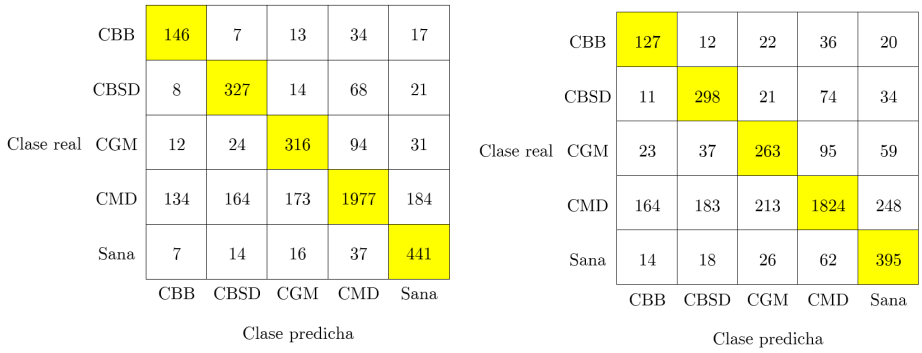
$$\text{IR} = \frac{13,158}{1,087} = 12.105. \tag{2}$$

3. Métricas de desempeño

Debido a la imposibilidad de utilizar accuracy como métrica de desempeño, es necesario aplicar otras más adecuadas derivadas de la matriz de confusión.

3.1. Precisión

Se asocia a la calidad del modelo y se refiere a la capacidad del modelo para identificar instancias positivas entre las instancias recuperadas. Matemáticamente, es la cantidad de verdaderos positivos (TP) entre la suma de verdaderos positivos (TP) y falsos positivos (FP), tal como se muestra en la ecuación 3:



(a) EfficientNet-B4, matriz de confusión (b) ResNet-50, matriz de confusión

Fig. 7. Matrices de confusión.

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{3}$$

3.2. Recall

Se asocia a la cantidad de instancias que el modelo es capaz de identificar y se refiere a la capacidad del clasificador de identificar instancias positivas. Se calcula de acuerdo con la ecuación 4:

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{4}$$

3.3. F1

La métrica F1 combina las métricas precision y recall en un solo valor. Es muy útil para comparar el rendimiento combinado de la calidad y exhaustividad entre varios algoritmos; el valor de F1 es la media armónica de precision y recall y se calcula de acuerdo con la ecuación 5:

$$F1 = \frac{2(\text{precision} \times \text{recall})}{\text{precision} + \text{recall}}. \tag{5}$$

4. Resultados

4.1. Parámetros de aprendizaje

Se utilizaron los mismos parámetros de aprendizaje en todas las arquitecturas. La función de activación seleccionada fue softmax y sparse categorical crossentropy como función de pérdida. El entrenamiento se realizó utilizando el método mini-batch Learning con tamaño 20 y una tasa de aprendizaje de 0.004. Tanto ResNet-50, como EfficientNet-B4 fueron entrenadas durante 200 épocas. Las imágenes fueron redimensionadas a 200 x 250 pixeles.

Tabla 1. Resultados EfficientNet-B4.

Clase	Precision	Recall	F1-score
CBB	47.56	67.28	55.73
CBSD	61.01	74.66	67.15
CGM	59.40	66.25	62.64
CMD	89.46	75.11	81.66
Sana	63.54	85.63	72.95

Tabla 2. Resultados ResNet-50.

Clase	Precision	Recall	F1-score
CBB	37.46	58.53	45.68
CBSD	54.38	68.04	60.45
CGM	48.26	55.14	51.47
CMD	87.23	69.30	77.24
Sana	52.25	76.70	62.16

4.2. Desempeño

El banco de datos fue dividido en los conjuntos de entrenamiento, validación y prueba con proporción 70/10/20 y el proceso de entrenamiento se realizó utilizando el método de validación hold-out. En la figura 3.1, se muestra el desempeño obtenido en el conjunto de prueba por EfficientNet-B4 (3.1) y ResNet-50 (3.1). Se observa que ambos modelos están sesgados hacia la clase cassava mosaic disease (CMD), la cual es la clase mayoritaria. En el caso de EfficientNet-B4, el 15.66 %, 15.52 %, 19.71 % y 7.18 % de los patrones de las clases CBB, CBSD, CGM y Sana respectivamente fueron clasificados como CMD.

Por otro lado, el 16.59 %, 16.89 %, 19.92 % y 12.04 % de los patrones de las clases CBB, CBSD, CGM y Sana respectivamente clasificados a través de ResNet-50, fueron clasificados como CMD. En la tabla 1, se muestra el desempeño obtenido por EfficientNet-B4 en cada una de las clases del banco de datos. Como era de esperarse, la clase más dificil de clasificar por esta CNN es cassava bacterial blight (CBB), ya que es la clase minoritaria. Por el contrario, la clase más facil de clasificar para este modelo es cassava mosaic disease (CMD), ya que es la clase mayoritaria.

En la tabla 2, se muestran los resultados de la clasificación de imágenes de enfermedades foliares obtenidos a través de la CNN ResNet-50. Se observa un comportamiento similar al obtenido con EfficientNet-B4, sin embargo, ResNet-50 no logró alcanzar el mismo desempeño. En la figura 8, se muestra la matriz de confusión obtenida por el ensamble propuesto. Se observa que el modelo fue capaz de mejorar el desempeño obtenido de forma individual por las dos arquitecturas por las que está compuesto.

La tabla 3 muestra los valores de las métricas de desempeño seleccionadas obtenidas por el ensamble propuesto. La clase que resultó más beneficiada del uso del modelo combinado fue CBB, ya que fue la clase que obtuvo un mayor aumento proporcional de patrones correctamente clasificados, pasado de un valor F1 mínimo de 45.68 % obtenido a través de ResNet50 a 63.60 % con el uso del ensamble propuesto.

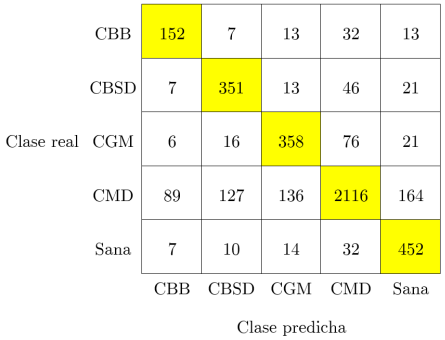


Fig. 8. Ensamble ResNet-50-EfficientNet-B4, matriz de confusión.

5. Conclusiones y trabajo a futuro

En este trabajo se presenta un método de detección de enfermedades foliares en plantaciones de yuca a través de la clasificación de imágenes usando redes neuronales convolucionales. Los modelos seleccionados fueron entrenados utilizando imágenes RGB reales y sintéticas. El modelo propuesto permite a los productores de yuca detectar cuatro enfermedades foliares sin la intervención de agrónomos, reduciendo así el tiempo y los costos asociados al diagnóstico de enfermedades foliares.

El productor podrá capturar imágenes de plantas sospechosas y presentarlas al modelo para obtener un diagnóstico probable. En comparación con varios métodos actualmente disponibles para el diagnóstico de enfermedades foliares basados en CNNs, el propuesto en esta investigación es capaz de realizar el diagnostico sin necesidad de segmentar las hojas ni realizar procesos intensivos de preprocesamiento, reduciendo así la cantidad de recursos computacionales necesarios para su ejecución.

En el estudio también se presenta la comparación del desempeño obtenido por varios modelos DL en el banco de datos cassava leaf disease, en concreto, el desempeño obtenido por ResNet-50, EfficientNet-b4 y un ensamble tipo stacking. En todos los casos, la clase más difícil de clasificar fue CBB, mientras que la clase CMD fue la que obtuvo el mejor resultado a través de la métrica F1.

A través del análisis de las matrices de confusión, se observa que, en cuanto a desempeño individual, EfficientNet-B4 supera a ResNet-50, alcanzando valores más altos en las métricas de desempeño para todas las clases. Además, se observa que el desempeño obtenido por el ensamble tipo stacking supera al desempeño individual obtenido por las CNNs utilizadas para crearlo, alcanzando valores F1 de 63.60 %, 73.97 %, 70.82 %, 85.77 % y 76.22 % para las clases CBB, CBSD, CGM, CMD y Sana respectivamente.

Tras analizar los resultados obtenidos por las arquitecturas seleccionadas en esta investigación, se considera que el desempeño exhibido por el modelo basado en ensambles de CNNs es lo suficientemente bueno como para poder mejorar el rendimiento de los cultivos de yuca a través del diagnóstico de enfermedades foliares sin necesidad de intervención humana.

Tabla 3. Resultados ensamble ResNet-50 - EfficientNet-B4.

Clase	Precision	Recall	F1-score
CBB	58.24	70.05	63.60
CBSD	68.69	80.14	73.97
CGM	67.04	75.05	70.82
CMD	91.92	80.40	85.77
Sana	67.36	87.77	76.22

Durante el desarrollo de la presente investigación, se observó que la cardinalidad de las clases es de gran importancia para el correcto aprendizaje de las redes neuronales convolucionales. Con el objetivo de mejorar los resultados obtenidos y como trabajo futuro, se propone aumentar la cardinalidad de las clases minoritarias a través de la adición de ejemplos reales y la creación de otros de forma sintética. Además, se propone utilizar modelos más profundos de las mismas familias de redes neuronales convolucionales seleccionadas para la creación del ensamble.

Es importante resaltar que, a pesar del uso del modelo presentado, la obtención de imágenes sigue siendo un trabajo intensivo, por lo que es necesario idear un método más eficiente para la captura de estas. Una posible solución sería montar una cámara sobre un dron. Con esta adición se reduciría mucho la cantidad de tiempo requerida para la obtención de diagnósticos probables.

Referencias

1. Albawi, S., Mohammed, T. A., Al-Zawi, S.: Understanding of a convolutional neural network. In: International Conference on Engineering and Technology (ICET), pp. 1–6 (2017) doi: 10.1109/ICEngTechnol.2017.8308186
2. Anyanwu, C., Ibeto, C., Ezeoha, S., Ogbuagu, N.: Sustainability of cassava (manihot esculenta crantz) as industrial feedstock, energy and food crop in Nigeria. Renewable Energy, vol. 81, pp. 745–752 (2015) doi: 10.1016/j.renene.2015.03.075
3. Balasubramaniam, S., Satheesh Kumar, K.: Optimal ensemble learning model for COVID-19 detection using chest X-ray images. Biomedical Signal Processing and Control, vol. 81, pp. 104392 (2023) doi: 10.1016/j.bspc.2022.104392
4. Basodi, S., Ji, C., Zhang, H., Pan, Y.: Gradient amplification: An efficient way to train deep neural networks. Big Data Mining and Analytics, vol. 3, no. 3, pp. 196–207 (2020) doi: 10.26599/BDMA.2020.9020004
5. Civilibal, S., Cevik, K. K., Bozkurt, A.: A deep learning approach for automatic detection, segmentation and classification of breast lesions from thermal images. Expert Systems with Applications, vol. 212, pp. 118774 (2023) doi: 10.1016/j.eswa.2022.118774
6. Goel, L., Nagpal, J.: A systematic review of recent machine learning techniques for plant disease identification and classification. Institution of Electronics and Telecommunication Engineers Technical Review, vol. 40, no. 3, pp. 423–439 (2022) doi: 10.1080/02564602.2022.2121772
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016) doi: 10.1109/CVPR.2016.90
8. Kumar-Sahu, S., Pandey, M.: An optimal hybrid multiclass SVM for plant leaf disease detection using spatial fuzzy C-Means model. Expert Systems with Applications, vol. 214, pp. 118989 (2023) doi: 10.1016/j.eswa.2022.118989

9. Latif, S., Müller, J.: Potential of cassava leaves in human nutrition: A review. Trends in Food Science and Technology, vol. 44, no. 2, pp. 147–158 (2015) doi: 10.1016/j.tifs.2015.04.006
10. Meunkaewjinda, A., Kumsawat, P., Attakitmongcol, K., Srikaew, A.: Grape leaf disease detection from color imagery using hybrid intelligent system. In: 5th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology, vol. 1, pp. 513–516 (2008) doi: 10.1109/ECTICON.2008.4600483
11. NaCRRI: Cassava leaf disease classification (2020) https://www.kaggle.com/competitions/cassava-leaf-disease-classification/data
12. Olayemi-Alebiosu, D., Dharmaratne, A., Hong Lim, C.: Improving tuberculosis severity assessment in computed tomography images using novel DAvoU-Net segmentation and deep learning framework. Expert Systems with Applications, vol. 213, pp. 119287 (2023) doi: 10.1016/j.eswa.2022.119287
13. Omondi, J. O., Yermiyahu, U.: Improvement in cassava yield per area by fertilizer application. Cassava, chapter 6 (2021) doi: 10.5772/intechopen.97366
14. Oyewola, D. O., Dada, E. G., Misra, S., Damaševičius, R.: Detecting cassava mosaic disease using a deep residual convolutional neural network with distinct block processing. PeerJ Computer science, vol. 7, pp. e352 (2021) doi: 10.7717/peerj-cs.352
15. Rashidpoor-Toochaei, M., Moeini, F.: Evaluating the performance of ensemble classifiers in stock returns prediction using effective features. Expert Systems with Applications, vol. 213, pp. 119186 (2023) doi: 10.1016/j.eswa.2022.119186
16. Rustam, F., Ashraf, I., Jurcut, A. D., Bashir, A. K., Zikria, Y. B.: Malware detection using image representation of malware data and transfer learning. Journal of Parallel and Distributed Computing, vol. 172, pp. 32–50 (2023) doi: 10.1016/j.jpdc.2022.10.001
17. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning, pp. 6105–6114 (2019) doi: 10.48550/arXiv.1905.11946
18. Tavana, P., Akraminia, M., Koochari, A., Bagherifard, A.: An efficient ensemble method for detecting spinal curvature type using deep transfer learning and soft voting classifier. Expert Systems with Applications, vol. 213, pp. 119290 (2023) doi: 10.1016/j.eswa.2022.119290
19. Tonukari, N., Ezedom, T., Enuma, C., Sakpa, S., Avwioroko, O., Eraga, L., Odiyoma, E.: White gold: Cassava as an industrial base. American Journal of Plant Sciences, vol. 6, pp. 972–979 (2015) doi: 10.4236/ajps.2015.67103

ISSN 1870-4069

Normalización de radiografías de tórax para la detección de neumonía mediante algoritmos tradicionales de aprendizaje de máquina

Salvador Ayala Raggi, Angel Ernesto Picazo Castillo,
Aldrin Barreto Flores, José Francisco Portillo Robledo

Benemérita Universidad Autónoma de Puebla,
México

{saraggi,a.picazo.2505}@gmail.com

Resumen. En este trabajo se presenta una propuesta para la normalización automática de la región de interés (pulmones) en las radiografías de tórax utilizando algoritmos tradicionales de aprendizaje de máquina, esto consiste en lograr que las imágenes sean semejantes en rotación, escala y contraste. Posteriormente se utiliza el PCA como método de reducción de características para las imágenes normalizadas. De estas características obtenidas de PCA se seleccionan las que tienen la mejor capacidad de discriminación de clase usando el criterio de Fisher.La normalización en conjunto con la selección de características demuestran formar un método capaz de lograr que clasificadores como el K-NN ponderado y el MLP puedan obtener precisiones de 89.79 % y 91.24 % respectivamente para la clasificación de imágenes de neumonía. El método propuesto no busca sustituir a los métodos de aprendizaje profundo pero demuestran ser opciones aceptables para la clasificación.

Palabras clave: K-vecinos más cercanos, clasificación de imágenes, determinante lineal de Fisher, neumonía viral.

Chest X-Ray Normalization for Pneumonia Detection Using Traditional Machine Learning Algorithms

Abstract. This work presents a proposal for the automatic normalization of the region of interest (lungs) in chest radiographs using traditional machine learning algorithms. This involves making the images similar in rotation, scale, and contrast. Subsequently, PCA is used as a method for feature reduction for the normalized images. From these PCA features, the ones with the best class discrimination capability are selected using Fisher's criterion. Normalization, together with feature selection, demonstrates a method capable of achieving precision of 89.79% and 91.24% for pneumonia image classification using weighted K-NN and MLP classifiers, respectively. The proposed method does not seek to replace deep learning methods but demonstrates acceptable options for classification.

Keywords: K-nearest neighbors, image classification, Fisher linear determinant, viral pneumonia.

1. Introducción

La Neumonía es una enfermedad pulmonar causada por bacterias y virus, una persona puede ser contagiada por medio del aire, saliva o moco. Además, los niños y las personas mayores tienen un mayor riesgo de ser contagiados, de acuerdo con [1]. En la actualidad, ya existen diversos métodos para la detección de esta enfermedad como las tomografías, radiografías de pecho y ultrasonidos.

Sin embargo, las tomografías son más caras que una radiografía y el ultrasonido no siempre está disponible o asequible, por esto, las radiografías resultan ser un método de detección más común[2, 3, 4, 5, 6]. En la actualidad, existen bancos de información que contienen radiografías ya etiquetadas y pueden ser utilizadas para el entrenamiento de diferentes algoritmos de aprendizaje de máquina[10]. La construcción de estos bancos ha sido un esfuerzo conjunto de instituciones y médicos expertos [7, 8, 9].

El problema con estas imágenes radica en la falta de uniformidad en la región de interés (pulmones), ya que hay algunas que contienen información no deseada o irrelevante para una clasificación, como otras partes del cuerpo u objetos que cubren el pecho; esto puede causar que los algoritmos de clasificación reduzcan sus métricas de precisión [11]. En este trabajo, se proponen dos procedimientos seriados para el tratamiento de las radiografías de tórax.

El primero consiste en la "normalización" de las imágenes, es decir que el banco de información contenga imágenes con su región de interés con la misma alineación, ubicación y escala tanto como sea posible, además de una mejora de contraste. Para el segundo procedimiento, después de obtener dicho banco se procederá a realizar un análisis de las características que las imágenes posean.

Todo esto con el objetivo de encontrar las características que mejor discriminen las clases para que el algoritmo de clasificación incrementa su métrica de precisión. Este trabajo comienza con la mención del banco de datos utilizado, continuando en la parte 2 con la descripción, teoría y resultados de nuestro "Algoritmo Localizador de Pulmones" (ALP), todo esto para el procedimiento de normalización.

Después en la parte 3, se menciona la teoría de "Eigenfaces" y del discriminante lineal de Fisher que se utilizaron en nuestro análisis para las características de las imágenes del nuevo banco de datos. Finalmente, se hará la comparación de las métricas de precisión al usar o no usar nuestro procedimiento propuesto utilizando los algoritmos "K -vecinos más cercanos ponderado" (Weighted K-NN) y el "Perceptrón multicapa" (MLP) como clasificadores [12].

1.1. Trabajo relacionado

En la actualidad ya existen diferentes metodologías para la clasificación de radiografías de tórax, como en[24, 25, 7, 16, 17, 18]. Utilizando algoritmos de aprendizaje profundo o clasificadores tradicionales de aprendizaje de máquina [19, 20], reportando precisiones de clasificación superiores al 96 %. Sin embargo, las arquitecturas utilizadas en los algoritmos aún son insuficientes para la clasificación precisa de neumonía causada por COVID-19 [21] .

Fig. 1. Ejemplos de la base de datos [24, 25]. Neumonía viral (Derecha), COVID-19 (medio) y Normal (Izquierda).

Esto abre la posibilidad a que se puedan hacer otras propuestas para la clasificación de radiografías y que no estén basadas en redes neuronales convolucionales (CNNs) como en [13] que utiliza un MLP y una arquitectura basada en involución de imágenes reportando una precisión de la clasificación máxima de 94.49 %. La selección de características ha sido capaz de incrementar la precisión de clasificación en otros trabajos.

Como en [14] para su máquina de vectores de soporte o en [15] para su algoritmo K-NN. Los resultados de nuestro trabajo no pretenden sustituir las CNNs para la clasificación de imágenes, sino proponer otra opción, como en [13], y demostrar que nuestro trabajo tiene resultados comparables con algoritmos de aprendizaje profundo.

1.2. Base de datos de imágenes radiográficas

La base de datos utilizada para este trabajo fue ÇOVID-19 Radiography Database"[24, 25] de Kaggle. Esta base de datos fue seleccionada, ya que ha sido utilizada en otros trabajos similares[26, 27]. El contenido de esta base es 3616 imágenes ya etiquetadas como opacidad pulmonar (otras enfermedades pulmonares), 1345 como Neumonía (algunas causada por COVID-19) y 10,192 como normal (saludable). Como puede verse en la figura 1, la región de interés de las imágenes (Pulmones) presentan diferencias notorias en cuanto a escala, traslación, rotación y contraste.

2. Algoritmo localizador de pulmones (ALP)

Este algoritmo tiene como objetivo localizar los pulmones en las radiografías, también cuenta con su etapa de entrenamiento y prueba como puede verse en 2. Para la primera parte, se seleccionaron 400 imágenes aleatoriamente de las clases Neumonía, COVID-19 y Normal de la base de datos. A todas las imágenes se les aplicó la ecualización del histograma [22, 23] para posteriormente realizar un etiquetado manual de las zonas de interés de las imágenes.

Después se hizo un aumento de datos creando diez imágenes nuevas por cada imagen ya etiquetada. Finalmente, se realizó una reducción de dimensionalidad aplicando el método de .[Ei]genfaces"basados en el análisis de componentes principales (PCA) [30, 31]. En la parte de prueba, a una nueva imagen en la entrada se le aplica la mejora de contraste y su proyección en el espacio de las "Eigenfaces" para poder realizar su comparación mediante el algoritmo de regresión de "K-NN ponderado".

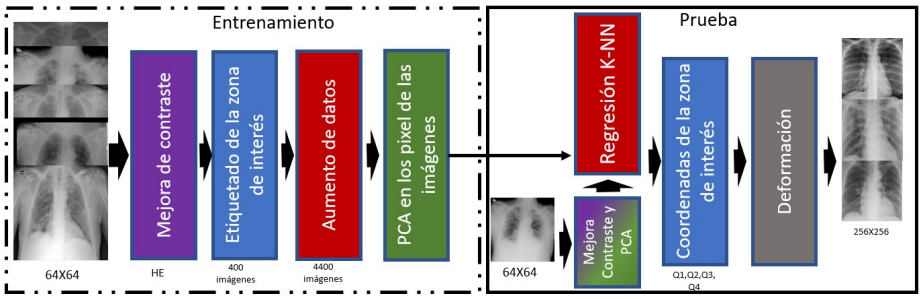


Fig. 2. Descripción del Algoritmo Localizador de Pulmones. En la entrada de la etapa de prueba se encuentra un ejemplo de radiografía y a la salida se tienen la región de interés ya extraída.

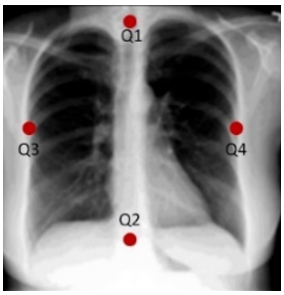


Fig. 3. Coordenadas Q1, Q2 ,Q3 y Q4 en una radiografía.

El objetivo es encontrar las imágenes que sean más parecidas. Las coordenadas de las zonas de interés de dichas imágenes serán utilizadas en la regresión para predecir las coordenadas de la imagen en la entrada, todo esto sucede de manera automática. Ya con las coordenadas calculadas, se utiliza la operación de deformación (Warping) y la interpolación para construir la nueva imagen, que solo contiene la región de interés, a partir de la imagen de entrada.

2.1. Ecualización del histograma

La ecualización del histograma (HE) es una técnica de procesamiento de imágenes que tiene como objetivo mejorar el contraste de una imagen al redistribuir los valores de los píxeles. Esta técnica se basa en la idea de que una distribución uniforme de valores de intensidad en una imagen proporciona una mejor representación visual de la misma.

Para lograr la ecualización del histograma, se calcula el histograma de la imagen original para obtener una representación gráfica de la distribución de los valores de intensidad. Después, se determina una función de transformación que redistribuye los valores de intensidad para lograr una distribución más uniforme. Esta función se aplica a la imagen original para obtener una versión ecualizada del histograma que presenta un mejor contraste.

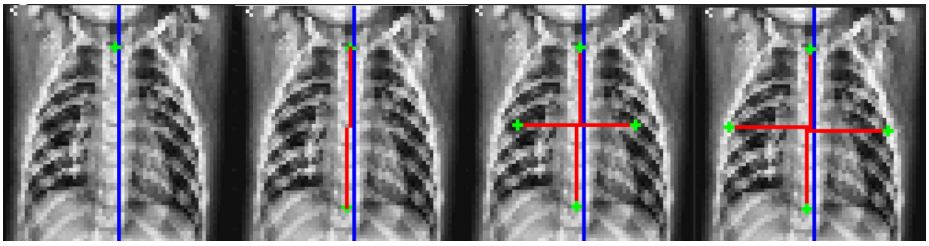


Fig. 4. Secuencia para la localización de los puntos Q, primero Q1 es localizado, después Q2 para que Q3 y Q4 aparezcan en la recta perpendicular que cruza por el punto medio de la recta Q1Q2. Finalmente, Q3 y Q4 son ajustados.

La ecualización del histograma es una técnica ampliamente utilizada en el procesamiento de imágenes y ha demostrado ser efectiva para mejorar la calidad visual de imágenes en una variedad de aplicaciones.

2.2. Etiquetado de coordenadas para la etapa de entrenamiento

Cada una de las imágenes seleccionadas para esta etapa necesita ser etiquetada manualmente con un arreglo de coordenadas que delimita la zona de interés de los pulmones. Estos puntos se convertirán en las nuevas características que el K-NN ponderado utiliza para su predicción de las imágenes de prueba. El arreglo de coordenadas puede verse en la figura 3, se trata de 4 puntos Q1 (x1,y1), Q2 (x2,y2), Q3 (x3,y3) y Q4 (x4,y4). Q1 y Q2 representan el largo de los pulmones y Q3 y Q4 el ancho de los mismos. El proceso del etiquetado puede verse en la figura 4.

Primero se localiza manualmente el punto Q1 en la parte superior de los pulmones, utilizando la columna vertebral como referencia. Posteriormente, Q2 es colocado en la parte inferior de los pulmones. Cuando Q1 y Q2 están colocados automáticamente aparece una recta que los une, en el punto medio de esta recta se coloca una recta perpendicular que contiene a los puntos Q3 y Q4. Estos últimos dos están limitados a solo moverse en la recta perpendicular y pueden tener una distancia diferente al punto medio de la recta Q1Q2, debido a que los pulmones no son simétricos entre ellos.

2.3. Aumento de datos

Esta técnica es utilizada en muchas tareas del aprendizaje de máquina como la clasificación de imágenes para incrementar una base de datos limitada y evitar el sobreajuste [28, 29]. Para nuestro algoritmo, existe una gran cantidad de imágenes en la base de datos utilizada [24, 25].

Sin embargo, con el objetivo de tener una distribución normal en los valores de las coordenadas de la región de interés, se optó por crear radiografías artificiales basadas en las imágenes previamente ya etiquetadas utilizando operaciones de traslación y rotación. Primeramente es necesario definir el rango para las operaciones de las imágenes, para la rotación fue de -10° a 10° (sugerido por [25] y para la traslación fue de -5 a 5 píxeles, estos valores fueron calculados analizando las coordenadas de las 400 imágenes originales.

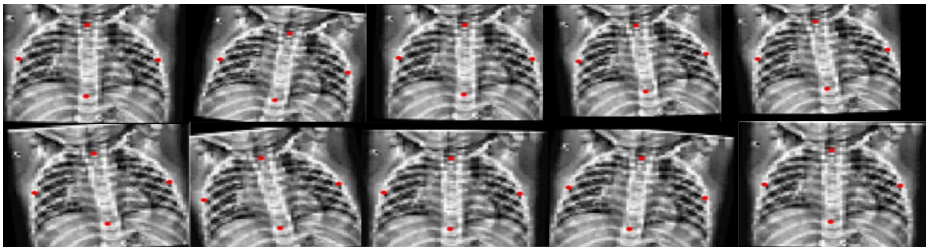


Fig. 5. Ejemplo de imágenes artificiales durante el aumento de datos.

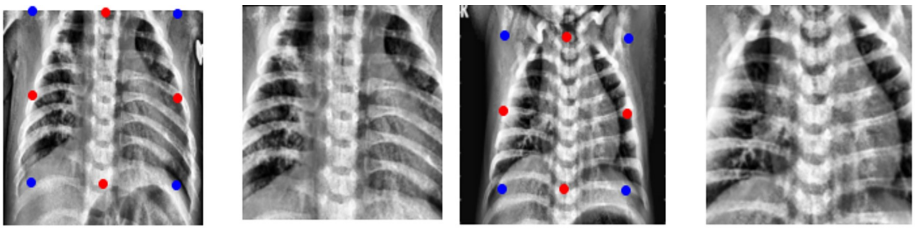


Fig. 6. Imágenes de ejemplo con sus regiones de interés ya extraídas.

Después se crearon 10 imágenes artificiales por cada imagen original tomando valores completamente aleatorios que estén dentro del rango predefinido. Las operaciones se aplicaron a la imagen y también a sus coordenadas de la zona de interés. En total la base de datos del ALP contiene 4400 imágenes con una distribución normal en las coordenadas de su arreglo de puntos. En la figura 5 puede verse un ejemplo de las imágenes artificiales creadas junto con su arreglo de puntos.

2.4. Regresión para calcular las coordenadas de los pulmones

Como se ve en la figura 2, para le etapa de prueba se tiene una nueva imagen a la que se le desea obtener su región de interés. Para esta parte y de manera automática, a la imagen de prueba de se le aplica la mejora de contraste y reducción de características mediante su proyección en las "Eigenfaces". Los pesos obtenidos en la proyección se utilizan en el "K-NN ponderado" para encontrar los vecinos más parecidos mediante la distancia euclidiana del espacio de las "Eigenfaces".

Ya conociendo los vecinos más cercanos se puede realizar una regresión utilizando las coordenadas de las regiones de interés de dichos vecinos para predecir las coordenadas de los pulmones de la imagen de prueba. Las ecuaciones de regresión (1 y 2) se tienen que utilizar para cada coordenada ya sea x o y de cada punto Q hasta que todo el arreglo de puntos (Q1, Q2, Q3 y Q4) se complete. Las ecuaciones se describen abajo:

$$x_i = \frac{1}{k} \sum_{i=1}^{k} x_{ni}, \tag{1}$$

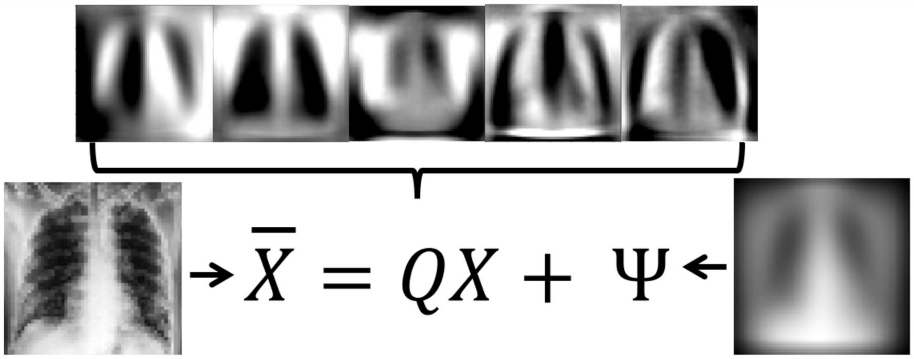


Fig. 7. Imagen de salida (Izquierda) como la combinación lineal de la matriz de Eigenfaces Q (medio) y la imagen de entrada más la cara media (derecha).

$$y_i = \frac{1}{k} \sum_{i=1}^{k} y_{ni}. \tag{2}$$

2.5. Image Warping

En el contexto del procesamiento de imágenes y visión por computadora, el término "Warping" se refiere a una transformación geométrica que se aplica a una imagen para cambiar su forma o perspectiva. La técnica de Warping se basa en la aplicación de una función de transformación a cada uno de los píxeles de la imagen original.

Esta función de transformación puede ser lineal o no lineal, y puede ser definida por diferentes parámetros dependiendo del tipo de transformación deseada. Algunas técnicas de Warping comunes incluyen la homografía, la transformación afín y la transformación de Fourier [33]. Después de tener las coordenadas mediante la regresión, se utiliza la operación Warping para la extracción de la zona de interés.

En la figura 6 pueden verse ejemplos de las imágenes originales y la obtención automática de sus coordenadas de la región de interés (puntos rojos) y las coordenadas que utiliza la operación Warping (puntos azules). En el lado derecho puede verse la imagen de salida del ALP.

3. Reducción y selección de características

Después utilizar el ALP en todo el banco de datos para extraer todas las regiones de interés, estas nuevas imágenes recibirá otro preprocesamiento antes de entrar a algún algoritmo clasificador. Para nuestro trabajo proponemos el uso de las Eigenfaces [30, 31] como método de reducción de características.

Además, incluimos el análisis estadístico de las mismas obtenido mediante el discriminante lineal de Fisher. Estos dos métodos en conjunto garantizan la obtención de un número pequeño de características que mejor discriminan las clases para una clasificación.

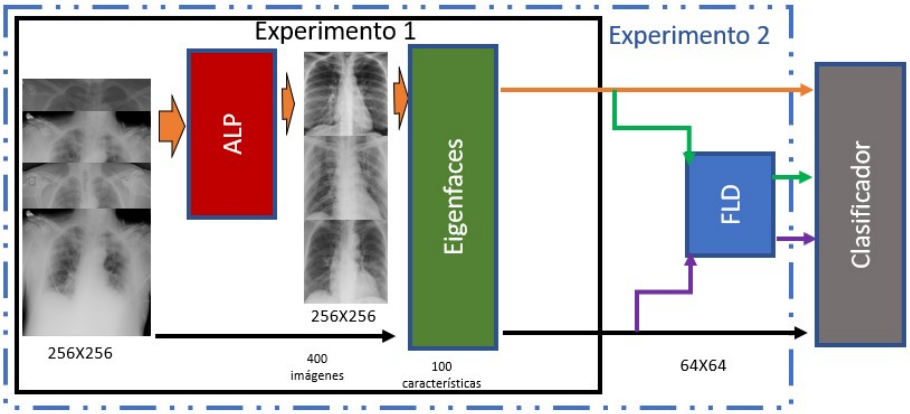


Fig. 8. Diagrama acerca de los dos experimentos. Cada flecha que llega al clasificador es una métrica de precisión y se realizará una comparación de todas.

3.1. Eigenfaces

El método de las Eigenfaces fue creado primeramente por Turk y Pentland [30] y por Sirovich y Kirby [31]. Este método está basado en el análisis de componentes principales (PCA) y su objetivo es la reducción de la dimensionalidad de las imágenes en el banco de datos [32]. Debido a que cada píxel se convierte en una dimensión o variable para analizar y que tenemos imágenes con una resolución de 256x256, cada imagen demandaría un largo tiempo de procesamiento.

Cada Eigenface es una imagen que muestra una estructura de rasgos o patrones que son comunes en el conjunto de imágenes utilizado. Estas Eigenfaces son ordenadas de acuerdo a la varianza de las imágenes de entrada, pueden ser utilizadas para reconstruir cualquier imagen como una combinación lineal de ellas y representarla también en un espacio de menor dimensión.

En la figura 7 puede verse la ecuación de las Eigenfaces y la matriz Q de Eigenfaces. Por otro lado, es necesario que las imágenes de entrada tengan condiciones de luz y ángulo semejantes. Por esta razón el ALP es utilizado antes como una manera de normalizar las radiografías.

3.2. El discriminante lineal de Fisher para la selección de características

El discriminante lineal de Fisher (FLD) busca encontrar una proyección lineal de las características que maximice la separación entre las clases del banco de datos. Esta proyección se realiza evaluando las características una a una y analiza que las medias de las observaciones de cada clase estén lo más alejadas posibles y que las varianzas dentro de cada clase sean lo más pequeñas posibles. Utilizando este análisis se pueden seleccionar las características obtenidas del método de las Eigenfaces que mejor discriminen las clases del conjunto de datos [34].

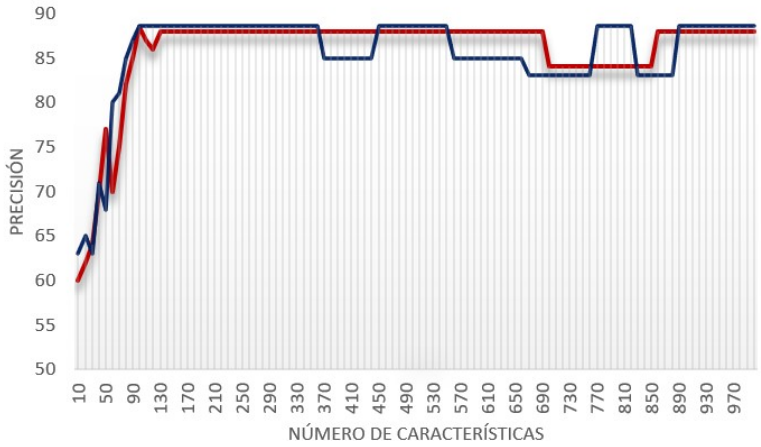


Fig. 9. Valores de precisión variando el numero de características. K-NN ponderado (Linea roja) y MLP (Línea azul).

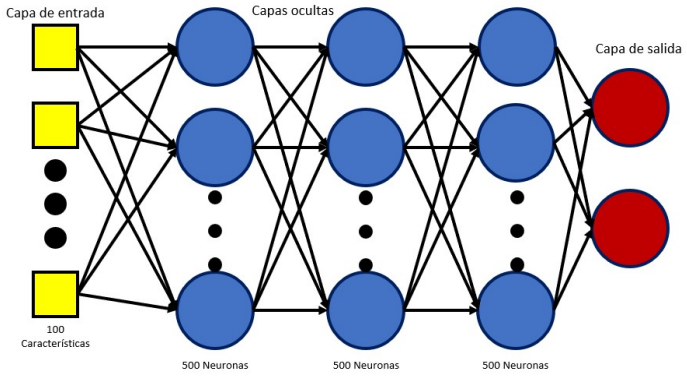


Fig. 10. Topología del MLP.

El FLD ha sido usado en traba como en [35], es denotado como J y su fórmula se encuentra en la ecuación 3:

$$J_i = \frac{\left(\mu_{ic_0} - \mu_{ic_1}\right)^2}{\sigma^2_{ic_0} + \sigma^2_{ic_1}}. \tag{3}$$

4. Resultados experimentales

Para esta sección, se dividieron los resultados en dos experimentos. El primero fue la comparación del banco de datos original contra el banco de imágenes ya normalizadas. Para ambos bancos se utilizaron las Eigenfaces como método de reducción de dimensiones. El segundo experimento implementará la selección de características para los dos bancos de imágenes previamente mencionados.

Tabla 1. Resultados del K-NN ponderado para diferentes valores de K.

Parámetro K	Banco normalizado	Banco no alineado
3	86.29 %	73.76 %
4	87.17 %	74.48 %
5	87.31 %	75.36 %
6	87.60 %	75.80 %
7	87.9 %	75.80 %
8	88.04 %	75.80 %
9	88.19 %	76.23 %
10	88.19 %	76.96 %
11	88.48 %	77.25 %
12	88.48 %	77.25 %

Dando un total de cuatro métricas de precisión en la clasificación, las cuales serian: sin normalizar y sin selección de características, sin normalizar y con selección de características, normalizadas sin selección de características y normalizadas con selección de características. Se utilizaron los algoritmos de K-NN ponderado y MLP como clasificadores. En la figura 8 pueden verse ambos experimentos y los diferentes bloques que los conforman.

4.1. Preparación para el primer experimento

Se utilizaron 1345 (256X256) imágenes de neumonía y otras 1345 (256X256) de normal, a todas estas imágenes se les extrajo la región de interés formando el banco de imágenes normalizado usando el ALP. Las imágenes se dividieron en 2000 imágenes para el entrenamiento, 1000 de cada clase. Para la etapa de prueba se tomaron 700 imágenes, 350 por cada clase. Para seleccionar un número adecuado de características, se realizaron pruebas variando el número de características en ambos clasificadores desde 10 características hasta 1000.

En la figura 9 se muestra el gráfico que representa el número de características contra el valor de precisión de cada clasificador. Con esto puede verse que para ambos clasificadores 100 características son suficientes para alcanzar el máximo valor. Para el MLP se utilizaron 3 capas ocultas con 500 neuronas cada una. En la figura 10 se puede observar la topología completa del MLP usado.

4.2. Resultados y discusión del primer experimento

Para el K-NN ponderado se utilizaron diferentes valores para el parámetro K en la tabla 1 puede verse los resultados para ambos bancos de datos. Puede verse que el mejor valor de K es 11, además el banco de datos normalizado tiene mejores resultados. Para el MLP el banco original obtuvo un 79.81 % de precisión y el alineado obtuvo un 88.64 %. Para ambos clasificadores es mejor utilizar las imágenes alineadas que las originales.

Tabla 2. Resultados de K-NN ponderado para diferentes valores de K.

K-NN ponderado (banco alineado)	
Sin selección	Selección de características
88.48 %	89.62 %
K-NN ponderado (banco no alineado)	
Sin selección	Selección de características
77.25 %	79.21 %
MLP (banco alineado)	
Sin selección	Selección de características
88.64 %	90.08 %
MLP (banco no alineado)	
Sin selección	Selección de características
79.81 %	80.32 %

Tabla 3. Resultados de K-NN ponderado y MLP para la validación cruzada.

Clasificador	**K-NN ponderado**	**MLP**
Test 1	88.62\ %	90.08\ %
Test 2	88.77\ %	90.08\ %
Test 3	89.79\ %	90.22\ %
Test 4	89.79\ %	91.24\ %
Test 5	89.65\ %	91.24\ %
Promedio	89.32\ %	91.10\ %
Desviación estándar	0.577\ %	0.577\ %

4.3. Preparación para el segundo experimento

Durante el segundo experimento, el FLD fue aplicado a las 1000 características de la figura 9 y se obtuvieron las mejores 100 características de ambos bancos para la clasificación. Las 100 características seleccionadas mediante el FLD se compararon con las obtenidas mediante las Eigenfaces. Los clasificadores del primer experimento fueron usados en este también.

4.4. Resultados y discusión para el segundo experimento

En la tabla 2 se pueden ver todas las métricas de precisión para los dos clasificadores. Primeramente el banco de datos normalizado obtiene los valores más altos. Además la selección de características logró incrementar el valor de precisión, ya que las características con poca relevancia para la clasificación fueron excluidas.

4.5. Resultados adicionales

El conjunto de datos alineados fue utilizado para realizar una validación cruzada para poder observar la consistencia del conjunto de algoritmos propuestos en este trabajo. En la tabla 3 pueden verse los resultados de las 5 pruebas realizadas además del promedio y la desviación estándar para cada clasificador.

5. Conclusiones y trabajo futuro

En este trabajo se introdujo un problema sobre la alineación de las región de interés en la radiografías de tórax. Nuestro método propuesto es un conjunto de algoritmos que toman las regiones de interés de las imágenes y logra normalizarlas además de representarlas con un número menor de características.

A lo largo de este trabajo pudo verse que la normalización de imágenes y la selección de características pueden generar mejores resultados en la clasificación. También puede verse que al usar dichas técnicas en conjunto generan una propuesta consistente, resultado de la validación cruzada. La normalización de imágenes genera que el método de las Eigenfaces, basado en el PCA, otorgue mejores resultados, ya que ahora la región de interés aparece en el mismo ángulo y la misma iluminación en el conjunto de datos. El FLD nos otorga características con mejores capacidades para la clasificación de las clases.

Nuestra propuesta logra alcanzar valores de precisión aceptables frente a otros trabajos del estado del arte sin la necesidad de utilizar técnicas basadas en CNNs. Para el trabajo futuro, el ALP puede ser utilizado en otras bases de datos y para la detección de otras enfermedades pulmonares, además de que otros clasificadores pueden utilizar las imágenes normalizadas o las mejores características que el FLD pueda dar para representar el conjunto de datos.

Referencias

1. World Health Organization: Pneumonia (2023) http://www.who.int/es/news-room/fact-sheets/detail/pneumonia
2. Alzahrani, S. A., Al-Salamah, M. A., Al-Madani, W. H., Elbarbary, M. A.: Systematic review and meta-analysis for the use of ultrasound versus radiology in diagnosing of pneumonia. Critical Ultrasound Journal, vol. 9, no. 1 (2017) doi: 10.1186/s13089-017-0059-y
3. Amatya, Y., Rupp, J., Russell, F. M., Saunders, J., Bales, B., House, D. R.: Diagnostic use of lung ultrasound compared to chest radiograph for suspected pneumonia in a resource-limited setting. International Journal of Emergency Medicine, vol. 11, no. 1 (2018) doi: 10.1186/s12245-018-0170-2
4. Moberg, A. B., Taléus, U., Garvin, P., Fransson, S. G., Falk, M.: Community-acquired pneumonia in primary care: clinical assessment and the usability of chest radiography. Scandinavian Journal of Primary Health Care, vol. 34, no. 1, pp. 21–27 (2016) doi: 10.3109/02813432.2015.1132889
5. Ticinesi, A., Lauretani, F., Nouvenne, A., Mori, G., Chiussi, G., Maggio, M., Meschi, T.: Lung ultrasound and chest x-ray for detecting pneumonia in an acute geriatric ward. Medicine, vol. 95, no. 27, pp. e4153 (2016) doi: 10.1097/md.0000000000004153
6. Niederman, M.: Community-acquired pneumonia annals of internal medicine. 2nd edition (2009)
7. Salvatore, C., Interlenghi, M., Monti, C. B., Ippolito, D., Capra, D., Cozzi, A., Schiaffino, S., Polidori, A., Gandola, D., Alì, M., Castiglioni, I., Messa, C., Sardanelli, F.: Artificial intelligence applied to chest X-ray for differential diagnosis of COVID-19 pneumonia. Diagnostics, vol. 11, no. 3, pp. 530 (2021) doi: 10.3390/diagnostics11030530
8. Qin, C., Yao, D., Shi, Y., Song, Z.: Computer-aided detection in chest radiography based on artificial intelligence: a survey. BioMedical Engineering OnLine, vol. 17, no. 1 (2018) doi: 10.1186/s12938-018-0544-y

9. Ghaderzadeh, M., Aria, M., Asadi, F.: X-Ray equipped with artificial intelligence: Changing the COVID-19 diagnostic paradigm during the pandemic. BioMed Research International, vol. 2021, pp. 1–16 (2021) doi: 10.1155/2021/9942873
10. Amatya, Y., Rupp, J., Russell, F. M., Saunders, J., Bales, B., House, D. R.: Diagnostic use of lung ultrasound compared to chest radiograph for suspected pneumonia in a resource-limited setting. International Journal of Emergency Medicine, vol. 11, no. 1 (2018) doi: 10.1186/s12245-018-0170-2
11. Cleophas, T. J., Zwinderman, A. H.: Machine learning in medicine. vol. 2, Springer (2013)
12. Ertel, W.: Introduction to artificial intelligence. Springer International Publishing (2017) doi: 10.1007/978-3-319-58487-4
13. Changawala, V., Sharma, K., Paunwala, M.: Averting from convolutional neural networks for chest X-Ray image classification. IEEE International Conference on Signal Processing, pp. 14–17 (2021) doi: 10.1109/SPICSCON54707.2021.9885316
14. Liu, W., Zheng, Y., Zhou, X., Chen, Q.: Axis orbit recognition of the hydropower unit based on feature combination and feature selection. Sensors, vol. 23, no. 6, pp. 2895 (2023) doi: 10.3390/s23062895
15. Lv, C., Lu, Y., Lu, M., Feng, X., Fan, H., Xu, C., Xu, L.: A classification feature optimization method for remote sensing imagery based on fisher score and mRMR. Applied Sciences, vol. 12, no. 17, pp. 8845 (2022) doi: 10.3390/app12178845
16. Hamza, A., Attique Khan, M., Wang, S. H., Alhaisoni, M., Alharbi, M., Hussein, H. S., Alshazly, H., Kim, Y. J., Cha, J.: COVID-19 classification using chest X-ray images based on fusion-assisted deep Bayesian optimization and Grad-CAM visualization. Frontiers in Public Health, vol. 10 (2022) doi: 10.3389/fpubh.2022.1046296
17. Nillmani, Sharma, N., Saba, L., Khanna, N. N., Kalra, M. K., Fouda, M. M., Suri, J. S.: Segmentation-based classification deep learning model embedded with explainable artificial inteligence for COVID-19 detection in chest X-ray scans. Diagnostics, vol. 12, no. 9, pp. 2132 (2022) doi: 10.3390/diagnostics12092132
18. Gazda, M., Plavka, J., Gazda, J., Drotar, P.: Self-Supervised Deep Convolutional Neural Network for Chest X-Ray Classification. IEEE Access, vol. 9, pp. 151972–151982 (2021) doi: 10.1109/access.2021.3125324
19. Do, T. N., Le, V. T., Doan, T. H.: SVM on top of deep networks for COVID-19 detection from chest X-ray images. Journal of information and communication convergence engineering, vol. 20, no. 3, pp. 219–225 (2022) doi: 10.56977/jicce.2022.20.3.219
20. El-Kenawy, E. S. M., Mirjalili, S., Ibrahim, A., Alrahmawy, M., El-Said, M., Zaki, R. M., Eid, M. M.: Advanced meta-heuristics, convolutional neural networks, and feature selectors for efficient COVID-19 X-Ray chest image classification. IEEE Access, vol. 9, pp. 36019–36037 (2021) doi: 10.1109/access.2021.3061058
21. Ridzuan, M., Bawazir, A. A., Navarette, I. G., Almakky, I., Yaqub, M.: Self-supervision and multi-task learning: challenges in fine-grained COVID-19 multi-class classification from Chest X-rays (2022) doi: 10.48550/ARXIV.2201.06052
22. Gonzales, R. C., Woods, R.: Digital image processing. Pearson, 4 ed (2018)
23. Moeslund, T. B.: Introduction to video and image processing. Springer London (2012) doi: 10.1007/978-1-4471-2503-7
24. Chowdhury, M. E. H., Rahman, T., Khandakar, A., Mazhar, R., Kadir, M. A., Mahbub, Z. B., Islam, K. R., Khan, M. S., Iqbal, A., Emadi, N. A., Reaz, M. B. I., Islam, M. T.: Can AI help in screening viral and COVID-19 pneumonia? IEEE Access, Institute of Electrical and Electronics Engineers (IEEE), vol. 8, pp. 132665–132676 (2020) doi: 10.1109/access.2020.3010287
25. Rahman, T., Khandakar, A., Qiblawey, Y., Tahir, A., Kiranyaz, S., Abul Kashem, S. B., Islam, M. T., Al Maadeed, S., Zughaier, S. M., Khan, M. S., Chowdhury, M. E. H.:

Exploring the effect of image enhancement techniques on COVID-19 detection using chest X-ray images. Computers in Biology and Medicine, vol. 132, pp. 104319 (2021) doi: 10.1016/j.compbiomed.2021.104319.
26. Muljo, H. H., Pardamean, B., Purwandari, K., Cenggoro T. W.: Improving lung disease detection by joint learning with COVID-19 radiography database. Communications in Mathematical Biology and Neuroscience, vol. 2022
27. Islam, N., Ebrahimzadeh, S., Salameh, J. P., Kazi, S., Fabiano, N., Treanor, L., Absi, M., Hallgrimson, Z., Leeflang, M. M., Hooft, L., van der Pol, C. B., Prager, R., Hare, S. S., Dennie, C., Spijker, R., Deeks, J. J., Dinnes, J., Jenniskens, K., Korevaar, D. A., Cohen, J. F., et al.: Thoracic imaging tests for the diagnosis of COVID-19. Cochrane Database of Systematic Reviews, vol. 2021, no. 3 doi: 10.1002/14651858.cd013639.pub4
28. Mikolajczyk, A., Grochowski, M.: Data augmentation for improving deep learning in image classification problem. International Interdisciplinary PhD Workshop IEEE (2018) doi: 10.1109/iiphdw.2018.8388338
29. Connor, S., Khoshgoftaar, T.: A survey on image data augmentation for deep learning. Journal of Big Data, vol. 6, no. 1 (2019) doi: 10.1186/s40537-019-0197-0
30. Turk, M., Pentland, A.: Eigenfaces for recognition. Journal of Cognitive Neuroscience, vol. 3, no. 1, pp. 71–86 (1991) doi: 10.1162/jocn.1991.3.1.71
31. Kirby, M., Sirovich, L.: Application of the Karhunen-Loeve procedure for the characterization of human faces. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 12, no. 1, pp. 103–108 (1990) doi: 10.1109/34.41390
32. Jolliffe, I.: Principal component analysis. Springer Series in Statistics (2002)
33. Szeliski, R.: Computer vision: Algorithms and applications, 2nd ed., Springer (2022)
34. Thalles, S.: An illustrative introduction to Fisher's Linear Discriminant (2023) https://sthalles.github.io/fisher-linear-discriminant/
35. Ibis, E.: Sistema de aprendizaje automático para la detección de neumonía, M. S. thesis (2022)

ISSN 1870-4069

XDApp: Clasificación de radiografías por medio de una aplicación móvil

Juan Eduardo Luján García, Areli Yesareth Guerrero Estrada, Cornelio Yáñez Márquez

Instituto Politécnico Nacional,
Centro de Investigación en Computación,
México

{jeduardolujan5,ayge.1904}@gmail.com, cyanez@cic.ipn.mx

Resumen. Existen una gran variedad de enfermedades respiratorias que pueden ser diagnosticadas por medio de estudios de imagen como lo son las radiografías de tórax. Por lo tanto, es posible analizar y clasificar distintas enfermedades con el uso de aprendizaje profundo, específicamente con la implementación de redes neuronales convolucionales. En este trabajo se presenta una aplicación móvil llamada XDApp que utiliza un modelo entrenado en un banco de datos multiclase de pacientes con COVID-19, neumonía, tuberculosis y sanos, para emitir un prediagnóstico de las radiografías. Esta aplicación permite realizar de manera accesible un prediagnóstico para ayudar a médicos especialistas y además, puede ser usada como recurso didáctico para estudiantes del área de la salud. Por último, la aplicación utiliza un modelo de TensorFlow Lite para realizar la inferencia, obteniendo así valores de AUC mayores a 95.18 % en cada clase.

Palabras clave: Aprendizaje profundo, red neuronal convolucional, radiografías, app móvil, clasificación.

XDApp: A Mobile App for Radiography Classification

Abstract. There are a wide variety of respiratory diseases that can be diagnosed through imaging studies such as chest radiographs. Therefore, it is possible to analyze and classify different diseases using deep learning, specifically with the implementation of convolutional neural networks. In this work, a mobile application is presented which uses a trained model on a multi-class dataset of patients with COVID-19, pneumonia, tuberculosis, and healthy subjects, in order to issue a pre-diagnosis of the radiographs. This application provides a pre-diagnosis accessible to help physicians and can also be used as a teaching resource for students in the health field. Finally, the application uses a TensorFlow Lite model to perform inference, obtaining AUC values greater than 95.18 % in each class.

Keywords: Deep learning, convolutional neural network, radiography, mobile app, classification.

1. Introducción

Las enfermedades respiratorias son de las más comunes en pacientes, generalmente ocasionadas por brotes estacionarios y virus altamente contagiosos. Dentro de las enfermedades respiratorias más frecuentes podemos encontrar influenza, síndrome agudo respiratorio (SARS), neumonía, tuberculosis y recientemente la enfermedad por coronavirus del año 2019 (COVID-19) ocasionada por el virus SARS-CoV-2.

La neumonía es una enfermedad infecciosa que puede ser ocasionada por virus, bacterias y hongos. Entre las bacterias más comunes que la producen está el Streptococcus neumoníae y la Haemophilus influenzae. Es la enfermedad respiratoria que causa más muertes en infantes, particularmente menores de 5 años, tan solo en 2019 ocasionó la muerte de más de 700 mil niños en todo el mundo [12, 14].

La tuberculosis es ocasionada principalmente por una bacteria llamada Mycobacterium tuberculosis. Esta bacteria ocasiona que el tejido infectado muera (necrosis) y puede manifestarse en varias partes del cuerpo, no necesariamente en los pulmones [13]. Es una de las principales causas de muerte en el mundo, provocando la muerte de 1.5 millones de personas en 2020 [13].

Por otra parte, el COVID-19 ha resultado ser una de las enfermedades infecciosas más mortíferas hasta la fecha. La reciente pandemia por COVID-19, que inició en el año 2019, ha provocado más de 191 millones de contagios y casi 3 millones de muertes tan solo en el continente americano [15].

Todas las enfermedades respiratorias mencionadas tienen algo en común, esto es que pueden ser diagnosticadas utilizando estudios de imagen como lo son radiografías de tórax (CXRAY) y tomografía computarizada (TC) [20].

Cada enfermedad presenta características radiológicas diferentes, por ejemplo, la neumonía comúnmente causa acumulación de líquido en los pulmones que se manifiesta como segmentos radiopacos (segmentos blanquecinos sobre fondo negro) [21] en las radiografías.

Por otro lado, el COVID-19 típicamente presenta, además de inflamación en los bronquios, patrones de vidrio deslustrado y de pavimento loco en la periferia de los lóbulos inferiores de los pulmones [22].

Estos pueden ayudar a los especialistas a diferenciarlo de la neumonía típica; por el contrario, la tuberculosis causa necrosis en el tejido, el cual se muestra como pequeñas bolitas obscuras dentro de los pulmones y otras partes del cuerpo [21]. Sin embargo, en necesario que un médico especialista se encargue de interpretar los estudios de imagen para poder proveer de un diagnóstico correcto.

El presente trabajo se encuentra estructurado de la siguiente manera: la sección 2 describe brevemente el estado del arte del aprendizaje profundo aplicado al área de la salud; en la sección 3, se describen los métodos y materiales utilizados para el desarrollo del presente trabajo.

En la sección 4, se detalla el modelo propuesto para resolver la tarea de clasificación multiclase; en la sección 5, se presentan los resultados y discusión de los mismos; finalmente, en la sección 6 se establecen las conclusiones y trabajo futuro de esta investigación.

Tabla 1. Particiones de cada banco de datos para conformar el banco multiclase.

Banco de datos	Partición	Número de imágenes
Tuberculosis	Entrenamiento	275
	Validación	38
	Prueba	81
Neumonía	Entrenamiento	828
	Validación	436
	Prueba	487
COVID-19	Entrenamiento	334
	Validación	47
	Prueba	97
Sanos	Entrenamiento	2133
	Validación	265
	Prueba	499

2. Aprendizaje profundo aplicado al área de la salud

Hoy en día, el aprendizaje profundo (DL, por sus siglas en inglés) es uno de los conjuntos de técnicas de aprendizaje automático (ML, por sus siglas en inglés), para tareas de visión por computadora más popular dentro del estado del arte. Más aún, la clasificación de imágenes médicas (incluyendo las de estudios radiológicos) se realiza principalmente con técnicas de DL con el uso de redes neuronales convolucionales (CNN, por sus siglas en inglés) [25, 5].

Al mismo tiempo, existen múltiples intentos por resolver tareas de clasificación de radiografías utilizando modelos ligeros de DL [19, 18, 16], que puedan ser más eficientes y que, incluso, puedan ser embebidos en dispositivos móviles como smartphones, tablets y tarjetas de desarrollo como Raspberry Pi, Jetson, entre otras. Asimismo, se han creado una gran variedad de aplicaciones móviles relacionados con temas de salud, bienestar y medicina.

En particular, se han desarrollado aplicaciones de asistencia médica ayudando a pacientes a mantener el control de sus enfermedades, mejorando el autocuidado y el seguimiento a pacientes que han mostrado una mejora en la calidad de vida, dolor y actividad en los pacientes [2, 23, 24]. Las aplicaciones móviles son una parte importante de las tecnologías del aprendizaje y conocimiento que facilitan el estudio de temas de alta complejidad, mejorando las competencias de los profesionales, ya que permiten una mayor interacción con escenarios virtuales.

El uso de imágenes para observar las estructuras es fundamental como complemento a las prácticas en el área de la salud [10]. También se han desarrollado aplicaciones para la creación de ambientes de aprendizaje que permitan comprender los factores que afectan la enseñanza de estudiantes de medicina [3]. De tal forma que, el uso e implementación de técnicas DL pueden ser de ayuda para el prediagnóstico de enfermedades por medio del análisis de radiografías.

Agregado a lo anterior, la disponibilidad de una aplicación móvil que ayude a realizar este tipo de prediagnósticos, serían útiles en lugares en los cuales no hay alta disponibilidad de médicos radiólogos.

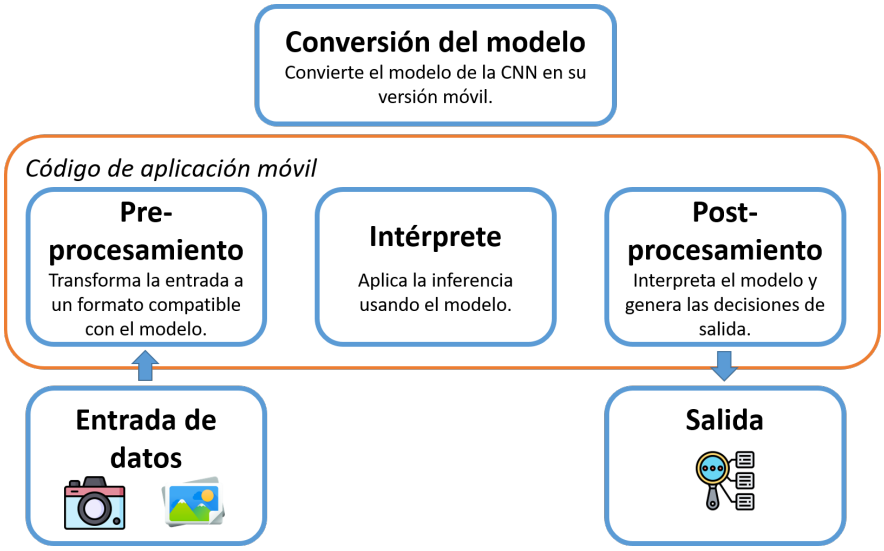


Fig. 1. Esquema de implementación de modelos de TensorFlow Lite en dispositivos móviles.

Por lo tanto, ayudaría en la práctica y enseñanza de diagnósticos que emplean imágenes radiológicas para estudiantes del área de la salud y las distintas especialidades.

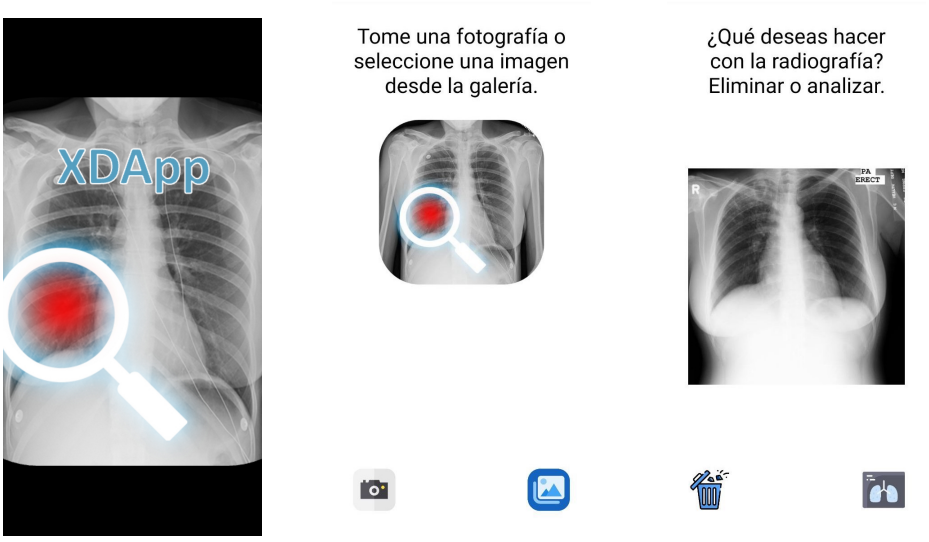
3. Métodos y materiales

En esta sección, se describen los bancos de datos de imágenes radiológicas de vista postero-anterior usados para entrenar el modelo de CNN. Después, se menciona el método de validación utilizado para dividir los bancos de datos y crear el nuevo banco de datos multiclase. Finalmente, se especifican las métricas usadas para evaluar el desempeño del modelo.

3.1. Bancos de datos

Se generó un banco de datos multiclase utilizando bancos de datos de tuberculosis, COVID-19 y neumonía para obtener a los pacientes enfermos; y los bancos de tuberculosis, neumonía y neumotórax [7] para los pacientes sanos. A continuación, se describen brevemente los bancos de datos empleados para obtener las imágenes de pacientes enfermos.

Tuberculosis. Es una colección de imágenes de rayos X del tórax de dos hospitales, recopilada por el Instituto Nacional de Salud de los Estados Unidos [8]. Está formado por el conjunto del condado de Montgomery, el cual posee 138 imágenes, de las cuales 80 son casos sanos y 58 son enfermos de tuberculosis; así como el conjunto de Shenzhen, que tiene 662 radiografías frontales con 326 pacientes sanos y 336 pacientes enfermos de tuberculosis.



(a) (b) (c)

Fig. 2. Capturas de pantalla de la aplicación móvil. a)Imagen de bienvenida a la aplicación; b)Pantalla principal para captura o carga de imágenes; c)Imagen seleccionada cargada en la aplicación.

COVID-19. Comprende una colección de imágenes de rayos X de tórax de pacientes enfermos con COVID-19 recopilados por Cohen et al. [4] durante el 2020, con un total de 478 imágenes.

Neumonía. Se emplearon dos bancos de datos de neumonía, el conjunto de datos pediátrico generado por Kermany et al. [9] y el conjunto RSNA Pneumonia Detection Challenge (RSNA-PC) disponible como una competencia de la plataforma Kaggle [1].

El conjunto publicado por Kermany et al. está formado por 5853 imágenes en el conjunto de entrenamiento, de los cuales 3883 son pacientes infectados con neumonía y 1349 son pacientes sanos; y 624 en el conjunto de prueba, dividido en 390 imágenes de pacientes enfermos y 234 de sanos.

El banco de datos RSNA-PC está formado por 3000 radiografías de tórax de adultos en el conjunto de prueba y 26684 en el conjunto de entrenamiento, de las cuales el 29 % corresponde a imágenes normales y 71 % de imágenes con áreas dañadas del pulmón causadas por neumonía. De este banco, se tomaron 478 imágenes de pacientes sanos y 478 de enfermos para balancearlo con el número de casos del banco de COVID-19.

3.2. Método de validación

Para la validación de los resultados se empleó el método Hold-out 70-10-20 para obtener el conjunto de entrenamiento, validación y prueba. Este método consiste en dividir aleatoriamente el conjunto de datos en 70 % para el conjunto de entrenamiento, 10 % para la validación y 20 % para el conjunto de prueba.

[1] kaggle.com/c/rsna-pneumonia-detection-challenge

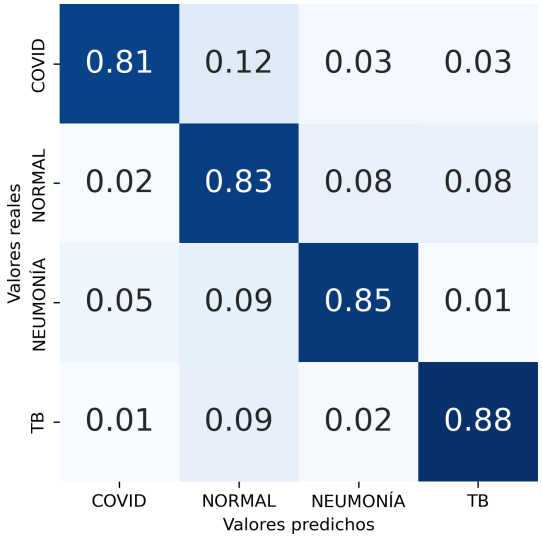


Fig. 3. Resultados de clasificación de enfermedades con NanoChest-Net.

Este método se empleó para todos los bancos de datos, excepto para el de Kermany que ya cuenta con un conjunto de prueba oficial. En este caso, el conjunto de entrenamiento se dividió en 90 % para entrenamiento y 10 % para la validación. En la tabla 1 se muestran las particiones para cada conjunto de datos.

3.3. Métricas de desempeño

Para bancos de datos desbalanceados, existen métricas útiles para evaluar el desempeño de un clasificador. Por ejemplo, la tasa de verdaderos positivos (TPR, por sus siglas en inglés), también conocida como sensibilidad (eq. (1)) evalúa la habilidad del clasificador para detectar la presencia de una condición entre el total de instancias:

$$\mathrm{TPR/sensibilidad} = \frac{\mathrm{tp}}{\mathrm{tp} + \mathrm{fn}}. \tag{1}$$

La tasa de falsos positivos (FPR, por sus siglas en inglés), indica cuando se produce una falsa alarma, es decir, la proporción de instancias predichas erróneamente como positivas por el clasificador (eq. (2)):

$$\mathrm{FPR} = \frac{\mathrm{fp}}{\mathrm{tn} + \mathrm{fp}}. \tag{2}$$

Por otro lado, el valor de predicción positiva (PPV por sus siglas en inglés) o precisión (eq. (3)), mide la proporción de instancias verdaderamente positivas entre las predichas positivas por el clasificador:

$$\mathrm{PPV/precision} = \frac{\mathrm{tp}}{\mathrm{tp} + \mathrm{fp}}. \tag{3}$$

Tabla 2. Métricas de desempeño en banco de datos multiclase.

Clase	Precisión	Sensibilidad	F1	AUC
COVID-19	0.6991	0.8144	0.7524	0.9786
NORMAL	0.8697	0.8297	0.8492	0.9518
NEUMONÍA	0.9035	0.8460	0.8738	0.9601
TB	0.5966	0.8765	0.7100	0.9829
Promedio macro	0.7673	0.8417	0.7964	0.9687

Otra medida muy utilizada con datos desbalanceados es la medida F_1. La medida F_1 (eq. (4)), evalúa la similitud entre las instancias positivas verdaderas y las predichas por el clasificador. Esta métrica es la media harmónica entre la precisión y la sensibilidad:

$$F_1 = \frac{2 \cdot \mathrm{tp}}{2 \cdot \mathrm{tp} + \mathrm{fp} + \mathrm{tn}} = 2 \cdot \frac{\mathrm{precision} \cdot \mathrm{sensibilidad}}{\mathrm{precision} + \mathrm{sensibilidad}}. \tag{4}$$

Por último, otra técnica para visualizar el desempeño de un clasificador es evaluar la relación entre TPR y FPR a diferentes umbrales, esto es conocido como curvas ROC [6]. De la misma manera, se puede calcular el área bajo la curva ROC (AUC), a mayor área, mejor desempeño por parte del clasificador.

4. Propuesta

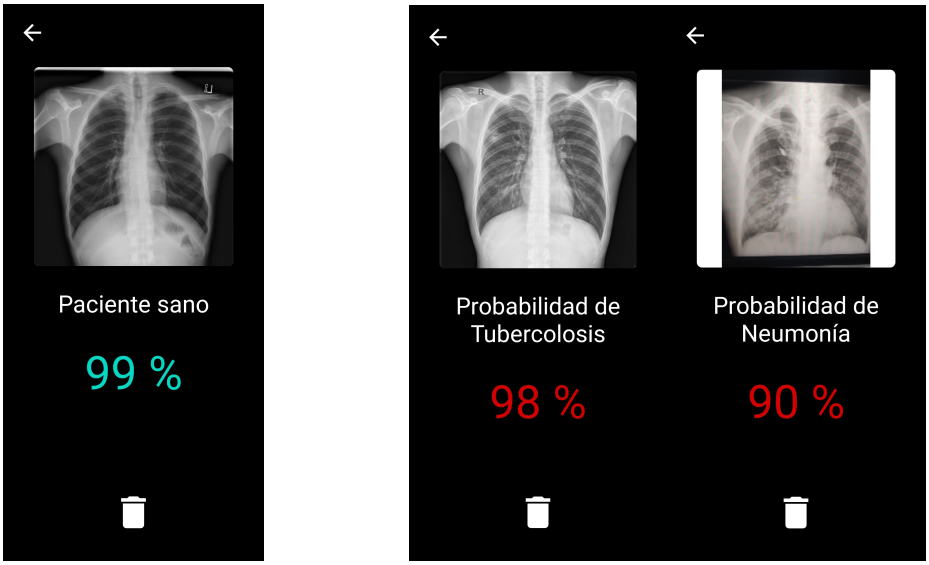
Para la clasificación de radiografías, se implementó el modelo convolucional llamado NanoChest-Net [11], que fue desarrollado exclusivamente para clasificar imágenes de estudios radiológicos de manera eficiente debido al tamaño del mismo (13.5 MB en su versión para móvil). Se entrenó dicho modelo en el banco de datos multiclase que incluye las distintas enfermedades y los pacientes sanos.

Una vez que la CNN fue entrenada, se convirtió el modelo a su versión de TensorFlow Lite para utilizarlo como modelo de inferencia dentro de la aplicación móvil XDApp (X-ray Diagnosis App).

La Figura 1 representa el esquema de implementación de TensorFlow Lite en kotlin, los datos de entrada se refieren a las imágenes que se convierten a mapa de bits, redimensionado a $1 \times 250 \times 250 \times 3$ para ser procesados por el intérprete que utiliza la CNN para generar la salida.

La salida es un tensor de tamaño 4×1 que contiene las probabilidades de que la imagen pertenezca a cada clase. Por lo tanto, se selecciona la clase con mayor probabilidad para mostrarla al usuario e indicar si el paciente tiene alguna enfermedad o está sano.

La aplicación móvil permite analizar imágenes adquiridas desde la cámara del dispositivo o cargar imágenes desde la galería. Al analizar las imágenes, la aplicación emitirá un prediagnóstico, indicando la probabilidad de que una enfermedad esté o no presente como se muestra en la Figura 2.



(a) Sano (b) Enfermos

Fig. 4. Capturas de pantalla del resultado de clasificación en la app móvil. a)Pacientes sanos; b)Pacientes enfermos de tuberculosis y neumonía.

5. Resultados y discusión

5.1. Experimentación

Se utilizó Python 3.11 como lenguaje de programación, se empleó la librería TensorFlow 2.9.1 [1] para el procesamiento de aprendizaje profundo con Keras como API de alto nivel y la biblioteca scikit-learn [17] para obtener las métricas de desempeño.

El entrenamiento de la CNN se realizó en un equipo con las siguientes características: procesador Intel Core i7 5930K; 16 GB RAM; tarjeta de gráficos discretos Nvidia RTX 2070S. Para el diseño y programación de la aplicación móvil se empleó como lenguaje kotlin y la biblioteca TensorFlow Lite para integrar el modelo entrenado usando Android Studio como entorno de desarrollo.

Las imágenes fueron normalizadas y redimensionadas al tamaño de entrada del modelo NanoChest-Net (250x250x3), además se aplicaron técnicas de aumento de datos con el fin de que el modelo trate de generalizar al entrenar con un mayor número de ejemplos diferentes. De tal forma, se aplicaron de manera aleatoria volteo horizontal; alejamiento-acercamiento, en un rango de 0.9 a 1.2; rotación aleatoria de ± 20 grados; traslación horizontal y vertical, con un factor de 0.2; y cambio de brillo en un rango de 0.80 a 1.05.

Para la parte del entrenamiento, se utilizaron los siguientes hiperparámetros: optimizador Adam; tasa de aprendizaje de 5×10^{-4}; 100 épocas; categorical cross-entropy como función de costo; penalización para el desbalance de las clases, calculado con la función compute class weight de scikit-learn; y tamaño de lote de 32.

5.2. Resultados de clasificación

Una vez entrenado el modelo de CNN, se obtuvieron las predicciones para el conjunto de prueba y se generó la matriz de confusión normalizada, para las 4 clases (Figura 3). De la matriz de confusión (Figura 3), se calcularon las métricas de desempeño mostradas en la Tabla 2.

En la Tabla 2, se puede observar que en la métrica de sensibilidad, el modelo es capaz de clasificar de una manera más efectiva los ejemplos de tuberculosis, con un valor de 0.8765; en segundo lugar, la neumonía, con un valor de 0.8460; seguido de los pacientes normales, con un valor de 0.8297; y finalmente, los pacientes con COVID-19 con 0.8144.

Sin embargo, al calcular TPR y FPR a diferentes umbrales, los valores de AUC son bastantes altos, con un valor de 0.9687 para tuberculosis; 0.9786 para COVID-19; 0.9601 para neumonía; y finalmente, 0.9518 para la clase normal. En la Figura 4, se puede observar un ejemplo de clasificación utilizando XDApp.

6. Conclusiones y trabajo a futuro

Se ha entrenado un modelo de red neuronal convolucional para la clasificación de radiografías de múltiples enfermedades. En concreto, se logró clasificar neumonía, tuberculosis, COVID-19 y pacientes sanos, usando el modelo NanoChest-Net con sensibilidades de 84.60 %, 87.65 %, 81.44 % y 82.97 %. A su vez, por ser un modelo con parámetros reducidos, permite utilizarlo dentro de una aplicación para dispositivos móviles que no necesariamente necesitan ser de gama alta.

La implementación de un modelo de clasificación en XDApp, permite tener un evaluador de radiografías móvil, que puede ser utilizado para consultar prediagnósticos, en situaciones donde no exista un médico radiólogo especialista. De la misma manera, XDApp puede ser utilizada como una herramienta para la enseñanza educativa en ciencias de la salud, donde el estudiante tenga dudas al observar una radiografía.

Es importante mencionar, que una de las desventajas que tiene el uso de la cámara para adquirir fotografías de radiografías se debe de hacer en un ambiente con iluminación controlada para evitar que los reflejos y efectos de luz, alteren el resultado de clasificación.

Como parte del trabajo futuro, es importante considerar ampliar el número de enfermedades que el modelo puede clasificar. Además, este tipo de aplicaciones podría ser implementado de manera remota en un servidor para ser accesible por medio de una computadora con acceso a internet e incluir envío de notificaciones con las imágenes y predicciones a los médicos encargados de realizar el diagnóstico.

Agradecimientos. Agradecemos al Instituto Politécnico Nacional por su apoyo para la realización de este trabajo; de igual manera, al apoyo del gobierno mexicano a través del Consejo Nacional de Ciencia y Tecnología (CONACYT).

Referencias

1. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., et al.: TensorFlow: Large-scale machine learning on heterogeneous systems (2015) doi: 10.48550/ARXIV.1603.04467
2. Anderson, K., Burford, O., Emmerton, L.: Mobile health apps to facilitate self-care: A qualitative study of user experiences. PLOS ONE, vol. 11, no. 5, pp. e0156164 (2016) doi: 10.1371/journal.pone.0156164
3. Atkinson, R. B., Sidey-Gibbons, C., Smink, D. S., Askari, R., Pusic, A. L., Cho, N. L., Robertson, J. M., Rangel, E. L.: Real-time student feedback on the surgical learning environment: Use of a mobile application. Journal of Surgical Education, vol. 80, pp. 817–825 (2023) doi: 10.1016/j.jsurg.2023.02.017
4. Cohen, J. P., Morrison, P., Dao, L.: COVID-19 image data collection (2020) doi: 10.48550/ARXIV.2003.11597
5. El-Shafai, W., El-Nabi, S. A., M. El-Rabaie, E. S. M., Ali, A. M., Soliman, N. F., Algarni, A. D., El-Samie, F. E. A.: Efficient deep-learning-based autoencoder denoising approach for medical image diagnosis. Computers, Materials & Continua, vol. 70, no. 3, pp. 6107–6125 (2022) doi: 10.32604/cmc.2022.020698
6. Fawcett, T.: An introduction to ROC analysis. Pattern Recognition Letters, vol. 27, no. 8, pp. 861–874 (2006) doi: 10.1016/j.patrec.2005.10.010
7. Gavrysh, V.: Pneumothorax binary classification task (2021) https://www.kaggle.com/datasets/volodymyrgavrysh/pneumothorax-binary-classification-task
8. Jaeger, S., Candemir, S., Antani, S., Wáng, Y. X. J., Lu, P. X., Thoma, G.: Two public chest X-ray datasets for computer-aided screening of pulmonary diseases. Quant Imaging in Medicine and Surgery, vol. 4, no. 6, pp. 475–477 (2014) doi: 10.3978/j.issn.2223-4292.2014.11.20
9. Kermany, D. S., Goldbaum, M., Cai, W., Valentim, C. C., Liang, H., Baxter, S. L., McKeown, A., Yang, G., Wu, X., Yan, F., Dong, J., Prasadha, M. K., Pei, J., Ting, M., Zhu, J., Li, C., Hewett, S., Dong, J., Ziyar, I., Shi, A., et al.: Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell, vol. 172, no. 5, pp. 1122–1131.e9 (2018) doi: 10.1016/j.cell.2018.02.010
10. Lucero-Mueses, J. E., Álzate-Mejía, O. A.: Aplicaciones móviles para el estudio de la anatomía humana. International Journal of Morphology, vol. 38, no. 5, pp. 1365–1370 (2020) doi: 10.4067/s0717-95022020000501365
11. Luján-García, J. E., Villuendas-Rey, Y., López-Yáñez, I., Camacho-Nieto, O., Yáñez-Márquez, C.: Nanochest-net: A simple convolutional network for radiological studies classification. Diagnostics, vol. 11, no. 5, pp. 775 (2021) doi: 10.3390/diagnostics11050775
12. OMS: Neumonía infantil (2022) who.int/es/news-room/fact-sheets/detail/pneumonia
13. OMS: Tuberculosis (2023) who.int/es/news-room/fact-sheets/detail/tuberculosis
14. PAHO: La neumonía es la causa principal de muerte de niños (2011) paho.org/es/noticias/11-11-2011-neumonia-es-causa-principal-muerte-ninos
15. PAHO: Brote de enfermedad por el Coronavirus COVID-19 (2023) paho.org/es/temas/coronavirus/brote-enfermedad-por-coronavirus-covid-19
16. Pasa, F., Golkov, V., Pfeiffer, F., Cremers, D., Pfeiffer, D.: Efficient deep network architectures for fast chest X-Ray tuberculosis screening and visualization. Scientific Reports, vol. 9, no. 1, pp. 6268–6268 (2019) doi: 10.1038/s41598-019-42557-4
17. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in python. Journal of Machine Learning Research, vol. 12, pp. 2825–2830 (2011) doi: 10.48550/ARXIV.1201.0490

18. Polsinelli, M., Cinque, L., Placidi, G.: A light CNN for detecting COVID-19 from CT scans of the chest. Pattern Recognition Letters, vol. 140, pp. 95–100 (2020) doi: 10.1016/j.patrec.2020.10.001
19. Siddiqi, R.: Efficient pediatric pneumonia diagnosis using depthwise separable convolutions. Computer Science, vol. 1, no. 6, pp. 343–343 (2020) doi: 10.1007/s42979-020-00361-2
20. Suetens, P.: Fundamentals of medical imaging. Second edition (2009)
21. Sutton, D.: Textbook of radiology and imaging (2003)
22. Varadarajan, V., Shabani, M., Ambale-Venkatesh, B., Lima, J. A. C.: Role of imaging in diagnosis and management of COVID-19: A multiorgan multimodality imaging review. Frontiers in Medicine, vol. 8 (2021) doi: 10.3389/fmed.2021.765975
23. Vialart-Vidal, M. N., Vidal-Ledo, M. J., Sarduy-Domínguez, Y., Delgado-Ramos, A., Rodríguez-Díaz, A., Fleitas-Estévez, I., Muñoz-Morejón, M., Gavilondo-Mariño, X., Pérez-Matar, R.: Aplicación de la eSalud en el contexto cubano. Revista Panamericana de Salud Pública, pp. 1–9 (2018) doi: 10.26633/RPSP.2018.19
24. Weigandt, W. A., Schardt, Y., Bruch, A., Herr, R., Goebeler, M., Benecke, J., Schmieder, A.: Impact of aneHealth smartphone app on quality of life and clinical outcome of patients with hand and foot eczema: Prospective randomized controlled intervention study. Journal of Medical Internet Research mHealth and uHealth, vol. 11, pp. e38506 (2023) doi: 10.2196/38506
25. Yacin-Sikkandar, M. Y., Sabarunisha-Begum, S., Alkathiry, A. A., Alotaibi, M. S. N., Manzar, M. D.: Automatic detection and classification of human knee osteoarthritis using convolutional neural networks. Computers, Materials and Continua, vol. 70, no. 3, pp. 4279–4291 (2022) doi: 10.32604/cmc.2022.020571

ISSN 1870-4069

Sistema de imagen táctil para la representación de texto e imágenes en relieve: Etapa de conversión

Oscar Daniel Martínez-Nambo, Guillermo Rey Peñaloza-Mendoza, Alicia Campos-Hernández

Instituto Tecnológico Superior de Pátzcuaro,
Ingeniería Biomédica,
México

{ciidtbiomedica, grey, acampos}@itspa.edu.mx

Resumen. Actualmente, existen tecnologías que apoyan a las personas con discapacidad visual, sin embargo, estas emplean herramientas sonoras lo que implica apartarlas de la lectura y la escritura, elementos esenciales en el proceso educativo. En el presente trabajo, se desarrolla la implementación de la tecnología de procesamiento digital de imágenes y textos para la creación del sistema de imagen táctil (TIS Tactile Imaging System), que tiene como objetivo brindar una herramienta que pueda mejorar la comunicación y, por tanto, el aprendizaje de las personas con discapacidades visuales. Esto se realiza mediante el diseño de un sistema capaz de procesar imágenes de texto alfanumérico o dibujos para obtener su equivalente en matriz Braille o en contornos, respectivamente, para posteriormente imprimir imágenes interpretables por tacto. En el caso del texto alfanumérico se realiza la detección del símbolo y se cambia por su equivalente en matriz Braille con respecto a una base de datos; en el caso de las imágenes, estas son binarizadas y posteriormente se aplica un filtrado para encontrar los contornos. De esta manera se busca dar solución a algunos problemas y mejorar la comunicación para personas con problemas visuales, fomentando la educación y dando oportunidades para un mejor desarrollo académico.

Palabras clave: Braille, discapacidad visual, imagen táctil, procesamiento de imágenes.

Tactile Imaging System for Embossed Text and Images Representation: Conversion Stage

Abstract. At present, there are technologies that support people with visual disabilities, however, they use sound tools, which implies separating them from reading and writing, essential elements in the educational process. In the present work, the implementation of digital image and text processing technology for the creation of the tactile image system (TIS) is developed, which aims to provide a tool that can improve communication and, therefore, learning. of people with visual disabilities. This is done by designing a system capable of processing alphanumeric text images or drawings to obtain their equivalent in Braille matrix or outlines, respectively, to subsequently print images that can be interpreted by touch. In the case of alphanumeric text, the detection of the symbol is carried out

and it is changed by its equivalent in Braille matrix with respect to a database; In the case of images, these are binarized and later a filter is applied to find the contours. In this way, it seeks to solve some problems and improve communication for people with visual problems, promoting education and giving opportunities for better academic development.

Keywords: Braille, visual disabilities, tactile image, image processing.

1. Introducción

Según la Organización Mundial de la Salud, el término discapacidad se define como "cualquier restricción o falta de capacidad para realizar una actividad de la misma manera o en el grado que se considera normal para un individuo" [1].

Datos del INEGI reflejan que en México la segunda discapacidad con mayor población es la visual y se estima que afecta a alrededor de 2.7 millones de personas [2], solamente en Michoacán se tienen cerca de 70 mil discapacitados visuales.

Durante mucho tiempo se apropió la idea que las personas ciegas eran incapaces de ser educadas debido a que su vida es excesivamente limitada en actividades como la lectura y escritura, sin embargo, en el siglo XVI se iniciaron los trabajos en búsqueda de una herramienta que permitiera integrar a las personas con discapacidad visual a la educación [3].

Actualmente existen múltiples herramientas que pretenden cubrir determinadas necesidades de las personas con discapacidad visual, tales como métodos, dispositivos y equipos que permiten el tratamiento, apoyo o compensación de la discapacidad, algunas de las más populares son la regla Braille, cuya función es apoyar la escritura con una organización por matrices Braille, la máquina de escribir Perkins, la cual es una máquina de escribir que en lugar de tener letras contiene las teclas para formar una matriz Braille, los traductores de texto a voz que permiten que el usuario con discapacidad visual pueda acceder a lo existente en un libro o en el celular, entre otras. Sin embargo, estas herramientas son costosas, escasas o, en el caso de los traductores de texto a voz, alejan al usuario de la lectura y la escritura.

Debido a esto, se plantea la necesidad de desarrollar herramientas auxiliares que sirvan para incrementar los recursos Braille existentes, debido a que los procesos de enseñanza-aprendizaje de este sistema no es fácil y se ven afectados por la falta de recursos, tales como libros, cuadernos y materiales didácticos, junto a la poca literatura y costosa tecnología existente.

2. Estado del arte

2.1. Impacto de la discapacidad visual

A nivel mundial, la Organización Mundial de la Salud (OMS) estima que aproximadamente 2200 millones de personas viven con alguna forma de deficiencia visual, de los cuales aproximadamente 1000 millones de personas padecen deterioro moderado o grave de la visión o ceguera [1]. Esta deficiencia tiene consecuencias que repercuten de manera muy preocupante en el desarrollo de las personas.

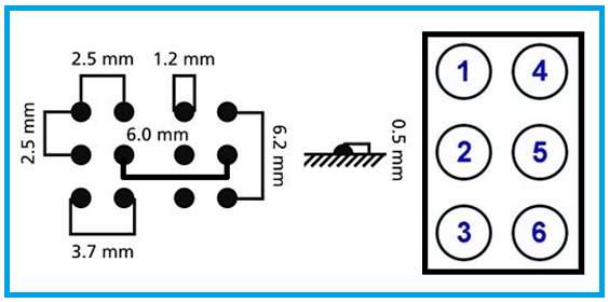


Fig. 1. Modelo de la matriz Braille y sus medidas

La sociedad como ente, es un ser que requiere de la interacción con su entorno, por lo cual, es indispensable el uso de nuestros sentidos para tener un desenvolvimiento integro en el día a día.

Esto se ve afectado si perdemos alguno de nuestros sentidos, principalmente la vista, ya que, una simple tarea como tomar un vaso de agua requiere ubicar en el espacio el vaso para tomarlo y llevarlo a nuestra boca.

En [4], se hace mención del gran impacto que tiene la discapacidad visual en las personas, por ejemplo: los niños pequeños que presentan deterioro grave desde temprana edad pueden sufrir retrasos en el desarrollo motor, lingüístico, emocional, social y cognitivo, con consecuencias para toda la vida.

El deterioro de la vista en una edad escolar disminuye el rendimiento académico. Así mismo, en adultos la participación en el mercado laboral y productivo se ve reducida, así como el incremento en la depresión.

2.1. Sistema Braille

Alrededor del año 1821, el oficial de artillería Charles Barbier inventó una técnica de escritura que servía para que los soldados se pudieran comunicar por la noche sin hablar. Esta técnica consistía en la creación de una matriz de puntos de 6x2 que representaba todos los sonidos del idioma francés.

Sin embargo, este método fue tomado y trabajado por Louis Braille y para el año 1824 lo redujo a una matriz 3x2, codificando los símbolos alfanuméricos en lugar de los sonidos del idioma [5].

El Braille es la representación de un alfabeto con letras, signos y números, a través de puntos en relieve que sirve como un sistema de lectura y escritura para personas ciegas. De manera general, la matriz Braille corresponde a una celda de seis puntos en relieve que representa un símbolo, se encuentra organizada como una matriz de tres filas y dos columnas, la cual, debe ser contenida con un tamaño tal que pueda caber en la yema de los dedos [6, 7], tal como se observa en la Figura 1.

Como referencia la Figura 2 muestra la correspondencia entre el alfabeto español y el alfabeto Braille.

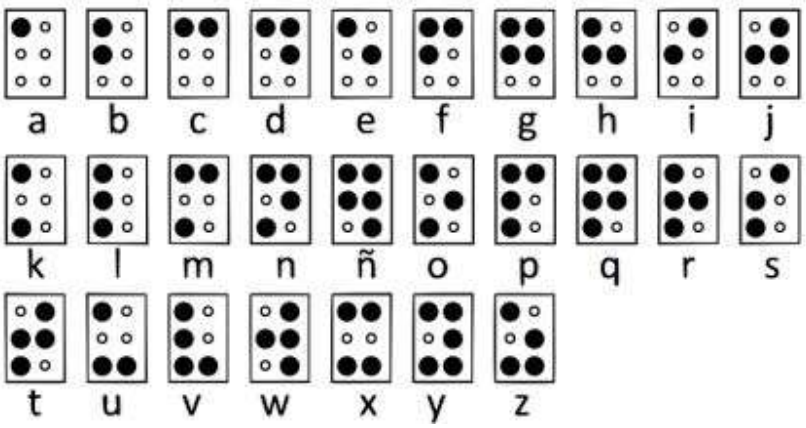


Fig. 2. Correspondencia entre el alfabeto español y el Braille

Para el proceso de enseñanza-aprendizaje de las personas con discapacidad visual existen dos herramientas fundamentales, los sistemas de traducción de texto a voz y el sistema Braille, sin embargo, el primero elimina las actividades de lectura y escritura, perjudicando el proceso cognitivo del aprendizaje [8].

Para emplear el sistema Braille como herramienta de enseñanza-aprendizaje, se requiere que los usuarios tengan acceso a elementos didácticos que les permitan entrenar su tacto ya que la lectura requiere emplear los dedos índices de ambas manos para desplazarlos al igual que en una lectura visual, por la línea de los símbolos de izquierda a derecha y de arriba abajo, leyendo matriz por matriz para asociarlas a su correspondiente símbolo y construir las palabras.

Como herramienta complementaria al sistema Braille, se han creado materiales didácticos que potencializan el uso del tacto, tales como imágenes en relieve, geometrías delimitadoras, juguetes con Braille, tablas de trabajo con canalillos para desplazar una plumilla y crear la representación de imágenes y símbolos, entre otros [9].

2.2. Tecnología para reproducción de imágenes a tacto

Las tecnologías existentes para percibir el mundo con el tacto han incrementado exponencialmente con el uso de la tecnología 3D, sin embargo, los recursos didácticos y herramientas en el proceso enseñanza-aprendizaje aún son escasos, ejemplo de esto son las impresoras Braille, las cuales existen desde hace años, pero su costo ha impedido que llegue a toda la población.

Actualmente, existen investigaciones y tecnologías en desarrollo que tratan de mejorar la precepción del mundo para las personas con discapacidad visual, ejemplos son: El sistema denominado Graille, fue desarrollado por investigadores de la Universidad de Tsinghua en China, el cual incluye una computadora y una pantalla compuesta por una matriz de 7200 puntos táctiles que permite representar imágenes para ser percibidas por el tacto.

Fig. 3. Sistema de cámara 2C3D.

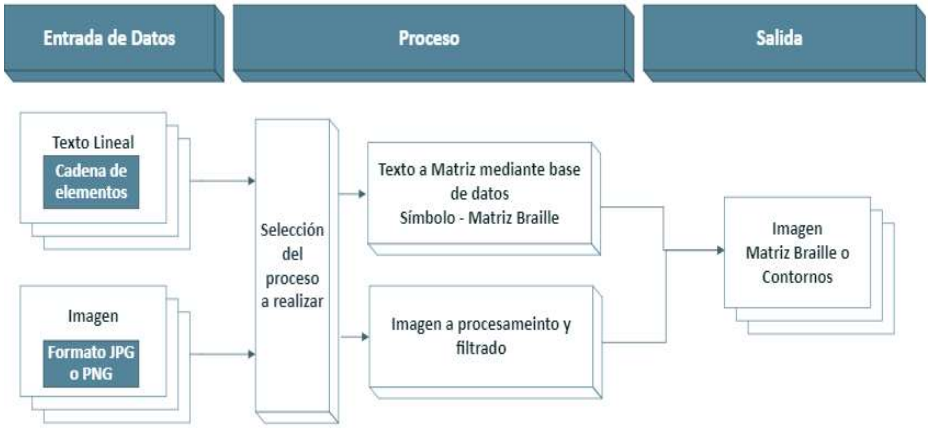


Fig. 4. Diagrama a bloques del funcionamiento del código.

La cámara 2C3D, Figura 3, desarrollada por Oren Geva del Shenkar College of Engineering and Design de Israel, permite a las personas con discapacidad visual percibir imágenes a través de pixeles 3D que cambian de posición de acuerdo a la imagen deseada [10].

En [11] se muestra el uso de la impresión 3D en la producción de materiales didácticos para la enseñanza de la historia del arte, la cultura popular y la literatura universal.

Considerando el desarrollo tecnológico realizado en Latinoamérica, la tecnología aplicada como apoyo para personas con discapacidad visual se basa en juguetes para enseñar formas geométricas u objetos, aplicaciones móviles para la traducción de texto a voz y algunos accesorios para múltiples, sin embargo, no son muchas empresas que invierten en nuevos desarrollos, una de ellas es Access Technology & Braille, mejor conocida como AT-BRAILLE, la cual tiene como objetivo dar a conocer tecnología auxiliar existente para personas con discapacidad, la cual es adaptada para Latinoamérica.

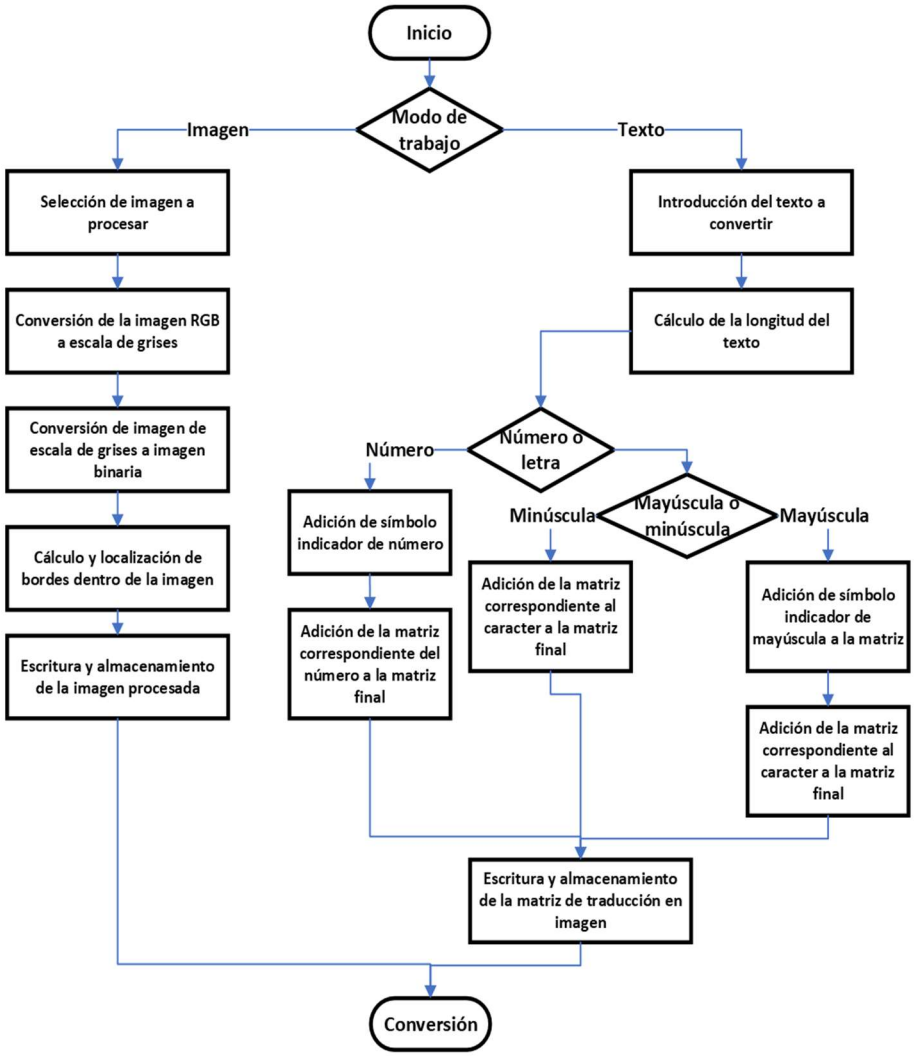


Fig. 5. Algoritmo de procesamiento del texto e imágenes.

En el proceso de enseñanza – aprendizaje de personas con discapacidad visual la tecnología es escasa, mucha de ella son trabajos de investigación y desarrollo tecnológico de universidades que han presentado interés en el tema y han colaborado con instancias de educación especial, ejemplo de esto lo tenemos en [12], donde en la Universidad Politécnica Salesiana de Ecuador se diseñó y desarrolló un prototipo electromecánico de línea Braille de bajo costo que tiene como función leer correos web, leer documentos PDF y convertirlos de texto a Braille o de texto a voz.

En [13] se presenta el diseño y fabricación de un dispositivo de lectoescritura del Braille que, junto con un tutor, se busca determinar y analizar los factores que facilitan el proceso de aprendizaje.

```
# Carácter a
a=np.array([
    [1,1,0,0,0],
    [1,1,0,0,0],
    [0,0,0,0,0],
    [0,0,0,0,0],
    [0,0,0,0,0],
    [0,0,0,0,0],
    [0,0,0,0,0],
    [0,0,0,0,0]])
```

```
#Carácter b
b=np.array([
    [1,1,0,0,0],
    [1,1,0,0,0],
    [0,0,0,0,0],
    [1,1,0,0,0],
    [1,1,0,0,0],
    [0,0,0,0,0],
    [0,0,0,0,0],
    [0,0,0,0,0]])
```

Fig. 6. Matriz numérica para la conversión de letra a Braille.

3. Planteamiento del problema y solución propuesta

Datos del INEGI muestran que más del 58% de la población de todo México presenta algún tipo de deficiencia visual [2], las tecnologías actuales ayudan a estas personas a integrarse a la sociedad, sin embargo, es preocupante el casi inexistente proceso de enseñanza - aprendizaje del Braille junto al limitado uso de materiales inclusivos en la educación, limita el desarrollo de las personas con discapacidad visual.

Con esto se tienen los problemas siguientes principales, la existencia casi nula de recursos didácticos para personas con discapacidad visual y los costos elevados de las herramientas para elaborar materiales Braille, imposibilitan brindarles una oportunidad a las personas con discapacidad visual de tener la experiencia de la lectura y la percepción de las imágenes del mundo.

Por lo tanto, se propone el desarrollo de un sistema capaz de tomar como entrada del mismo un texto escrito alfanuméricamente o una imagen, para obtener de salida una imagen que represente el texto en sistema braille o la imagen en bordes, de tal forma que cualquier dispositivo CNC pueda grabarlo sobre una superficie.

Con esto, se busca generar un impacto social positivo alto, ayudando a mejorar la percepción, comunicación y el aprendizaje de las personas con discapacidad visual.

4. Metodología de desarrollo

El desarrollo del proyecto, en esta etapa de conversión, se divide en dos partes, la primera consiste en realizar un código de programación capaz de hacer la transformación de un texto a una imagen de matrices Braille equivalente, mientras que la segunda es el procesamiento de imágenes para obtener una representación de contornos de la misma.

Como se puede observar en el diagrama de la Figura 4, el sistema toma como entrada un texto o una imagen a convertir, selecciona el procesamiento a realizar y entrega como salida una imagen que representa la transformación realizada.

De manera general, el algoritmo del procesamiento de texto e imágenes se muestra en la Figura 5, como se observa, el primer paso es determinar si se desea procesar un texto o una imagen, después se realiza las operaciones necesarias sobre el producto de entrada acuerdo al caso que se determine; si fuese texto se analiza la cadena de

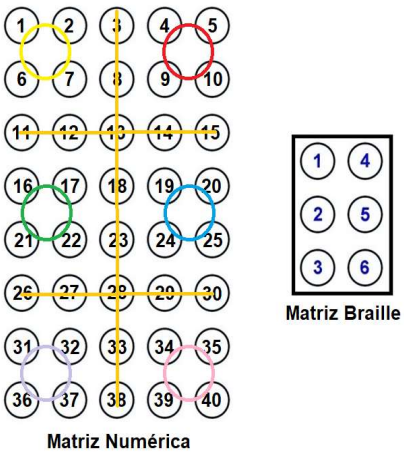


Fig. 7. Relación entre la matriz numérica y la matriz Braille.

```
for cont in range(0,len(text)):
    print(str(text[cont]) + '\n' + str(eval(text[cont])))
```

Fig. 8. Ciclo de impresión de caracteres.

caracteres para encontrar letras en minúscula, mayúscula y números, esto con la finalidad de asignarle una matriz Braille correspondiente a cada símbolo y para agregar los indicadores de cada caso como lo indica la nomenclatura Braille, para que el lector pueda saber si el siguiente símbolo es una letra minúscula, mayúscula o un número.

Si la entrada fuese una imagen que se desea procesar, a esta se le aplica una conversión a escala de grises, se binariza y se aplica un filtrado para detectar bordes/contornos.

El procesamiento de las imágenes de texto y figuras se realiza a través del uso de diversas matrices, en el primero cada matriz representa cada uno de los caracteres alfanuméricos utilizados en los textos y en el segundo caso la matriz corresponde a la propia imagen en contornos o bordes.

Por último, los resultados del procesamiento, se almacenan y se guardan en formato de imagen, la cual podrá ser tomada por un sistema CNC comercial y grabada para ser leída.

El código realizado para la lectura del texto, se encarga de leer una cadena proporcionada por el usuario y en base al análisis de cada uno de sus caracteres, selecciona una matriz correspondiente, en la Figura 6 se muestra un ejemplo de la matriz correspondiente para la letra a y b, como se puede observar se crea una matriz de 5 columnas por 8 renglones, cada punto de la matriz Braille está representado por 4 espacios en la matriz, tal como se observa en la Figura 7 los elementos 1, 2, 6 y 7 de la matriz numérica representan el elemento 1 de la matriz Braille, además la columna 3 y los renglones 3 y 6 son elementos para generar espacios entre los puntos de la matriz.

En el caso de los números, en Braille se agrega un identificador para indicar que en seguida se estarán leyendo números, esto es una matriz adicional, para esto se analizan

```
Texto a traducir:   abc
a
[[1 1 0 0 0]
 [1 1 0 0 0]
 [0 0 0 0 0]
 [0 0 0 0 0]
 [0 0 0 0 0]
 [0 0 0 0 0]
 [0 0 0 0 0]
 [0 0 0 0 0]]
b
[[1 1 0 0 0]
 [1 1 0 0 0]
 [0 0 0 0 0]
 [1 1 0 0 0]
 [1 1 0 0 0]
 [0 0 0 0 0]
 [0 0 0 0 0]
 [0 0 0 0 0]]
c
[[1 1 0 1 1]
 [1 1 0 1 1]
 [0 0 0 0 0]
 [0 0 0 0 0]
 [0 0 0 0 0]
 [0 0 0 0 0]
 [0 0 0 0 0]
 [0 0 0 0 0]]
```

Fig. 9. Ejemplo de cadena abc en matriz numérica.

```
Texto:
prueba 123
```

Fig. 10. Texto de entrada para prueba.

los elementos siguientes y anteriores de la cadena para asegurar que el indicador de número sólo se colocará al principio del mismo; caso similar con las letras en mayúsculas, se analiza la cadena por palabras para comprobar si alguna está escrita completamente en mayúsculas, si es así se utiliza el indicador correspondiente al inicio de la misma.

Los indicadores se agregan al inicio de una secuencia de números o letras mayúsculas, sin embargo, si no hay una secuencia y son símbolos alternados, estos indicadores se utilizarán en cada cambio existente.

Cada carácter alfanumérico que compone la cadena del texto de entrada, representa una matriz numérica que es obtenida de una base de datos, al tener la conversión de cada carácter, estas matrices se concatenan para obtener una matriz resultante equivalente al texto de entrada.

En la Figura 8 se muestra la sección del código del ciclo de impresión de las matrices Braille para una palabra o secuencia de letras, mientras que en la Figura 9 se muestran las matrices numéricas de una cadena de letras de entrada.

Como producto de salida, la matriz concatenada de respuesta se guarda como una imagen. Para el procesamiento de imágenes, se realiza la selección de la imagen de

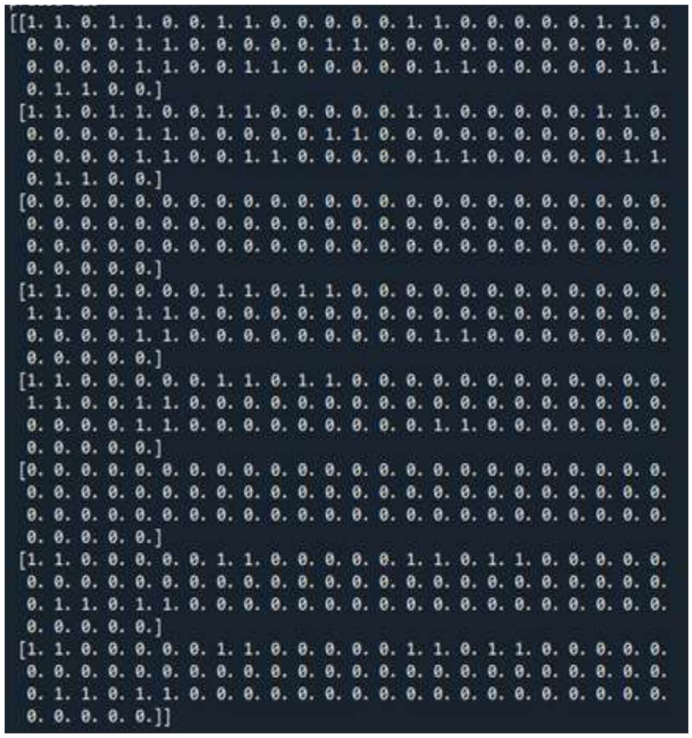


Fig. 11. Matriz numérica correspondiente al texto “prueba 123”.

entrada indicando el nombre del archivo con su extensión, y se procede a realizar el procesamiento de la imagen para calcular u obtener sus bordes.

Este proceso consiste en transformar la imagen a escala de grises, aplicar un filtro para eliminar el ruido de la imagen, detectar bordes e invertir el color de la imagen; para lo cual los métodos cvtColor, GaussianBlur, Canny y bitwise_not que ofrece la biblioteca implementada de OpenCV de Python.

El método cvtColor se encarga de transformar la imagen de RGB a escala de grises. Canny es un método que calcula automáticamente los bordes de una imagen, recibiendo solo tres parámetros: imagen, el valor mínimo del umbral y el valor máximo del umbral para generar una umbralización de histéresis.

Por último, la función bitwise_not que permite invertir los colores de la matriz que conforma la imagen.

En este sistema se utiliza un filtrado de tipo Gaussiano, este filtro permite dar más importancia a los píxeles que se encuentran más cerca del centro eliminando los puntos dispersos, para realizar este filtrado se utiliza el método GaussianBlur, el cual recibe como valores: la imagen y el tamaño del núcleo que se necesita, ya que debe hacerse en dos dimensiones, y como último valor el ancho de la curva de campana, sin embargo, al indicar el valor 0, OpenCv se encarga de calcular automáticamente este valor para el kernel elegido.

Posteriormente se muestra la imagen original y la procesada en el software para observar los cambios. La imagen se almacena en la computadora en una carpeta

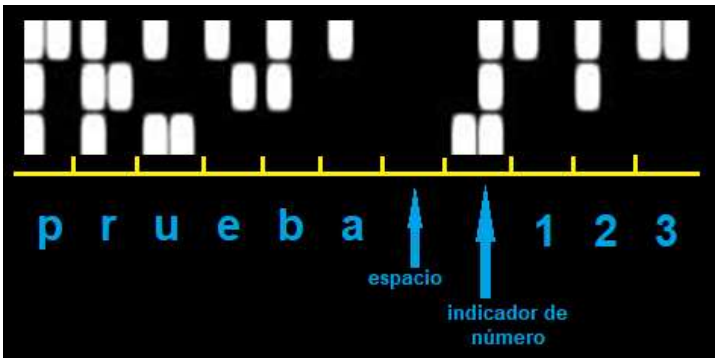


Fig. 12. Imagen en Braille equivalente al texto de entrada.

definida automáticamente por el usuario. Esta carpeta contiene todas las imágenes de salida, listas para ser enviadas a un sistema de CNC comercial para imprimir como relieve en papel.

Al obtener la imagen final procesada con las matrices Braille o con la imagen en bordes o contornos, se pretende como trabajo futuro, que esto se imprima de manera simple en papel para obtener el recurso táctil, para esto se trabajará con el diseño de una maquinaria que para la impresión tanto del Braille en escala real, así como la recreación de imágenes en contornos.

Esta maquinaria requiere trabajar en los tres ejes X, Y para crear la correspondencia y en Z para determinar lo alto del relieve, como base se usará el sistema de funcionamiento de las máquinas CNC que tomará como entrada la imagen procesada y realizará el relieve correspondiente.

5. Resultados

El prototipo se encuentra en una etapa temprana de desarrollo, teniendo como avance el sistema de conversión de texto o imagen a una imagen de matriz Braille o de contornos. Actualmente se permite la traducción de texto digital a código braille; Para lograrlo se utilizaron matrices, una matriz por cada letra y símbolo que se encuentra dentro del sistema braille.

Se analiza el texto que se ha dado para su traducción y en base a la ortografía se utilizarán las matrices correspondientes para cada letra, número o signo de puntuación. En la Figura 10 se muestra el texto “prueba 123” de entrada como prueba para la ejecución del código, en la Figura 11 se muestra la matriz obtenida que representa al texto introducido, cada sección representa un renglón de la matriz de conversión.

En la Figura 12 se muestra la imagen resultante con la matriz Braille de todo el texto. En otro código que se ha realizado a modo de prueba, utilizando la librería OpenCV, se logró el procesamiento de imágenes, que primero convierte la imagen a escala de grises para luego obtener el contorno de estas.

En la Figura 13 se muestra la imagen de entrada para la prueba del código y la imagen resultante para su impresión en relieve. Para el desarrollo de la interfaz gráfica se han investigado dos bibliotecas existentes para Python: PySimpleGUI y tkinter. Se

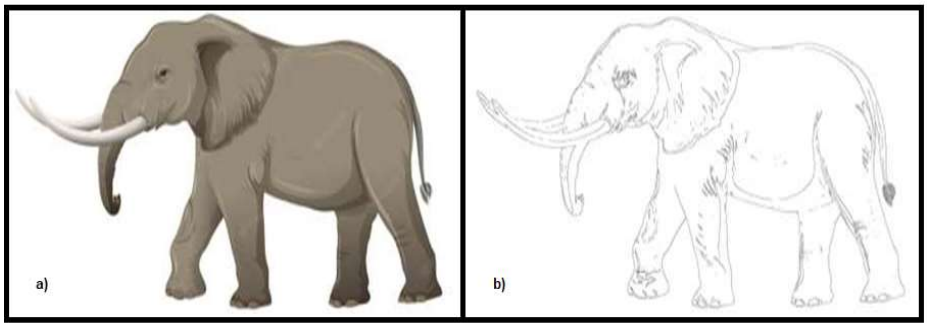


Fig. 13. Prueba con una imagen de elefante para obtener sus bordes o contornos, a) imagen original, b) imagen resultante del procesamiento.

continúa investigando sobre los beneficios de cada uno de estos con el fin de utilizar el adecuado para el proyecto.

En lo concerniente a la estructura física para la impresión, no se ha llevado a cabo, pero es la siguiente etapa de trabajo.

Para probar las transformaciones realizadas, se implementó un sistema CNC comercial, cambiando la herramienta por un punzón e introduciendo la imagen obtenida después del proceso.

No se ha empleado con usuarios finales que padezcan discapacidad visual. Como resultados de estas pruebas, con respecto al texto se tiene lo mostrado en la Tabla 1, el tamaño de los relieves de la matriz Braille correspondiente al texto de entrada son 23% más grandes, debido al punzón usado, al imprimir múltiples matrices se tiene un desplazamiento horizontal de una matriz con respecto a la otra.

Esto último es debido a la colocación superficial de la hoja de impresión, la cual no cuenta con una bandeja específica. Las imágenes de contorno fueron impresas en su totalidad de manera adecuada, sin embargo, como se ve en la Tabla 2 las imágenes con altos detalles no son bien procesadas o contienen mucha información que al obtener su relieve no son fácilmente detectable su forma.

6. Conclusiones y trabajo a futuro

El sistema desarrollado permite en primera instancia realizar el procesamiento de imágenes naturales y convertirlas a imagen de relieves, así mismo se logra realizar la acción de traducción de una forma óptima, llevando un texto a su equivalente en matriz Braille y guardar el resultado como imagen.

En términos de efectividad, el sistema funciona para letras y números, faltando implementar los caracteres especiales. En las imágenes, el proceso es efectivo cuando la imagen presenta un fondo sin detalles, ya sea color liso o con elementos simples.

Actualmente no se tiene la capacidad de una máquina de impresión, pero se tiene el sistema que permite tomar una entrada y obtenerla lista para imprimir en una máquina CNC comercial, destinando el sistema a satisfacer las necesidades en términos de educación, de las personas que sufren de discapacidad visual. El sistema, aun no

Tabla 1. Características de impresión de matriz Braille.

Letras	Números	Caracteres especiales	Error en Tamaño	Tamaño de cadena
A – Z y a - z	0 - 9	No aplica	23%	29 símbolos

Tabla 2. Características de impresión de la imagen de contornos.

Formato de entrada	Color de imagen	Tipo de fondo	Grosor de contorno
JPG y PNG	RGB	Liso	1.5mm

terminado, permite obtener recursos digitales adaptados en términos de Braille y relieves, que pudieran ser implementados en actividades de enseñanza – aprendizaje.

A futuro se pretende continuar con la investigación, adaptando una interface amigable al sistema de traducción, así como desarrollar el prototipo de la máquina de impresión.

Referencias

1. World health organization (WHO): Blindness and vision impairment (2022) www.who.int/en/news-room/fact-sheets/detail/blindness-and-visual-impairment.
2. Instituto nacional de estadística y geografía (INEGI): Discapacidad: Porcentaje de la población con algún tipo de discapacidad por grupo de edad (2020) www.inegi.org.mx/temas/discapacidad/
3. González-Saucedo, A. C., García-Heredia, F. J., Ramírez-Martínez, R.: Discapacidad visual. Cultura Científica y Tecnológica, vol. 10, no. 51 (2016) erevistas.uacj.mx/ojs/index.php/culcyt/article/view/954
4. Steinmetz, J. D., Bourne, R. R., Briant, P. S., Flaxman, S. R., Taylor, H. R., Jonas, J. B., Abdoli, A. A., Abrha, W. A., Abualhasan, A., Abu-Gharbieh, E. G., Adal, T. G., Afshin, A., Ahmadieh, H., Alemayehu, W., Alemzadeh, S. A., Alfaar, A. S., Alipour, V., Androudi, S., Arabloo, J., Arditi, A. B.: Causes of blindness and vision impairment in 2020 and trends over 30 years, and prevalence of avoidable blindness in relation to VISION 2020: The right to sight: An analysis for the global burden of disease study. The Lancet Global Health, vol. 9, no. 2, pp. e144–e160 (2021) doi: 10.1016/s2214-109x(20)30489-7
5. Liesen, B.: El braille: Origen, aceptación y difusión. Entre dos mundos, Revista de traducción sobre discapacidad visual, no. 19, pp. 5-35 (2002)
6. Martínez-Liébana, I. y Polo-Chacón D.: Guía didáctica para la lectoescritura braille. Organización Nacional de Ciegos Españoles (2004)
7. Discapnet: El alfabeto Braille. Fundación Once. www.discapnet.es/innovacion/productos-apoyo/alfabeto-braille (2023)
8. Baciero, A., Perea, M., Gómez, P.: Tocando tus palabras: Por qué la lectura braille es especial. Ciencia Cognitiva, vol. 13, no. 2, pp. 54–57 (2019)
9. Fuentes-Nieves, F. M.: Diseño de imágenes para ciegos, material didáctico para niños con discapacidad visual. Universitat Politècnica de València, Departamento de Dibujo (2014) doi: 10.4995/Thesis/10251/37882
10. Oregan Geva industrial design: 2C3D tactile camera for the blinds (2018) https://www.orengeva.com/2c3d

11. Martín-Blas, Á. D.: La impresión de figuras en 3D como incentivo a la lectura para personas con discapacidad visual. Integración: Revista digital sobre discapacidad visual, no. 75, pp. 184–203 (2019)
12. Cabrera-Hidalgo, J. C.: Diseño y desarrollo de un prototipo de línea Braille de bajo costo para personas no videntes en el marco de Cátedra UNESCO "Tecnologías de apoyo para la inclusión educativa" de la Universidad Politécnica Salesiana. Universidad Politécnica Salesiana, Ecuador (2019)
13. Hernández-Suarez, C. A., Jiménez-Hernández, L. A.: Prototipo de tecnología en asistencia para la enseñanza del braille. Redes de Ingeniería, vol. 2, no. 2, pp. 105–115 (2012) doi: 10.14483/2248762x.7168

ISSN 1870-4069

Eliminación de ruido en sonidos cardíacos mediante técnicas de aprendizaje profundo

Cristóbal González Rodríguez[1], Miguel A. Alonso Arévalo[2], Eloísa García Canseco[1]

[1] Universidad Autónoma de Baja California,
Facultad de Ciencias,
México

[2] Centro de Investigación Científica y de Educación Superior de Ensenada,
Departamento de Electrónica y Telecomunicaciones,
División de Física Aplicada,
México

{a351269, eloisa.garcia}@uabc.edu.mx,
aalonso@cicese.edu.mx

Resumen. Las enfermedades cardiovasculares son la principal causa de mortalidad en todo el mundo. La auscultación cardiaca es un método de diagnóstico prometedor; sin embargo, uno de sus principales inconvenientes es que es altamente propensa al ruido durante la grabación del sonido, lo que dificulta el diagnóstico. En este trabajo proponemos un algoritmo de eliminación de ruido para las señales de audio cardíaco. El ruido se elimina en la representación tiempo–frecuencia de la señal. Específicamente, calculamos la transformada de Fourier de tiempo corto (STFT) de la señal de FCG contaminada y entrenamos una red neuronal de tipo U-Net para que reconozca los sonidos cardíacos, ya sean normales o patológicos, del ruido. En nuestras pruebas, el método propuesto muestra un alto desempeño incluso en escenarios altamente desfavorables, ya que puede eliminar el ruido de una señal FCG contaminada con una relación señal a ruido (SNR) de -5 dB con mejoras promedio del orden de $\approx$15 dB.

Palabras clave: Fonocardiograma, transformada de Fourier, red neuronal convolucional, separación de fuentes.

Noise Removal in Heart Sounds Using Deep Learning Techniques

Abstract. Cardiovascular diseases are the leading cause of mortality worldwide. Cardiac auscultation is a promising diagnostic method; however, one of its main drawbacks is that it is highly susceptible to noise during sound recording, which hinders diagnosis. In this study, we propose a noise removal algorithm for cardiac audio signals. The noise is eliminated in the time-frequency representation of the signal. Specifically, we calculate the Short-Time Fourier Transform (STFT)

of the contaminated FCG signal and train a U-Net neural network to recognize cardiac sounds, whether they are normal or pathological, in the presence of noise. In our tests, the proposed method demonstrates high performance even in highly unfavorable scenarios, as it can remove noise from a contaminated FCG signal with a signal-to-noise ratio (SNR) of -5 dB, with average improvements of $\approx$15 dB.

Keywords: Phonocardiogram, Fourier transform, convolutional neural network, source separation.

1. Introducción

La fonocardiografía (FCG) es la representación gráfica de los sonidos producidos por el corazón y tradicionalmente ha generado mucho interés por el potencial que tiene como herramienta para el diagnóstico clínico. La señal de FCG proporciona información sobre la duración, la frecuencia y otros parámetros importantes de los sonidos cardíacos para determinar la funcionalidad y el estado actual de las válvulas cardíacas [14].

La identificación de síntomas patológicos mediante la auscultación de sonidos cardíacos con la ayuda de un estetoscopio es una gran habilidad y adquirirla es una tarea difícil que puede llevar muchos años de práctica clínica. Además, el oído humano tiene limitaciones fisiológicas para percibir completamente los sonidos producidos por el corazón ya que la mayor parte de la energía del FCG se encuentra por debajo del umbral de audición [7].

Las enfermedades cardiovasculares son la primera causa de mortalidad en México y en el mundo [18, 24]. Aunque existen muchas técnicas modernas de diagnóstico, como el electrocardiograma (ECG), la resonancia magnética (RM) o el ecocardiograma, la auscultación cardiaca es probablemente el método no invasivo más económico, práctico y rápido. Los recientes avances en informática médica en combinación con la cada vez mayor capacidad de procesamiento de los dispositivos electrónicos han impulsado el análisis automático de las señales de sonido cardíaco.

El objetivo principal de este análisis es clasificar con precisión la presencia o ausencia de sonidos patológicos en el ciclo cardíaco [14]. Si se confirma la presencia de un evento de este tipo, lo ideal sería que un sistema automatizado también sea capaz de identificar el tipo de patología. La presencia de ruido en la señal de FCG es uno de los problemas más frecuentes durante la auscultación cardiaca. Este problema es particularmente delicado para un sistema automático ya que no es capaz de discriminar entre verdaderos sonidos cardíacos y espurios.

Las fuentes de ruido pueden ser numerosas y de naturaleza muy distinta: sonidos originados en el entorno, como el habla o el uso de aparatos cercanos al sitio de auscultación; por sonidos fisiológicos, como los gástricos y respiratorios; o por la fricción producida entre el estetoscopio y la piel.

Bajo estos escenarios, realizar un diagnóstico de enfermedades resulta especialmente difícil. La mayoría de los métodos de eliminación de ruido del FCG presentados en la literatura se evalúan contaminando una señal de FCG con ruido blanco Gaussiano aditivo (AWGN por sus siglas en inglés).

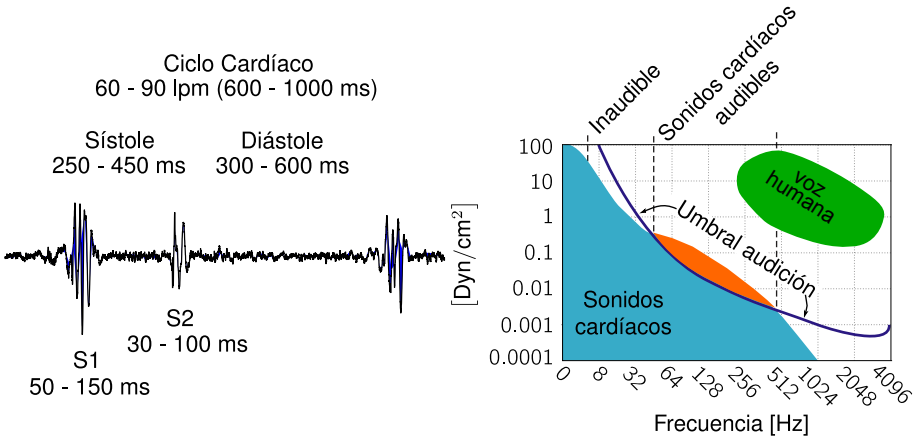


Fig. 1. Características principales de la señal de fonocardiograma (FCG).

Algunos otros trabajos también consideran la contaminación de la señal utilizando ruido rosa Gaussiano aditivo. En ambos casos, una vez eliminado el ruido de la señal de FCG contaminada, se procede a compararla con la señal original.

Tradicionalmente, la métrica más utilizada en la literatura para medir el rendimiento de un sistema de eliminación de ruido es la relación señal a ruido (SNR por sus siglas en inglés). Sin embargo la SNR tiene el problema de ser altamente susceptible a problemas de escalamiento debido a la energía de la señal.

Dicho escalamiento puede aumentar artificialmente la SNR resultante, abriendo así la posibilidad de obtener resultados engañosos que no representen adecuadamente el verdadero rendimiento de un método de eliminación de ruido. Un análisis a profundidad sobre este tema y posibles soluciones se presenta en [13].

1.1. Antecedentes

Una señal de FCG normal consta de dos sonidos cardíacos fundamentales llamados S1 y S2. Estos sonidos se producen durante el ciclo cardíaco cuando las válvulas semilunar y auriculoventricular se cierran [2]. La sístole (contracción), que marca el inicio de S1, y la diástole (relajación), que marca el inicio de S2, comprenden el comportamiento cuasi periódico del ciclo cardíaco. Existen dos periodos de silencio en los individuos sanos, ya que estos sonidos sólo comprenden una pequeña parte de cada sección del ciclo cardíaco: los silencios sistólico (s-Sys) y diastólico (s-Dia).

Lo más frecuente es que los ciclos cardíacos de individuos con afecciones cardíacas contengan ruidos adicionales, como sonidos S3 y S4, soplos, fricciones o chasquidos. Los sonidos cardíacos S1 y S2 suelen durar entre 25 y 150 ms, y la mayor parte de su contenido espectral se sitúa entre 24 y 144 Hz [5]. La duración de las patologías (chasquidos, fricciones y soplos) varía significativamente a lo largo del ciclo cardíaco, y su contenido espectral se sitúa entre 25 Hz y 700 Hz [6]. En la Figura 1 se resumen las características principales de la señal de FCG y en la Figura 2 se presentan dos ejemplos correspondientes a a) un sonido cardíaco normal y b) un sonido cardíaco anormal.

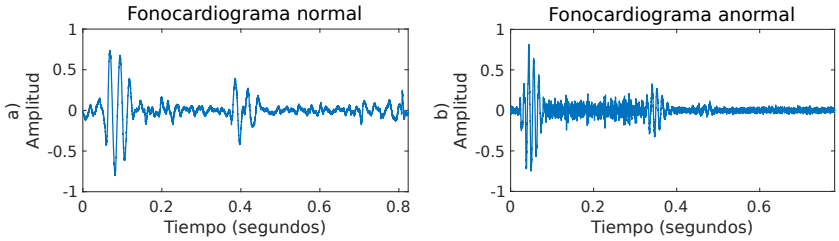


Fig. 2. Dos ejemplos de sonidos cardíacos, a) presenta el sonido de un corazón normal, mientras que b) presenta un sonido correspondiente a una persona con una patología cardíaca.

Se ha buscado eliminar el ruido en de la señal de FCG abordando el problema de múltiples maneras. Uno de los enfoques más populares es el uso de la transformada ondeleta discreta (DWT por sus siglas en inglés), el cual consiste en descomponer la señal en coeficientes ondeleta. A lo largo del tiempo se han propuesto variantes y mejoras de este método [11, 3, 10, 9, 17].

Este enfoque parte de la idea de que el ruido puede reordenarse en coeficientes en distintas bandas de frecuencia y que es posible aislar el sonido FCG mediante el uso de umbrales que eliminen los coeficientes correspondientes al ruido [16]. Se han realizado varias mejoras a este método básico, como probar diferentes ondeletas madre así como niveles de descomposición [16, 11, 3, 9]; evaluando los modelos propuestos con diferentes tipos de ruido [11, 10].

En otros casos, implementando redes neuronales para reconstruir la señal con un umbral adaptativo según los coeficientes obtenidos [10]. Más recientemente hay variantes que utilizan la transformada ondeleta síncrona [8] (SSWT por sus siglas en inglés) como la propuesta de [17]. Otro método reciente se basa en el algoritmo de mínimos cuadrados (LMS por sus siglas en ingés) [19]. Este enfoque utiliza un filtro adaptativo y consigue eliminar la mayor parte del ruido introducido en la señal.

Algunos ejemplos mencionados anteriormente implementan redes neuronales para mejorar el rendimiento de sus algoritmos [10], pero la aplicación de la inteligencia artificial en la eliminación del ruido presente en la señal de FCG aún puede ser explorada más a fondo, como se muestra a continuación.

1.2. Enfoque propuesto

La inteligencia artificial (IA) es una herramienta moderna que ha dado grandes resultados en numerosos ámbitos. Tal vez una de las aplicaciones más sobresalientes de la IA ha sido en el procesamiento de imágenes. En particular, la implementación de redes neuronales convolucionales (CNN por sus siglas en inglés) ha mostrado ventajas significativas sobre los métodos tradicionales para la eliminación de ruido de imágenes [25]. Algunas técnicas desarrolladas para imágenes han sido adaptadas para procesar señales de audio.

Un ejemplo notable de una arquitectura de CNN fue propuesta en [20], y se denomina U-Net. Este método fue originalmente propuesto para la segmentación de imágenes biomédicas, sin embargo variantes de la arquitectura U-Net han sido exitosamente utilizadas en la separación de fuentes de sonido.

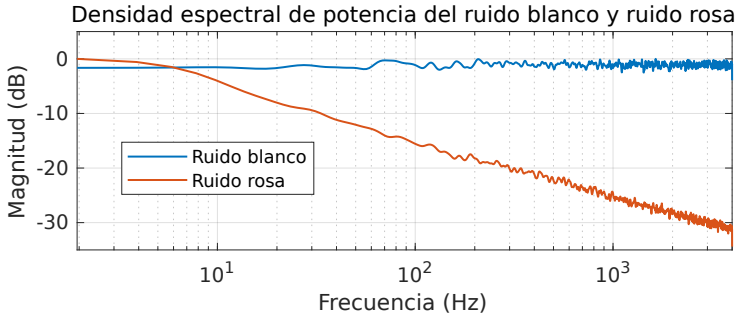


Fig. 3. Densidad espectral de potencia normalizada de los dos tipos de ruido utilizados en este trabajo.

Específicamente, en desagregar una señal grabada mediante un solo micrófono en las múltiples fuentes que la conforman, como se muestra en [4, 12]. En estos dos trabajos, las voces y los diferentes instrumentos de una canción se separan en diferentes formas de onda. Estos métodos funcionan obteniendo primero los espectrogramas del audio original mediante la transformada de Fourier de tiempo corto (STFT por sus siglas en inglés).

Los espectrogramas se utilizan después para entrenar una U-Net que identifique un tipo específico de sonido; esto se hace procesando los espectrogramas para crear una máscara. La máscara contendrá la información sobre qué partes del espectrograma corresponden a la señal que se desea aislar y cuáles al ruido; multiplicando la máscara por el espectrograma inicial se eliminará el ruido.

Dado que la U-Net está entrenada para identificar un tipo específico de sonido, el tipo de ruido con el que está contaminada la señal es irrelevante para la separación. En el presente trabajo mostramos cómo adaptar esta técnica de separación de fuentes al contexto de eliminación del ruido aditivo Gaussiano y rosa en la señal de FCG.

2. Metodología

Esta sección describe la metodología para la eliminación de ruido en señales de FCG y está divida en tres partes. Primeramente se describe la base de datos de sonidos cardíacos utilizados y los tipos de ruido utilizados para contaminarlas. En la segunda parte se describe brevemente el algoritmo de la Transformada de Fourier de Tiempo Corto (STFT por sus siglas en inglés) mediante el cual se convierten los sonidos en imágenes. En la tercera parte se describe la arquitectura de la U-Net y el proceso de entrenamiento de la red.

2.1. Base de datos y tipos de ruido

Como fuente de sonidos cardíacos se utilizó la base de datos propuesta en [22]. Esta base de datos pública es altamente popular en el área de análisis del FCG y se caracteriza por tener señales de sonidos cardíaco particularmente limpias. El uso de señales de sonido limpias permite un mejor entrenamiento de la red.

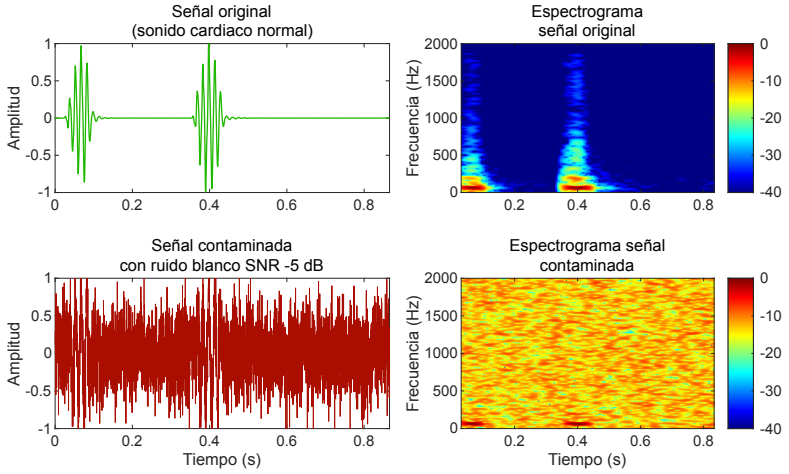


Fig. 4. Ejemplo de las imágenes de espectrograma que se utilizan para el entrenamiento de la red.

Este conjunto de datos tiene cinco categorías de señales: señales con sonidos de FCG normales (N) y cuatro señales con sonidos de FCG patológicos, que incluyen estenosis aórtica (EA), regurgitación mitral (RM), estenosis mitral (EM) y soplo en sístole (MVP). Estas cinco categorías contienen 200 grabaciones cada una; todas las señales se componen de tres ciclos cardíacos completos con una duración del FCG de aproximadamente tres segundos.

En total, la base de datos contiene 1,000 sonidos cardíacos muestreados a 8 kHz. En procesamiento de señales, es relativamente común utilizar ruido blanco aditivo como modelo para imitar el efecto de muchos procesos aleatorios que se dan en la naturaleza y que afectan el rendimiento de un sistema.

El concepto de ruido blanco se refiere a la idea de que tiene una potencia uniforme en toda la banda de frecuencias del sistema de información. En este trabajo se propone como primer tipo de señal de interferencia el uso de ruido blanco Gaussiano aditivo, porque tiene una distribución normal en el dominio del tiempo con una media de cero.

El segundo tipo de señal de interferencia utilizado en este trabajo es el ruido rosa. Este tipo de ruido es útil para aplicaciones de sonido y sistemas de audio ya que muchos sonidos musicales y naturales tienen espectros que contienen todas las frecuencias audibles, pero que disminuyen en intensidad a una razón de $\approx$3 dB por octava en función de la frecuencia (f) siguiendo un comportamiento de la forma $1/f$. La Figura 3 muestra la densidad espectral de potencia normalizada de los dos tipos de ruido utilizados en este trabajo.

2.2. Transformada de Fourier de tiempo corto

Para entrenar la red neuronal lo mejor es proporcionar tanta información sobre el comportamiento de la señal como sea posible. En [4, 12], los autores sugieren utilizar la transformada de Fourier de tiempo corto (STFT) para generar los espectrogramas de las señales de audio y utilizarlos como imágenes para entrenar el modelo.

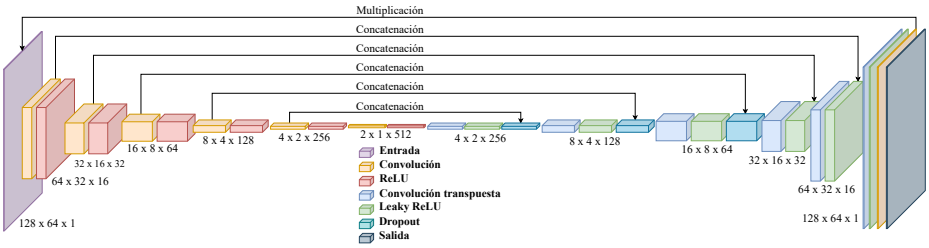


Fig. 5. Arquitectura de la U-Net con las dimensiones obtenidas a través de cada capa.

Este enfoque ha demostrado un excelente rendimiento en problemas de separación de fuentes, ya que contiene una representación precisa de la información temporal y frecuencial de la señal; en el presente trabajo, adoptamos el mismo enfoque.

Existen representaciones tiempo-frecuencia más precisas que la STFT. Sin embargo, en la práctica la STFT tiene una alta popularidad por múltiples razones, entre ellas su sencillez, velocidad de cálculo y un rendimiento altamente satisfactorio. En la práctica, las grabaciones de FCG que analizamos son de tiempo discreto, y los datos a transformar, que se muestrean en el tiempo, se procesan mediante la STFT discreta.

Un grupo de muestras forma un segmento o bloque de tiempo, y para cada segmento se calcula la transformada discreta de Fourier (DFT). El resultado es un vector complejo que se añade como columna a una matriz, que registra la magnitud y la fase de todos los segmentos de tiempo y bins de frecuencia. La STFT discreta de una señal s se calcula como sigue [21]:

$$X\left(k,m\right)=\sum_{n=-\infty}^{\infty}x\left(n\right)w\left(n-mH\right)e^{-\imath 2\pi nk/N}, \tag{1}$$

donde m representa cada marco temporal a lo largo de la señal de la que se calcula la STFT, y como se ha mencionado anteriormente, para cada marco temporal m se calcula localmente la transformada discreta de Fourier (DFT por sus siglas en inglés) a través de una señal $x\left(n\right)$, obteniendo un vector de frecuencias $X\left(k\right)$.

El conjunto de intervalos de frecuencia se define como $f_k = kf_s/N$, para $k = 0,1,2,\ldots,N/2$, siendo f_s la frecuencia de muestreo de la señal y N el número de muestras temporales en la DFT; n representa cada muestra a través de la señal local $x\left(n\right)$; H la longitud de salto entre dos segmentos temporales consecutivos m y $m+1$ que delimitan $x\left(n\right)$, con $m = 0,1,2,\ldots,M-1$, donde M es el número total de segmentos temporales en que se divide la señal original. Cada DFT es ponderada con la función de ventana de Hann:

$$w\left(n'\right)=0.5-0.5\cos\left(\frac{2\pi n'}{L_w-1}\right),\ 0\leq n'\leq L_w-1. \tag{2}$$

De longitud L_w muestras. La forma de la matriz resultante $X\left(k,m\right)$ sería entonces de dimensiones $K\times M$, con $K = N/2+1$ y $M = \lceil\left(S+R+1\right)/H\rceil$, donde $R=\left(-\left(S-L_w\right)\ \%H\right)\ \%L_w$, $\lceil\cdot\rceil$ representa la función techo (ceil en inglés) y $\%$ el residuo de la división; siendo $S = Tf_s$ el número total de muestras que conforman la señal original.

Tabla 1. Desempeño promedio del algoritmo de eliminación de ruido. La métrica utilizada es la SI-SDR para una validación cruzada de 10 iteraciones, los tipos de ruido considerados son blanco y rosa.

Ruido de entrenamiento	Tipo/Nivel de ruido (dB)	-5	0	5	10
Blanco	**Blanco**	9.93	13.41	17.21	20.71
	Rosa	-0.57	4.38	9.20	13.21
Rosa	**Blanco**	1.76	9.37	15.29	19.15
	Rosa	7.28	11.28	14.86	18.19

Los espectrogramas de las señales limpias y contaminadas fueron necesarios para entrenar la red. Además, para ajustar las imágenes (espectrogramas) a las dimensiones de entrada de la red, pero también para reducir la carga computacional de procesar miles de espectrogramas, hubo que ajustar su tamaño, eligiendo una FFT de tamaño de 256 muestras, longitud de salto de 32 muestras y una longitud de ventana de 128 muestras.

Utilizando estos parámetros obtuvimos los espectrogramas que se muestran en la Figura 4. Las gráficas presentadas en dicha figura también ilustran el tipo de imágenes PCG ruidosas que se alimentan a la red neuronal, en este ejemplo en particular, con señales contaminadas a -5 dB de relación señal a ruido (SNR por sus siglas en inglés).

Durante el entrenamiento de la red, solamente se utiliza la magnitud de la STFT. La fase de cada espectrograma obtenida durante el cálculo de la STFT se almacena para su reconstrucción posterior. A continuación, se procede a procesar los espectrogramas con la red entrenada para obtener la máscara. Luego se multiplican elemento por elemento con los espectrogramas originales. Finalmente se utiliza la fase para reconstruir el audio mediante la transformada de Fourier inversa.

2.3. Arquitectura de la U-Net

En este trabajo, utilizamos una red neuronal convolucional con una arquitectura U-Net similar a la que se propuso originalmente en [12]. La Figura 5 ilustra la arquitectura U-Net que utilizamos. Las imágenes de entrada y salida de la red tienen un tamaño de $(128, 64, 1)$.

Cada capa de convolución 2D y de transposición 2D utiliza un tamaño de núcleo de 5×5 y pasos de 2×2, rellenando con ceros después de cada convolución. Para la capa Leaky ReLU, se utilizó un parámetro $\alpha = 0.2$. Antes de multiplicar la máscara de salida por la entrada, sigue una última capa de convolución 2D con un tamaño de núcleo de 4×4, una tasa de dilatación de (2,2) y una función de activación de sigmoide.

Para entrenar la red, es necesario un conjunto de imágenes de entrada lo más amplio y diverso posible. En nuestro caso, las imágenes utilizadas son las obtenidas a partir del valor absoluto de la STFT (i.e., el espectrograma) de las señales de FCG de tamaño $K \times M$.

En procesamiento de imágenes, cuando se consideran matrices multidimensionales, las imágenes suelen tener un parámetro de tercera dimensión que representa el número de canales de color de la imagen. El valor típico es tres cuando nos referimos a imágenes de color verdadero o RGB (rojo, verde y azul).

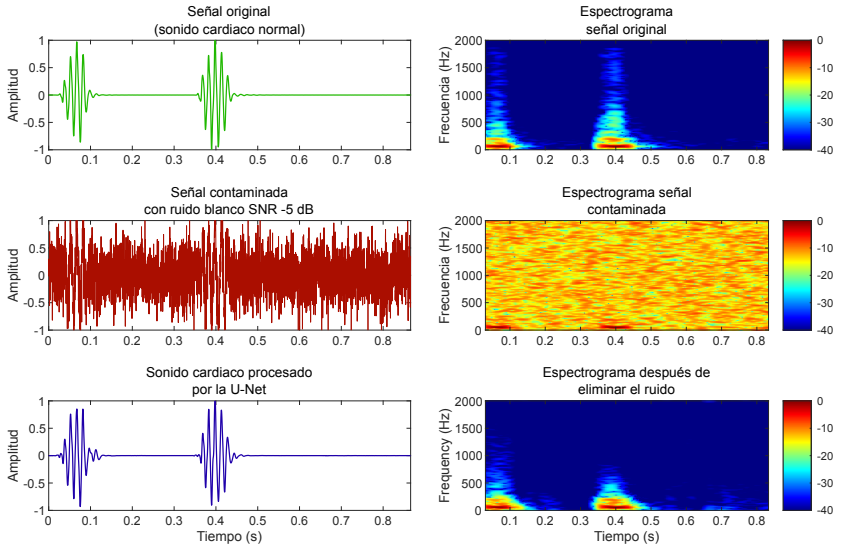


Fig. 6. Ejemplo del funcionamiento de la red para limpiar una señal de FCG normal contaminada con ruido blanco a -5 dB de SNR.

Sin embargo, como estamos utilizando espectrogramas compuestos por una matriz bidimensional, podemos considerarlos como si fueran imágenes con un solo canal de color; es decir, el eje de tercera dimensión tiene una longitud de uno.

La forma de cada imagen de espectrograma se fijó entonces de la siguiente manera $(128, 64, 1)$, como se muestra en la Figura 5. La red original propuesta en [12] se implementó considerando imágenes más grandes generadas por señales de audio de alta fidelidad (Hi-Fi).

Sin embargo, nuestra versión modificada también muestra un rendimiento notable para la eliminación de ruido en las señales de FCG que tienen un contenido espectral de tipo de pasa-bajas. Dado que las dimensiones de entrada de la red son 128×64, cada espectrograma se dividió en bloques de tiempo que coinciden con las dimensiones de entrada de la red.

De los mil sonidos de FCG, 900 audios fueron seleccionados aleatoriamente a partir de las de las cinco clases (una de FCG normal y cuatro tipos de FCG patológicos) para el entrenamiento de la red y se generaron más de cinco mil bloques de espectrograma utilizados únicamente para el entrenamiento. Internamente, las bibliotecas de inteligencia artificial que utilizamos dividieron las mil señales de la siguiente manera:

El 72 % de las señales se utilizaron para el entrenamiento de la red, el 18 % para la validación durante el entrenamiento, y el último 10 % se utilizó para la evaluación final de prueba de la red entrenada, ya que se trataba de datos que no había sido previamente vistos por la red.

Seleccionamos ADAM como algoritmo de optimización, y los parámetros de la red que mostraron el mejor rendimiento de convergencia para la eliminación de ruido fueron una tasa de aprendizaje de 0,0001 y un tamaño de lote de 64.

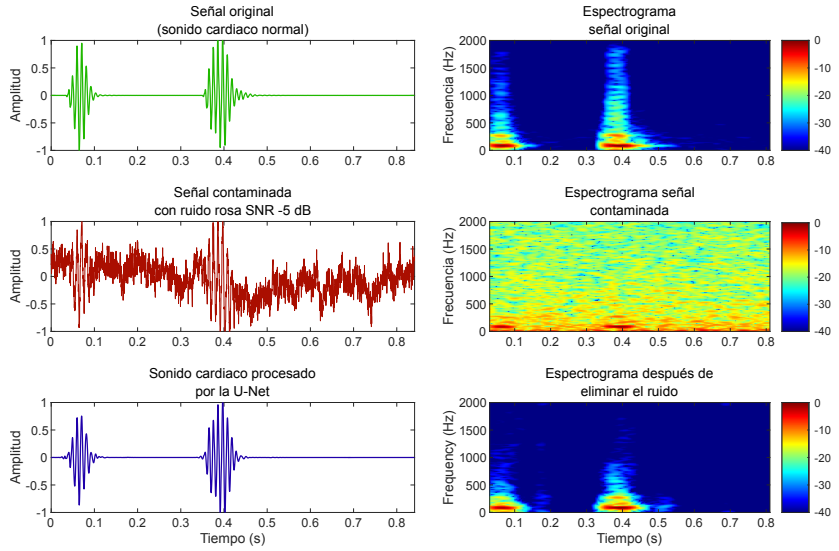


Fig. 7. Ejemplo del funcionamiento de la red para limpiar una señal de FCG normal contaminada con ruido rosa a -5 dB de SNR.

3. Resultados

El desempeño de la red propuesta fue evaluado mediante una validación cruzada de 10 iteraciones, es decir se entrenaron diez redes para cada uno de los dos tipos de ruido (blanco y rosa). Las redes entrenadas se evaluaron contaminando las señales del conjunto de prueba con una relación señal-ruido (SNR) de -5, 0, 5 y 10 dB. Estos valores de SNR se seleccionaron ya que cubren el intervalo en el que es más necesaria la eliminación de ruido.

Las redes se entrenaron para los cuatro niveles de SNR, pero en nuestras pruebas observamos que el entrenamiento a 0 dB SNR muestra los mejores resultados para cualquier nivel de ruido. Durante la evaluación, cada conjunto de prueba contenía un total de 100 señales, 20 para cada uno de los sonidos cardíacos disponibles en la base de datos [22]. Los programas para llevar a cabo el algoritmo propuesto fueron desarrollados en Python. Para implementar la U-Net se utilizaron las librerías de software TensorFlow y Keras [1].

Los bloques de procesamiento de la señal se implementaron utilizando SciPy [23] y Librosa [15]. La métrica para medir el desempeño de la red es una versión ligeramente modificada de la SNR, la cual fue originalmente propuesta en [13]. Esta métrica da lugar a una medida más sencilla y robusta denominada relación señal a distorsión invariante a la escala (SI-SDR por sus siglas en inglés) y que está definida como:

$$\text{SI-SDR} = 10\log_{10}\left(\frac{\|\frac{\hat{s}^T s}{\|s\|^2}s\|^2}{\|\frac{\hat{s}^T s}{\|s\|^2}s-\hat{s}\|^2}\right), \tag{3}$$

donde s es la señal objetivo (original) y $\hat{s}$ es la señal obtenida a la salida del algoritmo de eliminación de ruido. Esta redefinición de la SNR tiene en cuenta un posible reescalado de la señal y es invariante frente a variaciones de amplitud, evitando así resultados engañosos que pueden obtenerse utilizando únicamente la SNR [13].

La Tabla 1 muestra los resultados obtenidos usando la metodología propuesta en la sección anterior. La capacidad de la red para eliminar el ruido en escenarios difíciles es bastante notable, brindando mejoras de $\approx$15 dB para señales contaminadas con ruido blanco a -5dB de SNR y de $\approx$12 dB para señales contaminadas con ruido rosa.

Como se esperaba, los resultados muestran que la red elimina mejor el tipo de ruido para el que fue entrenada. Sin embargo, el desempeño es más alto para ruido blanco que para ruido rosa. Este fenómeno también tiene sentido, ya que al ser la señal de FCG de tipo pasa-bajas, el ruido rosa (que tiene mucho mayor energía en las bajas frecuencias que en las altas frecuencias) tiene un efecto más importante en la disminución de la calidad de la señal.

Conforme el nivel de ruido presente en la señal disminuye, también disminuye ligeramente el desempeño de la red neuronal. Este mismo comportamiento es visible para ambos tipos de ruido. Las Figuras 6 y 7 muestran dos ejemplos particulares del desempeño de la red para limpiar señales contaminadas con un nivel de -5 dB y utilizando ruido blanco y rosa respectivamente. Del lado izquierdo de la figura se presentan las formas de onda en el dominio del tiempo y del lado derecho los respectivos espectrogramas.

La forma de onda de la señal original es prácticamente indistinguible en la versión contaminada de la gráfica (segunda fila). En las gráficas de la tercera fila es posible apreciar el gran trabajo de limpieza que realiza el algoritmo propuesto. En los espectrogramas es posible apreciar que es en las altas frecuencias ($>$500 Hz) donde la red tiene más dificultades para reconstruir la señal original.

4. Conclusiones

Las enfermedades cardiovasculares son una de las principales causas de mortalidad en todo el mundo; su detección precoz es fundamental para mejorar los resultados en materia de salud a largo plazo. El análisis automático de los sonidos cardíaco es un método de diagnóstico prometedor, pero es altamente susceptible al ruido durante la grabación de audio.

En este trabajo propusimos una metodología robusta para la eliminación de ruido en sonidos cardíacos, la cual está basada en el análisis de tiempo-frecuencia (mediante la transformada de Fourier de tiempo corto) y de una arquitectura de red neuronal de tipo U-Net. La metodología propuesta fue evaluada usando una base de datos pública que contiene mil sonidos cardíacos, incluyendo sonidos normales y patológicos.

Llevamos a cabo una evaluación exhaustiva del algoritmo propuesto para distintos valores de relación señal a ruido (SNR), que van desde calidad de sonido altamente desagradable (-5 dB) hasta niveles de calidad de audio aceptables (10 dB). La metodología propuesta presenta un alto desempeño ya que puede eliminar el ruido de una señal FCG contaminada a -5 dB de SNR con mejoras promedio del orden de $\approx$15 dB en el caso de ruido blanco y de $\approx$12 dB para ruido rosado.

Consideramos que el método propuesto tiene un gran potencial para mejorar significativamente el rendimiento de los algoritmos de clasificación automática de sonidos cardíacos en entornos ruidosos, pero también podría utilizarse en estetoscopios electrónicos. Versiones posteriores de este trabajo deberán enfocarse en entrenar y evaluar el desempeño de la red utilizando otras fuentes de ruido.

Específicamente, es importante entrenar la red utilizando fuentes de ruido adquiridas en condiciones reales de auscultación, tales como ruidos ambientales y señales de voz. También sería altamente deseable incrementar el número de audio limpios para entrenamiento y prueba de la red.

Referencias

1. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., et al.: TensorFlow: Large-scale machine learning on heterogeneous systems (2016) doi: 10.48550/ARXIV.1603.04467
2. Abbas, A. K., Bassam, R.: Phonocardiography signal processing. Synthesis Lectures on Biomedical Engineering, vol. 4, no. 1, pp. 1–194 (2009) doi: 10.1007/978-3-031-01637-0
3. Ali, M. N., El-Dahshan, E. S. A., Yahia, A. H.: Denoising of heart sound signals using discrete wavelet transform. Circuits, Systems, and Signal Processing, vol. 36, no. 11, pp. 4482–4497 (2017) doi: 10.1007/s00034-017-0524-7
4. Andreas, J., Eric, H., Nicola, M., Rachel, B., Aparna, K., Tillman, W.: Singing voice separation with deep U-Net convolutional networks. In: 18th International Society for Music Information Retrieval Conference, pp. 23–27 (2017)
5. Arnott, P., Pfeiffer, G., Tavel, M.: Spectral analysis of heart sounds: Relationships between some physical characteristics and frequency spectra of first and second heart sounds in normals and hypertensives. Journal of Biomedical Engineering, vol. 6, no. 2, pp. 121–128 (1984) doi: 10.1016/0141-5425(84)90054-2
6. Choi, S., Jiang, Z.: Cardiac sound murmurs classification with autoregressive spectral analysis and multi-support vector machine technique. Computers in Biology and Medicine, vol. 40, no. 1, pp. 8–20 (2010) doi: 10.1016/j.compbiomed.2009.10.003
7. Cruz-Gutiérrez, A.: Segmentación robusta de audio cardíaco mediante análisis tiempo-frecuencia y métodos de optimización. Master's thesis, Centro de Investigación Científica y de Educación Superior de Ensenada (2016)
8. Daubechies, I., Lu, J., Wu, H. T.: Synchrosqueezed wavelet transforms: An empirical mode decomposition-like tool. Applied and Computational Harmonic Analysis, vol. 30, no. 2, pp. 243–261 (2011) doi: 10.1016/j.acha.2010.08.002
9. Ghosh, S. K., Tripathy, R. K., Ponnalagu, R.: Evaluation of performance metrics and denoising of PCG signal using wavelet based decomposition. In: IEEE 17th India Council International Conference, pp. 1–6 (2020) doi: 10.1109/INDICON49873.2020.9342464
10. Gradolewski, D., Magenes, G., Johansson, S., Kulesza, W. J.: A wavelet transform-based neural network denoising algorithm for mobile phonocardiography. Sensors, vol. 19, no. 4, pp. 957 (2019) doi: 10.3390/s19040957
11. Gradolewski, D., Redlarski, G.: Wavelet-based denoising method for real phonocardiography signal recorded by mobile devices in noisy environment. Computers in Biology and Medicine, vol. 52, pp. 119–129 (2014) doi: 10.1016/j.compbiomed.2014.06.011
12. Hennequin, R., Khlif, A., Voituret, F., Moussallam, M.: Spleeter: A fast and efficient music source separation tool with pre-trained models. Journal of Open Source Software, vol. 5, no. 50, pp. 2154 (2020) doi: 10.21105/joss.02154

13. Le-Roux, J., Wisdom, S., Erdogan, H., Hershey, J. R.: SDR–half-baked or well done? In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 626–630 (2019) doi: 10.48550/arXiv.1811.02508
14. Mahnke, C. B.: Automated heart sound analysis/computer-aided auscultation: A cardiologist's perspective and suggestions for future development. In: Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 3115–3118 (2009) doi: 10.1109/IEMBS.2009.5332551
15. McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E., Nieto, O.: Librosa: Audio and music signal analysis in python. In: Proceedings of the 14th Python in Science Conference, vol. 8, pp. 18–25 (2015) doi: 10.25080/Majora-7b98e3ed-003
16. Messer, S. R., Agzarian, J., Abbott, D.: Optimal wavelet denoising for phonocardiograms. Microelectronics Journal, vol. 32, no. 12, pp. 931–941 (2001) doi: 10.1016/S0026-2692(01)00095-7
17. Mohan, N., Kumar, S., Soman, K.: Group sparsity assisted synchrosqueezing approach for phonocardiogram signal denoising. In: 11th International Conference on Computing, Communication and Networking Technologies, pp. 1–5 (2020) doi: 10.1109/ICCCNT49239.2020.9225320
18. Organisation for economic cooperation and development: Obesity update (2017) www.oecd.org/health/health-systems/Obesity-Update-2017.pdf
19. Pauline, S. H., Dhanalakshmi, S.: A robust low-cost adaptive filtering technique for phonocardiogram signal denoising. Signal Processing, vol. 201, pp. 108688 (2022) doi: 10.1016/j.sigpro.2022.108688
20. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 234–241 (2015) doi: 10.48550/arXiv.1505.04597
21. Smith, J. O.: Spectral audio signal processing, W3K (2011)
22. Son, G. Y., Kwon, S.: Classification of heart sound signal using multiple features. Applied Sciences, vol. 8, no. 12, pp. 2344 (2018) doi: 10.1016/j.procs.2015.08.045
23. Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., et al.: SciPy 1.0: Fundamental algorithms for scientific computing in python. Nature Methods, vol. 17, pp. 261–272 (2020) doi: 10.1038/s41592-019-0686-2
24. World health organization: World health statistics 2021: Monitoring health for the SDGs, sustainable development goals (2021) https://www.who.int/data/gho/publications/world-health-statistics
25. Xie, J., Xu, L., Chen, E.: Image denoising and inpainting with deep neural networks. Advances in Neural Information Processing Systems, vol. 25 (2012)

ISSN 1870-4069

Remoción de líneas en imágenes de textos manuscritos utilizando una red neuronal convolucional tipo U-Net

Diego A. Peralta Rodríguez, José E. Valdez Rodríguez,
Nahum Carlos Alexis Rangel, Francisco Hiram Calvo Castro

Instituto Politécnico Nacional,
Centro de Investigación en Computación,
México

{dperaltar2022, jvaldezr2018, hcalvo}@cic.ipn.mx,
rangelcarloss41@gmail.com

Resumen. Cuando se digitalizan documentos escritos a mano surgen diversos problemas, sobre todo en la etapa de preprocesamiento de imágenes. Uno de los problemas tiene que ver con la remoción de líneas horizontales que existen en las hojas, ya que muchas veces al traslaparse las líneas con las palabras se dificulta la extracción de información escrita. Anteriormente ya se han propuesto e implementado distintos enfoques para tratar de resolver este problema con técnicas y algoritmos clásicos de preprocesamiento. Sin embargo, pensamos que es importante el poder aprovechar las ventajas que ofrecen los métodos que se basan en Redes Neuronales Convolucioneles (RNC), ya que estas tienen el potencial de mejorar significativamente la precisión y eficiencia en la eliminación de líneas. En este trabajo presentamos un método para remover líneas en textos manuscritos, sin afectar la información textual o la comprensión del mensaje escrito, implementando una red convolucional tipo U-Net. Para esto se llevaron a cabo experimentos utilizando conjuntos de imágenes especializados para tareas de remoción de pentagramas y para identificación de rasgos de personalidad, ya que no existen conjuntos de imágenes específicos para el problema de eliminación de líneas.

Palabras clave: Remoción de líneas, texto manuscrito, RNC.

Ruled Line Removal in Handwritten Text Images Using a U-Net Type Convolutional Neural Network

Abstract. When handwritten documents are digitized, several problems arise, especially in the image preprocessing stage. One of the problems has to do with ruled line removal that exist on the sheets, since many times the overlapping of lines with words makes the extraction of written information difficult. Previously, different approaches have been proposed and implemented to try to solve this problem with classical preprocessing techniques and algorithms. However, we believe that it is important to take advantage of the benefits offered by methods based on Convolutional Neural Networks (CNN), since these have the

potential to significantly improve the accuracy and efficiency of line removal. In this paper we present a method to remove lines in handwritten texts, without affecting the textual information or the comprehension of the written message, by implementing a U-Net type convolutional network. For this purpose, experiments were carried out using specialized datasets for staff line removal in musical sheets and personality trait identification tasks, as there are no specific datasets for the ruled line removal problem.

Keywords: Ruled line removal, handwritten text, CNN.

1. Introducción

El reconocimiento automático de escritura a mano es una tarea importante en muchas áreas, como la educación, la medicina, la banca y la seguridad, entre otros. El éxito de los diversos métodos de reconocimiento que se puedan aplicar dependerá en gran medida de la precisión con la que se logre segmentar el texto. Esto implica una clara separación del texto, como primer plano, y el fondo para garantizar una alta tasa de precisión en el reconocimiento. Para mejorar el rendimiento en esta tarea, usualmente se aplican técnicas de procesamiento tales como la mejora de contraste, algoritmos de umbralado, eliminación de ruido, detección de manchas, etc.

La remoción de líneas en documentos con hojas rayadas es un problema común en el preprocesamiento de imágenes. El propósito de las líneas es ayudar a escribir en línea recta y mantener una cierta uniformidad en el tamaño y espaciado de las letras. Remover las líneas es importante para una variedad de aplicaciones, sobre todo en el Reconocimiento Óptico de Caracteres (OCR, por sus siglas en inglés) [1], por ejemplo, en tareas para la extracción de información escrita de documentos clínicos o la digitalización de documentos. Un trabajo en donde adquiere relevancia la remoción de líneas es en el reconocimiento de rasgos de personalidad a partir del procesamiento de imágenes de textos escritos a mano [2].

La identificación de personalidad a partir de la escritura es una tarea que va cobrando cada vez más importancia en el área de visión por computadora, ya que analizar los trazos y estilo, si es que se parte desde un enfoque grafológico, requiere poder separar las letras de las líneas de fondo; o si se aborda la identificación desde un enfoque de análisis léxico, entonces se necesitará tomar en cuenta únicamente las palabras para poder clasificarlas de acuerdo a un tipo de personalidad. En resumen, actualmente se sigue enfrentando con el desafío de detectar y remover líneas con un alto grado de eficacia en diferentes tipos de documentos.

Dicha eficacia dependerá en gran medida de factores como la calidad de la imagen, la complejidad del contenido y la resolución de la misma. A su vez, la implementación de técnicas de preprocesamiento y postprocesamiento puede mejorar significativamente los resultados finales. En este trabajo se propone un método para la remoción de líneas conservando el texto sin pérdida significativa de información, utilizando una combinación de técnicas de procesamiento de imágenes y una red neuronal convolucional [3] de tipo U-Net.

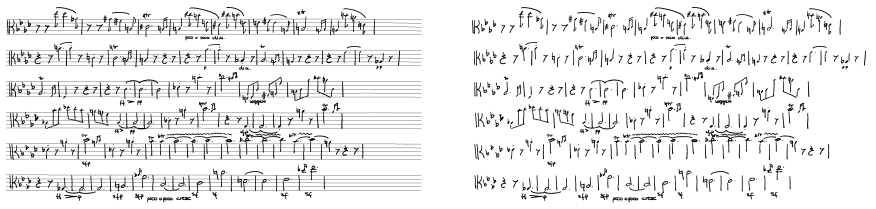


Fig. 1. Ejemplo de imágenes incluidas en el dataset CVC-MUSCIMA [13].

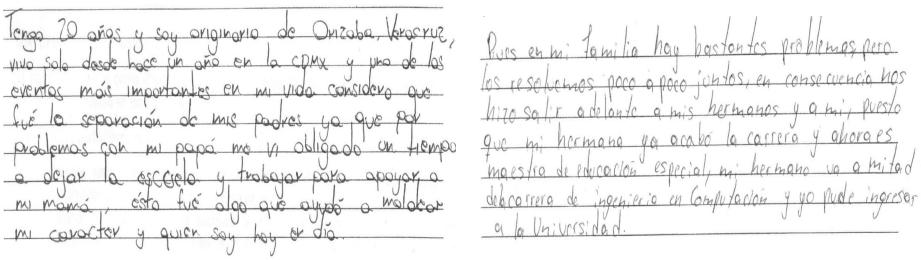


Fig. 2. Ejemplo de imágenes incluidas en el dataset HWxPI [12].

2. Trabajo previo

Diversos trabajos han propuesto métodos para lidiar con el ruido que añaden las líneas en imágenes de documentos. La mayoría de los trabajos, sobre todo los realizados antes del auge de las redes neuronales convolucionales, se centran en metodologías basadas en un preprocesamiento utilizando técnicas clásicas o estándar. Los métodos se dividían en tres importantes grupos [4]:

1. Métodos basados en morfología matemática [5].
2. Métodos que emplean la transformada de Hough para extraer características del texto y encontrar líneas en cualquier dirección [6].
3. Métodos que utilizan perfiles de proyección [7] para estimar líneas y reducir las dimensiones del problema.

No obstante, aun con la aparición de las redes neuronales convolucionales, hoy en día hay una exploración limitada en la remoción de líneas utilizando modelos neuronales. Un trabajo reciente que explora la remoción de líneas usando redes neuronales convolucionales es el presentado por [8]. Los autores crean un conjunto de datos sintéticos donde se generan líneas y se colocan en secuencias de imágenes concatenadas de palabras manuscritas.

La arquitectura propuesta consta de 3 capas convolucionales. Otros autores como Zhixin y sus colegas [9], proponen un método que se basa en un enfoque de perfilado local direccional para la detección de las ubicaciones de líneas en documentos árabes escritos a mano. Para eliminar los píxeles de las líneas de reglas, sin dañar demasiado el texto, realizan una búsqueda vertical adaptativa.

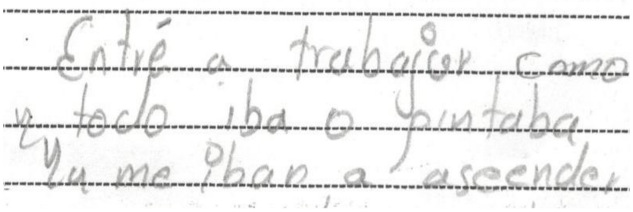


Fig. 3. Muestra parcial de imagen previa al filtro.

En el trabajo de Konstantinos [10], los autores presentan un sistema que implementa una técnica para el entrenamiento automático de la eliminación de líneas. Los parámetros de un algoritmo existente se ajustan en función de las características de una colección de documentos y se utiliza el algoritmo de recocido simulado para estimar los valores más adecuados de los parámetros.

El trabajo de Refay [11], utiliza la transformada de Hough en cuatro subventanas cuadradas. La etapa de eliminación emplea el histograma de intensidad y su entropía para aislar el texto. A la etapa de remoción le sigue una mejora mediante técnicas morfológicas.

3. Metodología

3.1. Datasets

Actualmente no hay un conjunto de datos diseñado específicamente para la tarea de remoción de líneas en documentos. Esto significa que no existe un conjunto de imágenes de documentos con y sin líneas disponibles para este propósito. La falta de un conjunto de datos de esta naturaleza limita la capacidad para mejorar y evaluar algoritmos y modelos de remoción de líneas en documentos.

Lo que más se le asemeja a datasets con este propósito son los utilizados para la remoción de pentagramas en partituras. Para superar esta carencia y limitación, se experimentaron con dos datasets diferentes en este trabajo. El primero es el conjunto de datos HWxPI [12], creado por investigadores de la UAM con el fin de identificar rasgos de personalidad a partir de ensayos escritos a mano.

Este conjunto de datos está compuesto por 418 imágenes de ensayos manuscritos (Fig. 1) y 418 transcripciones de los mismos. Aunque este dataset no fue creado específicamente para la remoción de líneas en documentos, se utilizó para entrenar y evaluar modelos debido a la presencia de líneas que contenían todas las imágenes de los ensayos. El segundo dataset utilizado en este trabajo es el CVC-MUSCIMA, diseñado para la eliminación de pentagramas en partituras [13].

Este conjunto de datos consta de 1000 imágenes de partituras con pentagramas y 1000 imágenes de partituras sin pentagramas, únicamente conformadas por las notas (Fig. 2). Aunque este conjunto de datos tampoco fue creado con el propósito específico de la remoción de líneas en documentos, se decidió utilizarlo porque las líneas de pentagrama pueden ser tratadas de manera similar a las líneas en documentos escritos a mano.

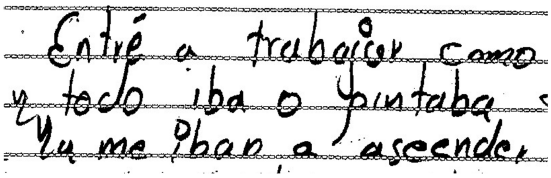


Fig. 4. Resultado después de pasar por el filtro.

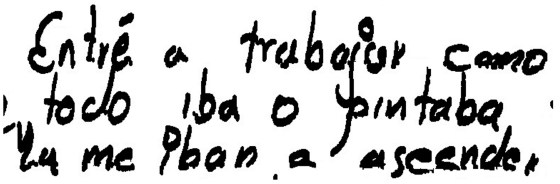


Fig. 5. Resultado después de aplicar apertura morfológica.

Por lo tanto, se determinó que este conjunto de imágenes podría resultar de gran utilidad en la realización de experimentos con el objetivo de desarrollar un modelo capaz de generalizar la tarea de remoción de líneas en diferentes tipos de documentos y no solo en partituras.

3.2. Preprocesamiento

El objetivo de trabajar con el dataset HWxPI no fue obtener la remoción de líneas del conjunto total de imágenes, sino solo de un porcentaje.

Lo que se buscaba era poder parear las imágenes originales con sus respectivas etiquetas (también conocidas como 'ground truth' en inglés) o imágenes limpias de líneas para entrenar un modelo de red neuronal. Para conseguir esto se aplicaron técnicas clásicas de preprocesamiento y morfología matemática. Se tomaron 100 imágenes como punto de partida, las cuales se recortaron para conservar únicamente el área donde había texto y se convirtieron a valores en escala de grises.

A continuación, se revisó el histograma y el valor de los píxeles que conformaban las líneas de los ensayos para definir un umbral adecuado. Es bien sabido que es complejo o casi imposible definir un valor único de umbralado para imágenes que difieran una de otra incluso en detalles que puedan parecer triviales. Por ejemplo, el hecho de que la persona al escribir haya recargado más o menos el lápiz, ocasiona un cambio grande en los valores de píxeles; si las líneas son continuas o son punteadas, si se escribe con bolígrafo o lápiz, si hay manchas en la hoja, etc.

Por lo tanto, se sabía que solo un porcentaje de las imágenes podrían tener éxito en la remoción total o parcial de las líneas, debido a la variabilidad en las características de las imágenes. Una vez realizadas las modificaciones necesarias, se determinó que el rango de píxeles más adecuado para umbralar se encontraba entre 120 y 230. En consecuencia, se diseñó un filtro para preservar únicamente los valores de gris dentro de ese rango. Tras aplicar el filtro, de las 100 imágenes se obtuvieron 40 con una remoción casi total de líneas.

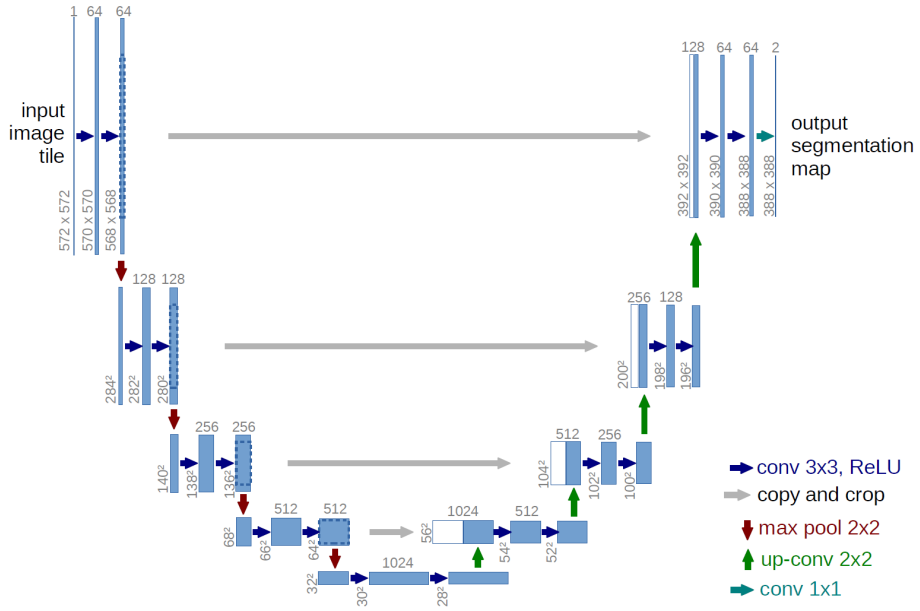


Fig. 6. Arquitectura U-Net [15].

En las figuras 3 y 4 se muestra el ejemplo de una sección de imagen antes y después de pasarse por el filtro. Gran parte del relleno en las líneas se pudo eliminar quedando solamente su contorno. Para eliminar el sobrante, se aplicó una operación morfológica de apertura, dada por la ecuación 1:

$$A \circ B = (A \ominus B) \oplus B. \tag{1}$$

Esta operación realiza una erosión seguida de una dilatación utilizando un mismo elemento de estructura para ambas operaciones. Se experimentó con diferentes elementos de estructura y el que mejor funcionó fue de forma circular y tamaño 3 x 3. La erosión eliminó por completo los contornos sobrantes y la dilatación ayudó a conservar la estructura original que tenían las letras (Fig. 5).

Por último, ya que en algunas imágenes quedaban puntos de cierta área que no se pudieron remover con una erosión, se aplicó una apertura por área. Esta es una operación útil de filtro que consiste en remover todos los componentes conectados cuya área en número de píxeles sea más pequeña que el valor de umbralado que se proponga [14].

Se agregaron 10 imágenes adicionales sin líneas al conjunto inicial de 40, pero estas fueron obtenidas mediante la eliminación manual de líneas en lugar de utilizar operaciones morfológicas. Como resultado, en total se generó un pequeño conjunto de datos compuesto por 100 imágenes: 50 imágenes originales y 50 imágenes sin líneas. Con respecto al conjunto de imágenes en el dataset CVC-MUSCIMA, primero se ensancharon las líneas del pentagrama utilizando una técnica de dilatación con un elemento de estructura rectangular de 3×5.

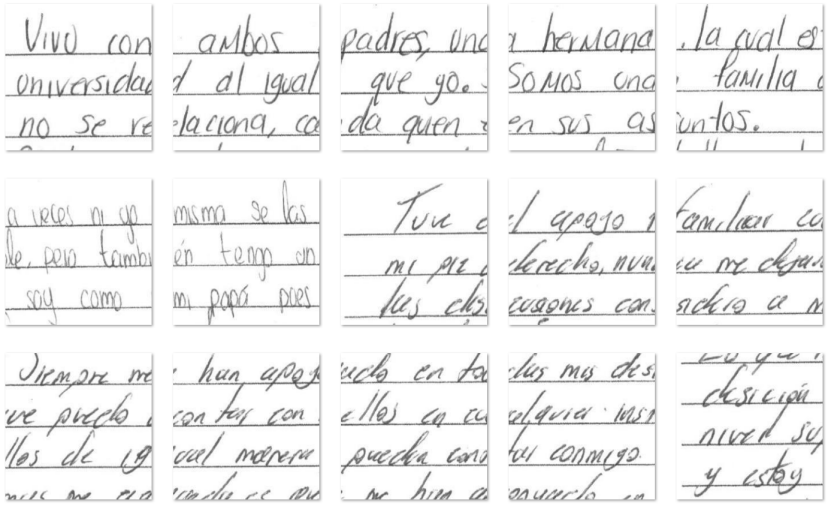


Fig. 7. Muestra de parches generados.

También se obtuvo el negativo (ya que originalmente el fondo era color negro) y se convirtieron a escala de grises. Estas modificaciones permitieron homologar, en cierto grado, las características con el conjunto de datos HWxPI, para poder después utilizarlos de manera conjunta en la tarea de entrenamiento y evaluación.

3.3. U-Net

La arquitectura U-Net (Fig. 6), ha demostrado ser altamente efectiva sobre todo en la segmentación de imágenes médicas debido a su capacidad para realizar predicciones precisas de la máscara de segmentación, incluso en situaciones en las que las regiones de interés son pequeñas o poco claras [15]. Esta es una Red Neuronal Convolucional (RNC) profunda que consta de dos partes principales: la ruta de codificación (encoder) y la ruta de decodificación (decoder).

La ruta de codificación es similar a la arquitectura de una CNN típica, donde la imagen de entrada se reduce gradualmente en tamaño mediante la aplicación de capas de convolución y submuestreo (pooling). Por otro lado, la ruta de decodificación aumenta gradualmente el tamaño de la imagen mediante la aplicación de capas de convolución y sobremuestreo (upsampling).

Además, U-Net utiliza una técnica llamada "salto de conexiones" (skip connections), que conecta las capas de codificación y decodificación, permitiendo que la información de alta resolución se transmita directamente a las capas de decodificación, lo que ayuda a evitar la pérdida de detalles importantes en la imagen.

El modelo U-Net es muy útil porque puede entrenarse con relativamente pocas imágenes, lo que la hace ideal para aplicaciones en las que se dispone de un conjunto de datos pequeño o se necesita una segmentación rápida en tiempo real. Debido a esta razón se optó por trabajar con esta arquitectura, ya que el conjunto de imágenes de HWxPI obtenido después de la eliminación de líneas fue limitado.

(a) (b) (c) (d) (e) (f)

Fig. 8. Muestras de resultados en la segmentación usando los dataset HWxPI y CVC-MUSCIMA, a) Parches original con líneas, b) Máscara, c) Predicción con HWxPI, d) Predicción con CVC-MUSCIMA, e) Intersección de a y c, f) Intersección de a y d.

4. Experimentos

Se realizaron experimentos con las imágenes obtenidas del dataset HWxPI y con el fin de ampliar el conjunto de entrenamiento, se propuso dividir cada imagen en parches de 256 x 256 píxeles, lo que permitió generar un conjunto de datos compuesto por 450 imágenes (Fig. 7). Además, se realizó un aumento de datos, que dio como resultado 64 imágenes adicionales. Este aumento incluyó técnicas de rotación, espejo, acercamiento, estiramiento en lo alto y ancho.

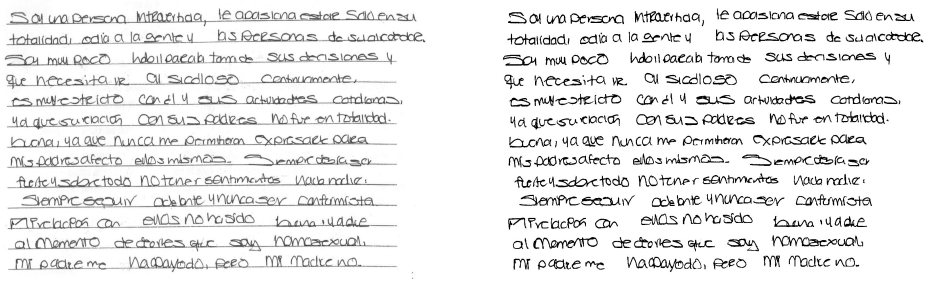
Una de las ventajas del aumento de datos es que ayuda a reducir el sobreajuste al introducir variaciones en el conjunto de entrenamiento, lo que hace que el modelo sea menos propenso a memorizar las muestras y tenga una mayor capacidad de generalización. Para la definición de hiperparámetros, se utilizaron 64 filtros, un dropout de 0.3, batch size de 32 y una función de activación sigmoide. Por otro parte, la métrica utilizada para evaluar el modelo fue la intersección sobre la unión (Ec. 2):

$$IoU = \frac{\text{Área de intersección}}{\text{Área de unión}}. \tag{2}$$

Esta métrica consiste en dividir el área de intersección entre las predicciones del modelo y las etiquetas reales por el área de unión de ambas. Es ampliamente utilizada en tareas de segmentación y permite medir la similitud entre la segmentación realizada por el modelo y la segmentación real, siendo un valor de 1 indicativo de una segmentación perfecta y un valor de 0 indicativo de una segmentación completamente errónea.

Tabla 1. Resultados cuantitativos (mayor es mejor).

	IoU del conjunto HWxPI de prueba
Entrenado con CVC-MUSCIMA	0.60
Entrenado con HWxPI	**0.73**



(a) Imagen original. (b) Imagen después de ser procesada.

Fig. 9. Resultado de modelo U-Net para una imagen.

Para la optimización del modelo se empleó el algoritmo de descenso de gradiente estocástico, junto con una función de pérdida de entropía cruzada binaria. Durante el entrenamiento, se obtuvo un valor de pérdida de 0.0667, mientras que el coeficiente de intersección sobre la unión (IoU) arrojó un resultado de 0.6847.

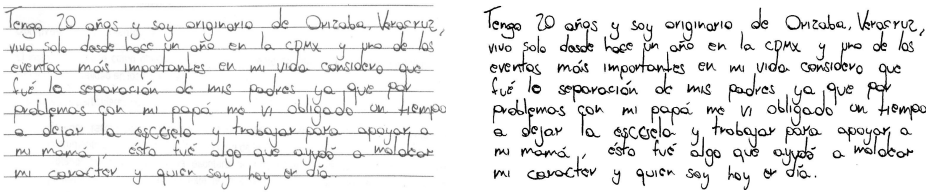
Al evaluar el modelo con el conjunto de pruebas, se logró obtener un valor de IoU de 0.733. Las imágenes resultantes se muestran en la figura 8(c). En cuanto a los experimentos realizados con CVC-MUSCIMA, de igual manera se generaron parches de cada imagen, 256 $\times$ 256 píxeles, y se realizó aumentación de datos. Con esto se pasó de tener 1000 imágenes a 5,955. Se cambiaron los valores de dropout a 0.4 y el tamaño de batch size a 64.

Todos los demás parámetros se conservaron igual que con el entrenamiento previo. En el entrenamiento se obtuvo un valor de pérdida de 0.02, mientras que el resultado de IoU fue 0.98. En cambio, con el conjunto de evaluación (test), que en este caso fueron los 450 parches obtenidos de las 50 imágenes de nuestro dataset que resultó del preprocesamiento, se consiguió un coeficiente de IoU de 0.60. Las imágenes resultantes se muestran en la figura 8(d).

5. Resultados

Con el dataset CVC-MUSCIMA en la fase de entrenamiento se lograron resultados por encima de 0.90 en la métrica IoU, pero en las pruebas y utilizando un conjunto de imágenes externo, no se pudo generalizar la tarea de eliminar únicamente las líneas dando como resultado una gran pérdida de texto.

Al contrario, con el dataset HW$\times$PI durante la fase de entrenamiento se obtuvieron resultados ligeramente superiores a 0.60 en la métrica IoU . Sin embargo, en las pruebas, se obtuvieron valores significativamente elevados y se tuvo una mayor precisión en la segmentación de objetos, lo cual se tradujo en una remoción de líneas correctamente y texto no distorsionado.



(a) Imagen original. (b) Imagen después de ser procesada.

Fig. 10. Resultado de modelo U-Net para una imagen.

En la Tabla 1 se puede observar el resultado cuantitativo de los experimentos antes mencionados. En la figura 8(e) y 8(f), el color rojo en las letras resalta una intersección exitosa entre la imagen original y la predicción. Como puede verse, el color rojo en las predicciones resultantes del dataset de partituras es casi imperceptible.

Una vez establecido el modelo, se utilizaron todas las imágenes del dataset HWxPI. Cada imagen se dividió en parches de 256 $\times$ 256 píxeles, generando un total de 6200 imágenes parche. Cada uno de estos parches fue procesado para remover las líneas presentes. Posteriormente, los parches de cada imagen se unieron nuevamente para restaurar la imagen original. Finalmente, todas las imágenes fueron binarizadas, ya que los resultados se obtuvieron en escala de grises. En la figura 9 y figura 10 se presentan ejemplos del resultado final.

6. Conclusiones y trabajo a futuro

Debido a que no existen datasets disponibles que contengan ejemplos de imágenes de textos con y sin líneas, se optó por experimentar con dos tipos de datasets, CVS-MUSIMA y HW$\times$PI. El primero a pesar de ser un conjunto de imágenes para la remoción de líneas de pentagrama y de su similitud con el objetivo del presente trabajo, no tuvo un resultado favorable. Se encontró que el segundo pudo realizar una mejor generalización y sin una pérdida significativa de información.

Utilizar una arquitectura tipo U-Net demostró ser una herramienta valiosa permitiendo obtener buenos resultados en la eliminación de líneas en textos manuscritos, sin la necesidad de un realizar un trabajo adicional de reconstrucción. Aunque en algunos casos hubo que hacer una dilatación seguida de una erosión para acabar de cerrar algunos espacios, en general no hubo alteraciones del texto. A diferencia de trabajos previos en los que, utilizando técnicas estándar de procesamiento, después de eliminar las líneas tuvieron que implementar algoritmos de reconstrucción.

Si bien este trabajo ha demostrado el potencial de las redes neuronales convolucionales en la eliminación de líneas, es importante destacar que aún existen áreas que pueden ser exploradas en futuras investigaciones. Por ejemplo, se pueden proponer diferentes arquitecturas de red para mejorar aún más la precisión del modelo. Para el trabajo a futuro, se considera combinar conjuntos de imágenes de ensayos y partituras para el entrenamiento, así como evaluar con diversas imágenes de documentos. Por último, se planea desarrollar un dataset específico para este tipo de tarea y ponerlo a disposición de aquellos que trabajen en áreas afines.

Referencias

1. Schantz, H. F.: History of OCR, optical character recognition. Recognition Technologies Users Association (1982)
2. Valdez-Rodríguez, J. E., Calvo, H., Felipe-Riverón, E. M.: Handwritten texts for personality identification using convolutional neural networks. Pattern Recognition and Information Forensics, pp. 140–145 (2018) doi: 10.1007/978-3-030-05792-3_13
3. LeCun, Y., Kavukcuoglu, K., Farabet, C.: Convolutional networks and applications in vision. In: Proceedings of IEEE International Symposium on Circuits and Systems (2010) doi: 10.1109/iscas.2010.5537907
4. Farahmand, A., Sarrafzadeh, A., Shanbezadeh, J.: Document image noises and removal methods. Lecture Notes in Engineering and Computer Science, vol. 2202, pp. 436–440 (2013)
5. Soille, P.: Morphological image analysis: Principles and applications. 2nd edn., Springer, Berlin (2004) doi: 10.1007/978-3-662-05088-0
6. Duda, R. O., Hart, P. E.: Use of the hough transformation to detect lines and curves in pictures. Communications of the Association for Computing Machinery, vol. 15, no. 1, pp. 11–15 (1972) doi: 10.1145/361237.361242
7. Namboodiri, A. M., Jain, A. K.: Document structure and layout analysis. Digital Document Processing, pp. 29–48 (2007) doi: 10.1007/978-1-84628-726-8_2
8. Gold, C., Zesch, T.: CNN-based ruled line removal in handwritten documents. Frontiers in Handwriting Recognition, pp. 530–544 (2022) doi: 10.1007/978-3-031-21648-0_36
9. Shi, Z., Setlur, S., Govindaraju, V.: Removing rule-lines from binary handwritten arabic document images using directional local profile. In: 20th International Conference on Pattern Recognition (2010) doi: 10.1109/icpr.2010.472
10. Prokopiou, K., Kavallieratou, E., Stamatatos, E.: An image processing self-training system for ruling line removal algorithms. In: 18th International Conference on Digital Signal Processing (2013) doi: 10.1109/icdsp.2013.6622767
11. Refaey, M. A.: Ruled lines detection and removal in grey level handwritten image documents. In: 6th International Conference on Information and Communication Systems (2015) doi: 10.1109/iacs.2015.7103230
12. Ramírez, G., Villatoro, E., Ionescu, B., Escalante, H. J., Escalera, S., Larson, M., Müller, H., Guyon, I.: Overview of the multimedia information processing for personality and social networks analysis contest. Pattern Recognition and Information Forensics, pp. 127–139 (2018) doi: 10.1007/978-3-030-05792-3_12
13. Fornés, A., Dutta, A., Gordo, A., Lladós, J.: CVC-MUSCIMA: A ground truth of handwritten music score images for writer identification and staff removal. International Journal on Document Analysis and Recognition, vol. 15, no. 3, pp. 243–251 (2011) doi: 10.1007/s10032-011-0168-2
14. Vincent, L.: Morphological area openings and closings for grey-scale images. Shape in Picture, pp. 197–208 (1994) doi: 10.1007/978-3-662-03039-4_13
15. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. Lecture Notes in Computer Science, pp. 234–241 (2015) doi: 10.1007/978-3-319-24574-4_28

ISSN 1870-4069

Sistema inteligente para la detección de ninfas de mosca blanca presentes en hojas de plantas

Diana Karina Jacobo-Rubio, Modesto Medina-Melendrez

Tecnológico Nacional de México,
Instituto Tecnológico de Culiacán,
México

{M20171595, modesto.mm}@culiacan.tecnm.mx

Resumen. La FAO revela en sus estadísticas que 2570 millones de persona dependen de la agricultura para su subsistencia. Un factor de riesgo para esta son las plagas, generando pérdidas de hasta un 40%. Una de las plagas más comunes en regiones tropicales y subtropicales es la mosca blanca Bemisia tabaci (Gennadius), que afecta a más de 600 especies de plantas. Para tener un control de esta plaga es necesario realizar su monitoreo, el cual es recomendable realizarlo en su etapa ninfal ya que en esta etapa se pueden obtener estimaciones precisas del nivel de infestación. Este artículo propone un monitoreo automático entrenando y usando un sistema inteligente para detectar ninfas de mosca blanca en hojas de plantas utilizando la red YOLO V2. El sistema se validó utilizando un banco de imágenes de hojas de tomate infestadas con mosca blanca, y se logró una precisión promedio de hasta 97% al detectar esta plaga.

Palabras clave: Plagas, mosca blanca, detección y sistemas inteligentes.

Intelligent System for the Detection of Whitefly Nymphs Presents on Plant Leaves

Abstract. FAO reveals that 2.57 billion people subsist on agriculture. A risk factor for it are pests, generating losses of up to 40%. One of the most common pests in tropical and subtropical regions is the whitefly Bemisia tabaci (Gennadius), affecting more than 600 plant species. To control this pest, it can be monitored in nymphal stage, since it estimates the level of infestation at this stage. This paper proposes an automatic monitoring by training and using an intelligent system for detecting whitefly nymphs present on leaves using the YOLO V2 network. After systems's validation using an image bank of infested tomato leaves, an average precision of 97% in detection was obtained.

Keywords: Pests, whitefly, detection and intelligent system.

1. Introducción

La Organización de las Naciones Unidas para la Alimentación y la Agricultura (FAO) revela en sus estadísticas que, a principios del nuevo milenio, 2570 millones de personas dependen de la agricultura, caza, entre otras actividades para su subsistencia.

Estas personas representan el 42% de la humanidad [1]. Un factor de riesgo para la agricultura son las plagas, estas pueden generar pérdidas de hasta un 40 por ciento de acuerdo con la FAO [2]. Una de las plagas mayormente distribuidas en regiones tropicales y subtropicales del mundo es la mosca blanca Bemisia tabaci (Gennadius), esta afecta a más de 600 especies de plantas cultivadas y silvestres [3].

En la última década la mosca blanca ha causado millones de dólares en pérdidas de cultivos en agroecosistemas a lo ancho del mundo [4]. Existen distintos métodos tanto químicos como biológicos para controlar las plagas y para poder aplicarlos es necesario mantener un monitoreo constante, en el caso de la mosca blanca es recomendable realizarlo en su etapa ninfal, ya que en la etapa adulta estas no suelen quedarse en las hojas y esto provoca estimaciones poco precisas del nivel de infestación [5].

El método más utilizado para monitorear poblaciones de plagas en estudios de salud de cultivos es mediante una inspección visual realizada por trabajadores. Este método resulta ser complicado, requiere mucho tiempo, suele ser extenuante para los trabajadores y por ello es propenso a errores humanos [6]. Por lo que es necesario desarrollar un instrumento automático para monitorear eficientemente las plagas en siembras e invernaderos.

La agricultura 4.0 busca integrar un conjunto de tecnologías, dispositivos, protocolos y paradigmas computacionales para mejorar los procesos agrícolas. En este nuevo enfoque se utilizan técnicas de procesamiento digital de imágenes como una herramienta que permite identificar de manera temprana las plagas o enfermedades en los cultivos [7].

Tanto los agricultores como las empresas buscan aumentar la producción y reducir los residuos, como resultado, la IA está emergiendo constantemente como parte de la evolución tecnológica de la industria agrícola [8].

El uso de técnicas de procesamiento de imágenes (PDI) e inteligencia artificial (IA) es una posibilidad para desarrollar sistemas de monitoreo automático. Para implementar este sistema es necesario contar con cámaras para capturar imágenes de las plantas o en su defecto de las hojas que son clave para detectar las ninfas de mosca blanca, luego es necesario aplicar técnicas de procesamiento de imágenes para poder detectar las ninfas.

2. Trabajos relacionados

En la literatura existen distintos sistemas desarrollados para la detección de mosca blanca adulta, pero son pocos los sistemas reportados para la detección de moscas blancas en su estado ninfal. Uno de los trabajos reportados se describe en el artículo escrito por Lino [9], el cuál utiliza técnicas de detección de patrones mediante algoritmos Haar-like Features (Haar) y Local binary patterns (LBP) para detectar y contar ninfas de mosca blanca en frijol ejotero.

En este artículo, Lino reporta resultados con un alto índice de falsos positivos provocados por la misma morfología de la planta, lo que afecta al conteo real de esta plaga. Una alternativa diferente fue reportada por Bernal [10], el cual propone un sistema de captura, detección de formas elípticas irregulares y técnicas de combinatoria para la detección y conteo de mosca blanca en estado ninfal sobre hojas de tomate. Este sistema detecta y cuenta las ninfas de mosca blanca sobre hojas de tomate de forma automática, sin embargo, este sistema cuenta con una precisión menor al 85%.

Tabla 1. Técnicas utilizadas para detectar plagas.

Técnicas utilizadas	Precisión
Red neuronal convolucional	92.4%
Codificación multitarea escasa, técnica de aprendizaje multinúcleo	Color:70.2%, textura:63.5, forma:80.2%
Clasificador multicapa basada en el clasificador de umbral aleatorio	Rango de reconocimiento: 85%
Aprendizaje de múltiples instancias	59.8%
Aprendizaje profundo con imágenes UAV	91.2%
Método de aprendizaje residual profundo	Clasificación de 98.6%
Clasificación con Máquinas vectoriales de soporte (SMV)	Error menor al 2.5%
Cascada Haar	83%
Red neuronal convolucional por regiones más rápida (RCNN)	98%

La aplicación de tecnologías e inteligencia artificial (IA) en la agricultura se ha desarrollado a lo largo de los años como parte de las soluciones hacia la mejora de la productividad agrícola y así poder satisfacer la enorme demanda alimentaria que aumenta año con año. Estas tecnologías e implementación de la IA no solo permiten a los agricultores mejorar la eficiencia, sino que también mejoran la calidad, la cantidad y garantizan una comercialización más rápida de los cultivos [11].

En la tabla 1 se muestra una recopilación de técnicas para detectar plagas mediante sistemas inteligentes realizada por Gómez [12] y Liakos [13]. De acuerdo con el estudio de los antecedentes, existen distintos trabajos relacionados con la detección de mosca blanca en cultivos, siendo pocos los sistemas relacionados con la detección de la mosca blanca en estado ninfal.

Entre los sistemas que detectan mosca blanca en su estado ninfal se encuentra el sistema propuesto por Bernal, este logra una precisión menor al 85%. En los sistemas desarrollados para la detección de mosca blanca en estado ninfal también se encuentra el sistema propuesto por Lino, el cual detecta esta plaga en hojas de frijol ejotero, el inconveniente es que presenta altos índices de falsos positivos como lo indica en su artículo [9].

En el caso de los sistemas que utilizan técnicas de IA y/o PDI, estos son capaces de detectar eficientemente plagas, pero aún no se ha desarrollado un sistema que utilice estas técnicas en conjunto para detectar moscas blancas en estado ninfal sobre hojas de plantas. Por ello, en este artículo se describe un sistema inteligente para la detección de ninfas de mosca blanca presentes en hojas de plantas cuyo propósito es mejorar la precisión respecto a los trabajos previamente reportados en la literatura.

3. Metodología propuesta

Como se puede observar en la tabla 1, existen distintas técnicas que pueden lograr una alta precisión para detectar plagas, una de las que mayor precisión presenta es la técnica RCNN. La red llamada You Only Look Once versión 2 (YOLO V2) es del tipo

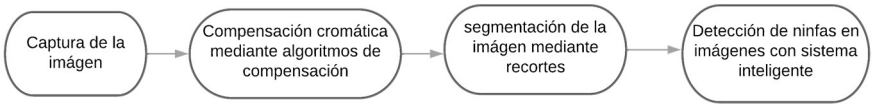
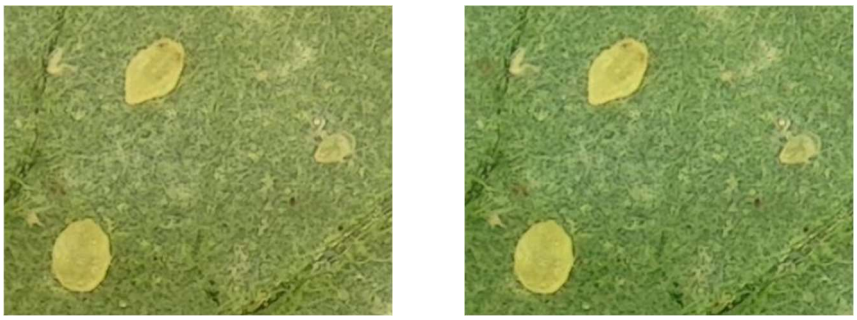


Fig. 1. Diagrama de flujo del sistema inteligente para la detección de mosca blanca en estado ninfal.



(a) (b)

Fig. 2. Compensación cromática. (a) Fragmento de una imagen original (b) Fragmento de una imagen compensada.

RCNN y se utiliza en el desarrollo de este proyecto, ya que esta se puede configurar, generar y entrenar solamente modificando ciertos parámetros de entrenamiento.

Existen distintas plataformas de programación como MATLAB y Python que cuentan con herramientas para configurar una red YOLO V2. Para crear una red de detección de objetos YOLO V2 es necesario especificar los parámetros del tamaño de imagen de entrada, número de clases, número de cajas de anclaje, número de épocas, tamaño del mini lote de entrenamiento y las épocas.

Primero, para especificar el tamaño de imagen de entrada el cuál es de 224x224x3 pixeles, se consideró el tamaño mínimo requerido de la red, ya que al momento de entrenar la red el coste computacional se ve reducido al procesar los datos. Conforme se avance en esta sección se estará brindando más información acerca de la configuración de los otros parámetros.

El sistema desarrollado está compuesto por cuatro módulos, la Fig. 1 representa el diagrama con las etapas de procesamiento que se consideran importantes para poder realizar la detección de ninfas de mosca blanca presentes en hojas de plantas. Las siguientes secciones están divididas en correspondencia a los módulos del sistema propuesto, por lo que en cada sección se detalla que es lo que se necesita para generar cada módulo, qué es lo que hace y el resultado de este.

3.1. Captura de imagen

En este módulo se implementa un sistema de captura de imágenes constituido por un sistema óptico de reducción, ya que la captura total de la escena en la imagen debe

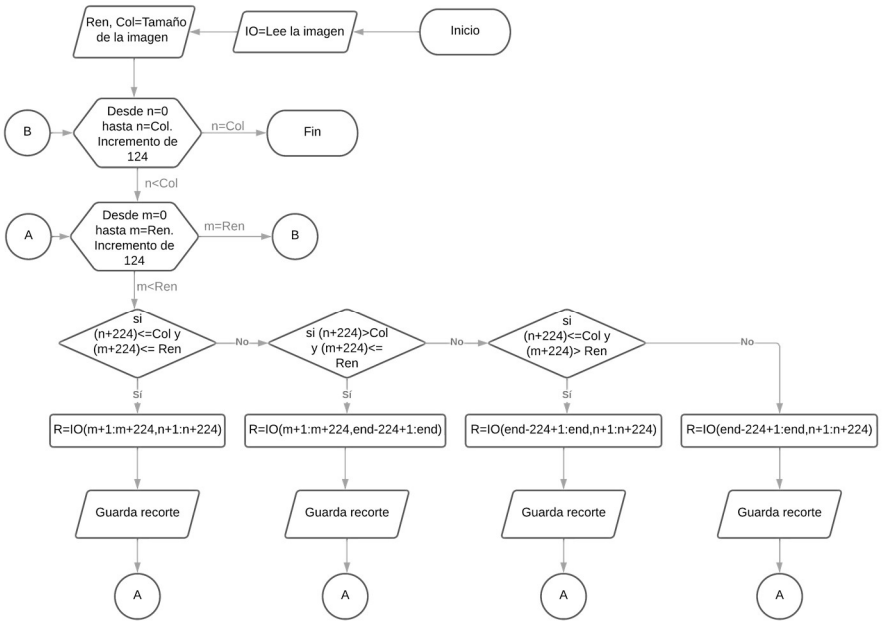


Fig. 3. Algoritmo utilizado para hacer recortes.

abarcar un área de 1 in^2; una cámara; un sistema de iluminación y una base para poder capturar las muestras.

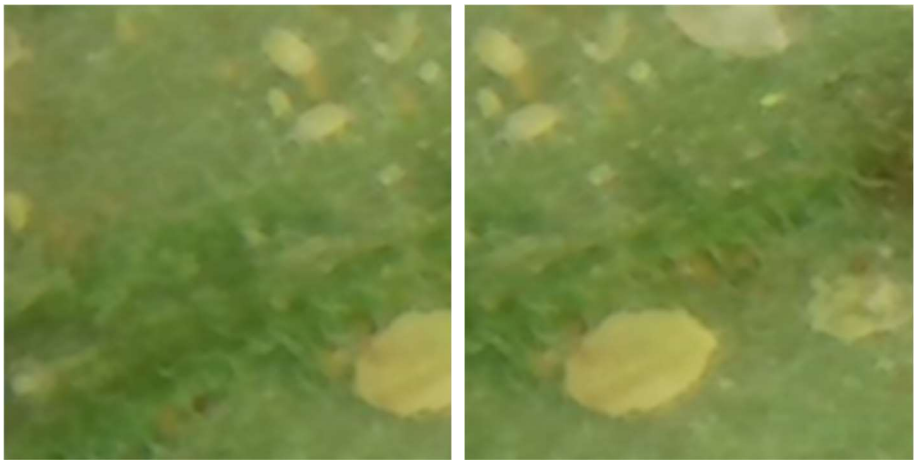
El sistema óptico de reducción y captura de imágenes está compuesto por una cámara con puerto USB para captura y transmisión de imágenes al ordenador y lentes de reducción para ajustar una escena relativamente grande al área sensible del sensor de la cámara. Para poder obtener un mejor enfoque en las capturas de imágenes. La cámara cuenta con una apertura para limitar la cantidad de luz que entra al lente.

En cuanto al sistema de iluminación se utilizó un anillo de leds que cumple con la iluminación estándar D65, ya que es la requerida por la paleta de colores de referencia. Las características técnicas de la cámara y el sistema de lentes utilizan un sensor True color CMOS, distancia entre pixeles de 2.2 um × 2.2 um, una resolución de 2592 columnas x 1944 renglones y un aumento de 0.12× – 2×. Se utiliza un ordenador con el software de MATLAB versión R2019a instalado.

Las especificaciones de la computadora son las siguientes: el modelo del ordenador es una Torre Dell Precision 3620, el procesador es Intel (R) Core (TM) i7-7700 CPU @ 3.60GHz, la RAM instalada es de 16.0 GB, utiliza un sistema operativo de 64 bits, y sistema operativo Windows 10 de 64 bits.

3.2. Compensación cromática

La compensación cromática estandariza las medidas de color y así poder caracterizar correctamente la croma del objeto de estudio (las ninfas de mosca blanca). Las imágenes capturadas dependen de diferentes variables de la cámara, ya sea por el tipo



(a) (b)

Fig. 4. Imágenes del banco de imágenes. (a) ninfa recortada (b) ninfa completa.

de sensor que utiliza (CMOS o CCD), la posición de los fotosensores, el patrón de Bayer qué utilicen, entre otros factores.

En este caso se realizó la compensación cromática mediante los polinomios caracterizados por Bernal [10], el cual utilizó como referencia la paleta de colores Macbeth Colorcheck para parametrizar el polinomio de la Ecuación 1 que se utiliza para compensar las imágenes capturadas:

$$\begin{aligned} \mathrm{Hr} = {} & -506.3\mathrm{Hm}_8 + 2060.7\mathrm{Hm}_7 - 3407\mathrm{Hm}_6 + 2926.4\mathrm{Hm}_5 - 1383.5\mathrm{Hm}_4 \\ & + 350.7\mathrm{Hm}_3 - 43.4\mathrm{Hm}_2 + 3.357\mathrm{Hm} - 0.0373, \end{aligned} \tag{1}$$

donde H_r es el valor de H (Hue) resultante y H_m es el valor medido de H. Un ejemplo de aplicar la compensación cromática se muestra en la Fig. 2, en la cual se observa la imagen original en el inciso (a) y la imagen compensada cromáticamente en el inciso (b).

3.3. Segmentación de la imagen

Algoritmo de recortes

El primer paso para entrenar una red es generar una base de datos. La base de datos necesita un banco de imágenes, por lo que posterior a la captura de imágenes de 2592×1944 pixeles se realizaron recortes de 224×224×3 de ellas para conformar el banco de imágenes.

Los recortes se realizaron con traslapes de 100 renglones y 100 columnas en sus extremos, esto para poder contener versiones completas de todas las ninfas, incluyendo las ninfas más grandes que tienen un diámetro máximo de hasta 99 pixeles. En la Fig. 3 se puede observar el diagrama de flujo del algoritmo utilizado para segmentar la imagen original.

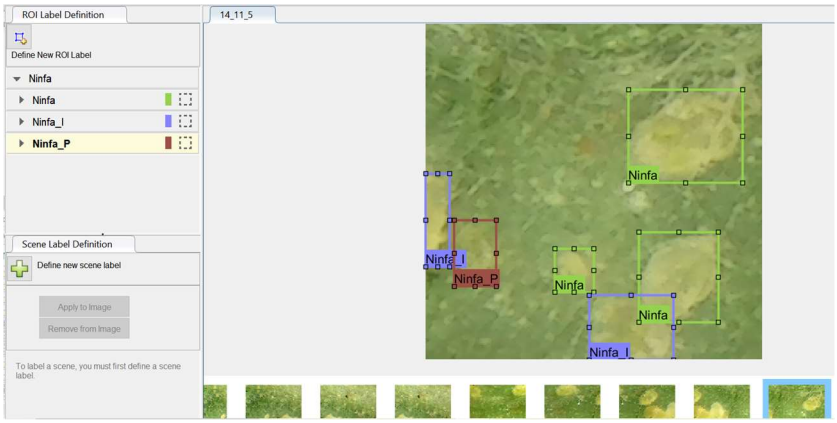


Fig. 5. Ejemplo de etiquetado de objetos a detectar.

Se constituyó un banco de imágenes de 2,016 imágenes de 224×224×3 pixeles. La Fig. 4 representa un ejemplo de recortes de imágenes, donde podemos ver a la misma ninfa en dos recortes vecinos, en el inciso (a) se observa a una ninfa cortada y en el inciso (b) se observa la ninfa completa.

3.4. Detección de ninfas con sistema inteligente

Para poder desarrollar el sistema inteligente capaz de detectar las ninfas de mosca blanca fue necesario crear un banco de imágenes de hojas infestadas y un banco de entrenamiento, configurar la red neuronal a utilizar, para posteriormente realizar su entrenamiento. Finalmente, se utiliza la red entrenada para lograr la detección de ninfas de mosca blanca.

Creación de base de datos

Se utilizó una interfaz amigable para cargar el banco de imágenes y etiquetar a las ninfas, un ejemplo de imagen cargada y objetos etiquetados se muestra en la Fig. 5. Se definieron 3 tipos de objetos a detectar y sus etiquetas correspondientes. Las etiquetas se definen por los tres casos morfológicos de las candidatas a ser ninfas que se presentan en las imágenes de hojas infestadas, y estas son: “Ninfa”, “Ninfa_I” y “Ninfa_P”.

La etiqueta “Ninfa” representa a aquellas que tienen la característica de la forma y color de una ninfa, la cuál debe ser ovalada y de un color más amarillo que el de la hoja; la etiqueta “Ninfa_I” representa a las ninfas que tienen las mismas características que la etiqueta “Ninfa”, pero esta se encuentra en los bordes de la imagen, por ello fue recortada; y la etiqueta de “Ninfa_P” representa a las candidatas a ser ninfas que cuenten con solo una de las característica deseadas, ya sea el color o la forma de la ninfa.

Creación de banco de entrenamiento

Para crear el banco de entrenamiento se necesita declarar la etiqueta, cargar el banco de imágenes, seleccionar la etiqueta que se quiere utilizar, y encerrar el objeto a etiquetar en un cuadro delimitador de región. Cuando se tienen todas las imágenes

1471x4 table

	1 imageFilename	2 Ninfa	3 Ninfa_I	4 Ninfa_P
1	'C:\Users\dikaj\D...	[188,48,34,46;62,92,26,20]	[163,155,6...	[]
2	'C:\Users\dikaj\D...	*3x4 double*	[]	[123,205,2...
3	'C:\Users\dikaj\D...	[128,191,30,22]	[164,32,61,...	[]
4	'C:\Users\dikaj\D...	*3x4 double*	[]	*3x4 double*
5	'C:\Users\dikaj\D...	[85,126,60,74;127,66,31,23]	[]	[82,111,15,...
6	'C:\Users\dikaj\D...	[110,116,56,84;75,47,31,21]	[]	[184,13,21,...
7	'C:\Users\dikaj\D...	[111,81,32,21]	[1,117,38,8...	[59.505527...
8	'C:\Users\dikaj\D...	[85,2,59,72]	[]	[]
9	'C:\Users\dikaj\D...	[173,190,21,34]	[109,1,56,7...	[]
10	'C:\Users\dikaj\D...	[125,81,56,81;49,191,22,33]	[1,1,39,75]	[]
11	'C:\Users\dikaj\D...	[1,81,57,82;129,51,30,26]	[205,211,2...	*4x4 double*
12	'C:\Users\dikaj\D...	[49,67,22,33;140,114,37,21]	[102,193,7...	*3x4 double*

Fig. 6. Banco de entrenamiento.

etiquetadas se genera la base de datos formando una tabla como se muestra en la Fig. 6, de la cual se extraen los diferentes elementos del banco de entrenamiento.

En la tabla de la Fig. 6, la primera columna representa la dirección de la imagen del banco de imágenes, en la segunda, tercera y cuarta columna se guardan los valores de las cajas delimitadoras de región de los objetos que se etiquetaron y que corresponden a las diferentes clases de objetos definidas (columna 2-Ninfa, columna 3-Ninfa_I y columna 4 Ninfa_P).

Cada caja delimitadora guarda los valores de su posición o coordenada (x, y) en la imagen, su ancho y su alto. Al utilizar la herramienta de etiquetado no se visualiza el tamaño de la caja de región de interés, lo que impidió discernir entre lo huevecillos y las ninfas de mosca blanca. Por lo que se etiquetaron todos los objetos que cumplen con forma y color.

Posterior al etiquetado, se utilizó un algoritmo de descarte de candidatas a ser ninfas si es que las cajas delimitadoras de región no cumplían con el tamaño mínimo requerido para contener las ninfas más pequeñas a ser detectadas.

Lo que se realizó en este caso fue un recorrido en la tabla de entrenamiento (Fig. 6) verificando el área de las cajas delimitadoras de región de interés (ancho por alto) y si estas resultaban más pequeñas que el área de la ninfa más pequeña (210 pixeles), estas candidatas fueron eliminadas. La lógica del algoritmo se muestra en la Fig. 7.

La Fig. 7 es un diagrama que representa la lógica del algoritmo de descarte de candidatas a ser ninfas, donde Ninfa representa la columna en la tabla de entrenamiento (Fig. 6) que contiene a las diferentes cajas delimitadoras que encierran a las ninfas y Ninfa_P representa a la columna en la tabla de entrenamiento que contiene a las diferentes cajas delimitadoras que encierran a las posibles ninfas y Am representa el área menor de la ninfa más pequeña.

Nn es el índice del ciclo que realiza el recorrido de los renglones en la tabla de entrenamiento y k es el índice que se utiliza para acceder a cada caja delimitadora contenida en la celda dentro del renglón Nn en la tabla de entrenamiento.

A cada caja delimitadora se le calcula el área, y se almacena en AN para la clase de objeto ninfa y en AN_P para la clase de objeto posibles ninfas. Posteriormente, estos

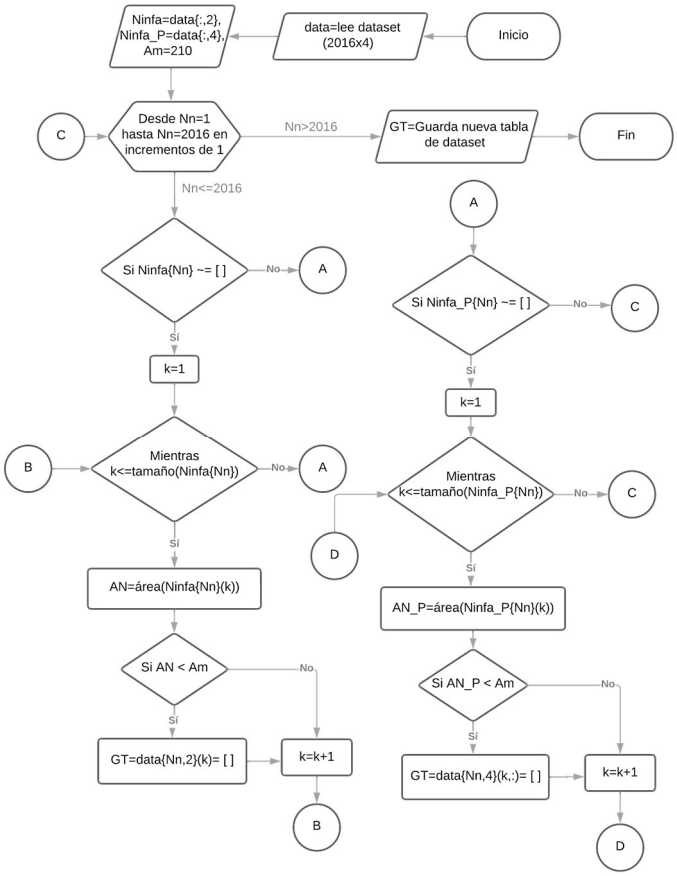


Fig. 7. Algoritmo de descarte de candidatas a ser ninfas.

valores se comparan con Am y en caso de ser menores a este, se eliminan las cajas delimitadoras a las que corresponden o en el caso contrario se mantienen en la tabla de entrenamiento nueva.

Parámetros de configuración

Los parámetros de configuración de la red YOLO V2 son las cajas de anclaje, las épocas a utilizar y el tamaño del mini lote de entrenamiento. Las entradas del sistema son el tamaño de las imágenes de entrada, el banco de entrenamiento, el número de etiquetas o clases de objetos, un modelo preentrenado como capa de extracción y una capa de extracción de características.

El tamaño de la imagen de entrada es de 224x224x3 como el tamaño de los recortes de imagen, el banco de entrenamiento que corresponde a la tabla de la Fig. 6, el número de etiquetas o clases de objetos corresponde a las que se muestran en la Fig. 3.

En el caso de la capa de extracción se seleccionó la resnet50, ya que es una red neuronal residual que ha demostrado ser útil para aumentar el rendimiento de las redes

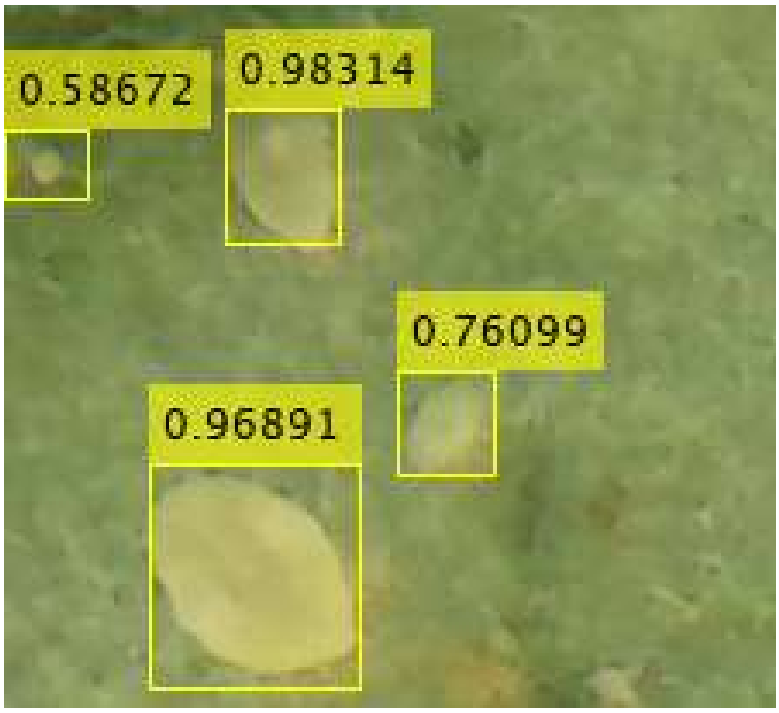


Fig. 8. Resultado del detector operando sobre una imagen del banco de imágenes.

neuronales con propósito de visión artificial y reconocimiento de objetos al realizar saltos entre capas.

Por último, para la capa de extracción de características se seleccionó una relu 40, esta genera mapas de características que se reducen en un factor de 16, la cual es una buena compensación entre la resolución espacial y la fuerza de las características extraídas. Un parámetro que se puede modificar es la cantidad de cajas de anclaje, este parámetro se utiliza para generar las cajas de anclaje con el tamaño adecuado que se posicionarán en cada celda de la imagen.

Teniendo los datos de entrada se definen las opciones de entrenamiento para la red neuronal, estas opciones se pueden modificar para generar diferentes configuraciones, en este proyecto lo que se modificó fueron las épocas y el mini lote de entrenamiento, ya que es recomendable que los otros parámetros de la configuración para la red YOLO V2 se mantengan es sus valores originales.

Una época representa un recorrido completo del algoritmo de entrenamiento a lo largo de todo el conjunto de entrenamiento, por lo que debe definirse el número de épocas a utilizar para el entrenamiento, lo que puede ayudar a mejorar la precisión del detector.

El mini lote define un subconjunto del conjunto de entrenamiento que se debe utilizar para evaluar el gradiente de la función de pérdida y actualizar los pesos, por lo que el parámetro mini lote define el tamaño que se desea usar para cada iteración de entrenamiento y este solamente se puede representar con un número entero positivo. La

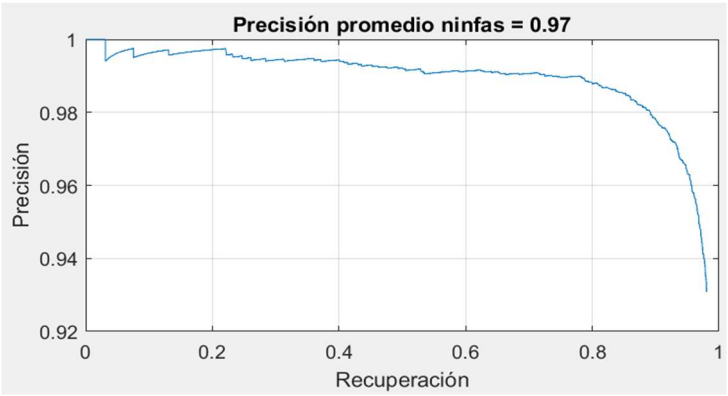


Fig. 9. Gráfica de precisión promedio al detectar ninfas de la mejor configuración.

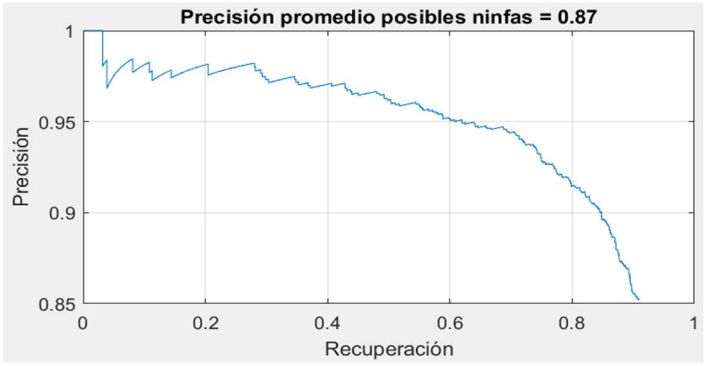


Fig. 10. Gráfica de precisión promedio al detectar posibles ninfas de la mejor configuración.

tabla 2 muestra las configuraciones que se exploraron para entrenar el sistema inteligente detector de ninfas de mosca blanca presentes en hojas de plantas.

Detector

El entrenamiento de la red entrega un objeto que contiene los parámetros de la red entrenada, aquí referido como "detector". Para validar el correcto entrenamiento del detector se pone en operación al procesar una imagen de referencia, en la Fig. 7 se muestra un ejemplo de funcionamiento correcto de la red entrenada.

En la Fig. 8 se puede observar que en la imagen de referencia se insertan los cuadros delimitadores obtenidos con la información como las cajas delimitadoras de la región de interés (x, y, ancho, alto) y el porcentaje de precisión.

4. Validación y análisis de precisión

El enfoque para validar este sistema fue mediante la configuración de la red neuronal cambiando los parámetros de épocas, cajas de anclaje y los mini lotes de entrenamiento.

Tabla 2. Configuraciones para entrenar la red.

Redes	Cajas de anclaje	Épocas	Mini lote
Red 1	5	60	2
Red 2	25	60	2
Red 3	50	60	2
Red 4	5	20	2
Red 5	5	40	2
Red 6	5	60	8
Red 7	5	60	16
Red 8	25	80	2
Red 9	25	90	2
Red 10	25	80	8

Tabla 3. Precisión promedio al detectar candidatas a ser ninfas de múltiples redes.

Red	Precisión promedio de ninfas	Precisión promedio de ninfas incompletas	Precisión promedio de posibles ninfas
Red 1	96	90	79
Red 2	96	91	83
Red 3	96	91	83
Red 4	93	74	66
Red 5	95	84	79
Red 6	96	90	79
Red 7	94	85	73
Red 8	97	92	87
Red 9	97	93	89
Red 10	97	91	85

Se analizó la precisión obtenida en la detección de las candidatas a ser ninfas. Se obtuvieron la precisión, recuperación y precisión promedio de acuerdo con la Ecuación 2, donde VP representa los verdaderos positivos, FP representa los falsos positivos y FN los falsos negativos:

$$\text{Precisión} = \frac{\text{VP}}{\text{VP}+\text{FP}}, \tag{1}$$

$$\text{Recuperación} = \frac{\text{VP}}{\text{VP}+\text{FN}}, \tag{2}$$

$$\text{Precisión Promedio} = 2 \times \frac{\text{Precisión} \times \text{Recuperación}}{\text{Precisión}+\text{Recuperación}}. \tag{3}$$

Se obtuvo la precisión promedio al utilizar cada uno de los detectores obtenidos de las configuraciones que se muestran en la tabla 2 y se recopiló la información de los resultados en la tabla 3. Al realizar pruebas con varios valores para losarámetros de cajas de anclaje, épocas y mini lote de entrenamiento se deduce la configuración del

sistema inteligente que arroja los resultados con mejor precisión promedio al detectar a las ninfas de mosca blanca presentes en hojas.

La red 9 cuenta con 150 capas y se configuró con 25 cajas de anclaje, 90 épocas y 2 mini lotes de entrenamiento, obteniendo precisiones que van desde 93% hasta un 97% al detectar candidatas a ser ninfas. En la Fig. 9 y 10 se muestran las gráficas de precisión promedio para la detección de las clases de objetos definidas.

A pesar de que los resultados que se reportan en esta sección fueron obtenidos utilizando un banco de imágenes que contiene exclusivamente hojas de tomate infestadas, la metodología propuesta puede utilizarse con imágenes de hojas de diferentes plantas infestadas de mosca blanca. En tal caso, deberá generarse una ecuación de compensación cromática diferente, así como realizar un nuevo entrenamiento de la red.

5. Conclusiones

El propósito del trabajo reportado es la detección de ninfas de mosca blanca sobre hojas de plantas, para su validación se utilizó un banco de imágenes de 224x224 píxeles de hojas de tomate infestadas con ninfas de mosca blanca.

De las diferentes pruebas de validación realizadas, se determinó que la configuración adecuada para entrenar la red consiste en utilizar 25 cajas de anclaje, 80 épocas de entrenamiento y un tamaño de mini lote de 2 para obtener los resultados con precisiones mayores. Se lograron precisiones que van del 93% al 97%, en la detección de ninfas nítidas.

En contraste, en la detección de ninfas que se ven borrosas y/o presentan formas irregulares se lograron precisiones que van desde 66% hasta 89%, en este caso, el entrenamiento y, en consecuencia, la detección, no se logra de manera eficaz debido a la falta de información característica de las ninfas en los objetos a detectar.

Por lo que, se puede concluir que si se cuenta con imágenes nítidas de las ninfas sobre hojas de plantas se pueden lograr detecciones con alta precisión al utilizar un sistema inteligente basado en la red neuronal YOLO V2.

Referencias

1. Organización de las naciones unidas para la agricultura y la alimentación. Agricultura y diálogos de cultura (2005) www.fao.org/home/es
2. El estado mundial de la agricultura y la alimentación (2021) www.fao.org/documents/card/es/c/cb4476es
3. DGSV-CNRF: Mosquita blanca Bemisia Tabachi (Gennadius). Estado de México (2020)
4. Elena-Cuéllar, M., Morales-Francisco J.: La mosca blanca Bemisia Tabaci (Gennadius) como plaga y vectora de virus en fríjol común (Phaseolus vulgaris L.). Revista Colombiana de Entomología, vol. 32, no.1 (2006)
5. Bernal, L., Pesca, L., Rodríguez, D., Cantor, F., Cure, J. R.: Plan de muestreo directo para Trialeurodes vaporariorum (Westwood) (Hemiptera: Aleyrodidae) en cultivos comerciales de tomate. Agronomía Colombiana, vol. 26, no. 2, pp. 266–276 (2008)
6. Legg, J.: Bemisia tabaci: The whitefly vector of cassava mosaic geminiviruses in Africa: An ecological perspective. African Crop Science Journal, vol. 2, no. 4, pp. 437–448 (1994)

7. Gómez-Camperos, J. A, Jaramillo, H. Y., Guerrero-Gómez, G.: Técnicas de procesamiento digital de imágenes para detección de plagas y enfermedades en cultivos: Una revisión. Ingeniería y Competitividad, vol. 24, no. 1 (2022) doi: 10.25100/iyc.24i1.10973
8. DASH Technologies: Towards Future Farming: AI is transforming the agriculture industry (2021) dashtechinc.com/towards-future-farming-ai-is-transforming-the-agriculture-industry/
9. Lino-Attyla, F. S., Silva-Brunna, C. R., Rocha-Danilo, P. C., Furriel-Geovanne, P., Calixto-Wesley, P.: Performance of haar and LBP features in cascade classifiers to whiteflies detection and counting. In: CHILEAN Conference on Electrical, Electronics Engineering, Information and Communication Technologies, pp. 1–6 (2017) doi: 10.1109/chilecon.2017.8229737
10. Bernal-López, O. F., Medina-Melendrez, M. G.: Sistema de captura y procesamiento digital de imágenes para el conteo de mosca blanca en estado ninfal sobre hojas de tomate. Tésis de Maestría (2020)
11. Organización de las naciones unidas para la agricultura y la alimentación. Can artificial intelligence help improve agricultural productivity? (2017) www.fao.org/e-agriculture/news/can-artificial-intelligence-help-improve-agricultural-productivity
12. Gómez-Camperos, J., Jaramillo, H., Guerrero-Gómez, G.: Técnicas de procesamiento digital de imágenes para detección de plagas y enfermedades en cultivos: una revisión. Ingeniería y Competitividad, vol. 24, no. 1 (2021) doi: 10.25100/iyc.v24i1.10973
13. Liakos, K., Busato, P., Moshou, D., Pearson, S., Bochtis, D.: Machine learning in agriculture: A review. Sensors, vol. 18, no. 8, pp. 2674 (2018) doi: 10.3390/s18082674

ISSN 1870-4069

Comparación de algoritmos de clasificación en el reconocimiento en ondas gravitacionales del tipo lenta, moderada y rápida

Miguel A. Avendaño-Bernal, Cesar Tiznado,
Javier M. Antelis, Claudia Moreno

Tecnológico de Monterrey,
Escuela de Ingeniería y Ciencias,
México

Universidad de Guadalajara,
Centro Universitario de Ciencias Exactas e Ingeniería,
Departamento de Física,
México

{A01733793, A00838226, mauricio.antelis}@tec.mx,
claudia.moreno@academico.udg.mx

Resumen. El 14 de septiembre de 2015 se confirmó a todo el mundo la primera detección de Ondas Gravitacionales, hoy en día muchos de los científicos teóricos, computacionales y experimentales han estado trabajando sin descanso en el desarrollo y mejora de nuevas formas de detectarlas, el siguiente objetivo es la prueba de las Ondas Gravitacionales Core-Collapse Super Novae (CCSNe). En el presente trabajo construimos un algoritmo de Machine learning, que clasifica en tres tipos de Ondas Gravitacionales (OGs) sobre CCSNe a partir de un DataSet de Espectrogramas, partiendo de la generación de 1500 Formas de Onda (FO) y con las herramientas de ML obtenemos un porcentaje de clasificación cercano al 100 % en distinción de dos estudios de las clases que se van generando, esto por la única ambición de aplicar a nuevos métodos de detección que incluyan ruido.

Palabras clave: Gravitational waves core-collapse supernovae g-mode feature phenomenological signal generation machine learning deep learning.

Comparison of Classification Algorithms in Recognizing Slow, Moderate, and Fast Type Gravitational Waves

Abstract. On September 14, 2015, the first detection of Gravitational Waves was confirmed worldwide. Nowadays, many theoretical, computational, and experimental scientists have been tirelessly working on the development and improvement of new ways to detect them, the next goal being the testing of Core-Collapse Super Novae (CCSNe) Gravitational Waves. In this work, we construct a machine learning algorithm that classifies three types of Gravitational Waves (GWs) on CCSNe from a Spectrogram DataSet, starting from the

generation of 1500 Waveforms (WFs). With ML tools, we obtain a classification percentage close to 100% in distinction of two studies of the classes that are generated. This is driven by the sole ambition to apply to new detection methods that include noise.

Keywords: Gravitational waves core-collapse supernovae g-mode feature phenomenological signal generation machine learning deep learning.

1. Introducción

En el año 1915, Einstein publicó un artículo con una nueva perspectiva de la gravedad [3], la idea era que la gravedad en sí no es una fuerza, utilizando nuevas herramientas matemáticas conocidas como Tensores para describir el movimiento de los cuerpos dando así la "Teoría de la Relatividad General" [9] con una idea particular, "El Principio de Equivalencia", llámese el pensamiento esencial que Einstein utiliza para construir su teoría [22]. En consecuencia se obtienen las Ecuaciones de Campo de Einstein o EFE (Einstein Field Equations por sus siglas en Inglés):

$$R_{\mu\nu} - \frac{1}{2} g_{\mu\nu} R = \frac{8\pi G}{c^4} T_{\mu\nu}. \tag{1}$$

Dando una solución particular conocida como Ondas Gravitacionales, que depende de la Deformación Espacio-Temporal [24]. Con la tecnología actual esta teoría se pone a prueba mediante la detección de Ondas Gravitacionales en base a la interferometría óptica que promete ser uno de las mejores herramientas a nuestra disposición [4].

Detectadas por primera vez el 14 de septiembre de 2015 por LIGO (Laser Interferometry Gravitational-Wave Observatory por sus siglas en Inglés) y Virgo [8], siendo la fuente el choque de 2 Agujeros Negros [11]. El presente trabajo exploraremos una nueva forma de detección de Ondas Gravitacionales generadas por CCSNe que se busca, sean aplicables a detectores más avanzados [7].

2. Marco teórico

Para comenzar con el análisis, necesitamos los conceptos básicos, la teoría de las supernovas de colapso del núcleo proviene de la muerte de estrellas masivas y el colapso gravitatorio que termina en una explosión de supernova, donde viene implicada la dinámica relativista del plasma en un fuerte campo gravitatorio que ambienta la ocasión [10]. La energía proviene de la masa que es superior de $M \geq 8M_{\odot}$, y el $99\,\%$ se convierte en neutrinos [26].

¿Qué ocurre con el $1\,\%$ restante de la energía gravitatoria? Esto se traduce como ondas gravitatorias, como mencionamos antes estas son deformaciones de energía que actuan sobre el espacio-tiempo propagándose a la velocidad de la luz [**?**]. En los casos de Estrellas de Neutrones, se trata de fusiones de sistemas binarios (SB) compactos (Pueden ser de Estrellas de Neutrones y/o Agujeros Negros) [13], en estos SB existe la probabilidad de detectar fuertes OGs emitidas durante un colapso gravitacional o una

explosión [6], añadiendo la búsqueda de modelos que puedan expresar el movimiento de las CCSNe, conocidos como "Modos-gτ "Modos-r"[2]. Los detectores de ondas gravitacionales de tercera generación se están preparando para la oscilación impulsada por neutrinos que se centra casi decenas de milisegundos después del rebote.

En consecuencia, la búsqueda de los OGs CCSNe en la nueva era del Detector Avanzado, debería tener un porcentaje de incertidumbre de $5\,\%$ [23], mejorando la sensibilidad al eliminar el ruido procedente de OGs de CCSNe reales [7]. Para esto, las investigaciones numéricas buscan probar simulaciones de CCSNe, que dependen fuertemente en la frecuencia que modelan la gravedad superficial (modo-g) [23].

La comprensión de los conceptos "Modo gτ sus frecuencias, aportan información útil. Ya que están extremadamente co-rrelacionados, el Modo g está asociado a la excitación de la Proto-Estrella de Neutrones (PEN), mostrando las permutaciones en el espacio por medio de su oscilación [17, 18], visto por ejemplo en el modelo [16], las oscilaciones del Modo g son convectivamente estables ya que muestran el cambio sobre su superficie.

Con la frecuencia, el análisis se acerca más a la idea anterior, con frecuencias más bajas, el estudio es menos detallado durante el post-rebote [27], ya qué no hay suficiente información que uno pueda llegar a obtener. En cambio las resoluciones de alta frecuencia son más reconocibles ya que permiten saber que tipo de fuente es y como evolucionó debido a su oscilación en ese momento [12].

El rango de frecuencias que vamos a estudiar lo tomamos de [21], siendo $f_{\text{Lenta}} = [100\,\text{Hz} - 500\,\text{Hz}]$ y $f_{\text{Rapida}} = [1000\,\text{Hz} - 2000\,\text{Hz}]$. Otro punto a tener en cuenta es la pendiente, que se calcula mediante la Ec: 2, en la que se consideran las frecuencias con el tiempo, y con una diferencia de 2 muestras tenemos la pendiente final que se desarrolla mediante los experimentos numéricos buscando valores pequeños y grandes [15]:

$$\text{Slope} = \frac{f_{\text{final}} - f_{\text{initial}}}{t_{\text{final}} - t_{\text{initial}}}. \tag{2}$$

La ecuación de la pendiente puede ser resuelta numéricamente o analíticamente en base a la complejidad del modelo utilizado [14]. Para los modelos Súper Nova la pendiente significa un resumen de la dinámica de la explosión y la información de fondo para el estudio OG [20]. La última herramienta que vamos a utilizar son los algoritmos de Machine Learning aplicados a OG, con la expectativa de mejora de las ejecuciones numéricas y disminución del ruido de las señales, analizando datos en el rango de tiempo y frecuencia [5].

Utilizando los datos de LIGO para buscar el comportamiento periódico o estocástico de los OGs y estudiarlos mediante métodos de Cross-Correlation con detecciones y eliminar las fluctuaciones y ruido que puedan aparecer y el análisis con arquitectura de Redes Neuronales de entrenamiento podría servir como ventaja de DeepClean de futuros detectores [19]. Hay algunos estudios modernos (Ver Referencia [1]), que presentan la idea de utilizar sets de datos de LIGO y ejecutarlo para reducir la tasa de falsa detección para así desarrollar la detectabilidad de CCSNe por el análisis de sensibilidad.

La búsqueda para la eliminación del ruido se ha incrementado sobre los acoplamientos no lineales y no estacionarios, incluyendo los supuestos de causalidad e invariancia temporal como fenómenos físicos que no pueden romperse [25].

3. Objetivos

Utilizando la información anterior nuestro objetivo será simular un algoritmo para generar y analizar un conjunto de muchas Formas de Onda, con esta información aplicar Herramientas de Aprendizaje Automático para clasificar en tres tipos de FO prestando atención al valor de la pendiente en el que esté, ya que esta determinará a que tipo pernecerá dado su comportamiento.

Para lograr esto, debemos definir pasos metodológicos buscando la creación del algoritmo final, además de una base datos para los casos. Dichos definirán si el objetivo tendrá éxito en un período relativamente corto de tiempo, para esta investigación hay 5 pasos para cada época que se muestra a continuación:

1. Optimizar y automatizar un código precursor que nos da un colaborador, la idea es generar muchas forma de onda de manera automática.

2. Crear todos los espectrogramas de las formas de ondas.

3. Re-dimensionar la Matriz de Datos de todos los espectrogramas a un tamaño de 28×28 píxeles y aplanarla a un vector.

4. Obtener un conjunto de datos que guarden el vector y clasificar el tipo de forma de onda en tres casos.

5. Aplicar las herramientas de machine y deep learning para obtener la precisión de la clasificación.

4. Implementación

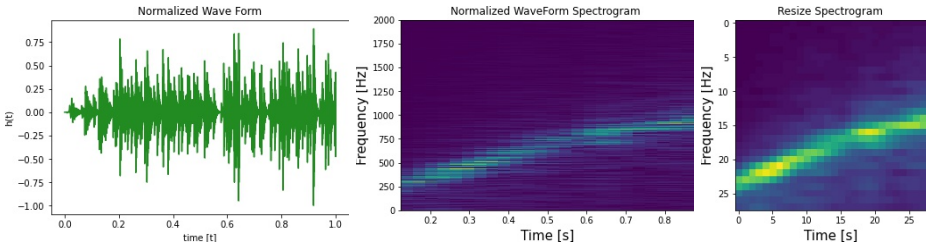
4.1. Matemáticas empleadas en el desarrollo analítico del modelo físico

La base para generar una forma de onda es una aproximación del oscilador armónico amortiguado, este modelo tiene que ser una forma exponencial que prediga el comportamiento real de una CCSNe real. Necesitamos un modelo como:

$$\ddot{x} + 2\beta\dot{x} + \omega_0^2 x = 0, \tag{3}$$

En este caso $\beta = \alpha/2m$ es el factor de amortiguamiento, el modelo necesita este parámetro para simular un efecto similar de una vibración estocástica del CCSNe, coincidiendo con la frecuencia angular ω_0 que es pieza clave para la solución de la ecuación diferencial. Obviamente esta satisface:

$$x(t) = \exp(-\beta t)\left[A_1 \exp\left(\sqrt{\beta^2 - \omega_0^2}t\right) + A_2 \exp\left(-\sqrt{\beta^2 - \omega_0^2}t\right)\right], \tag{4}$$



(a) (b) (c)

Fig. 1. Una forma de Onda junto a su espectrograma original y espectrograma resampleado.

donde A_1 y A_2 son constantes de amplitud , que necesitan ser definidas al principio de la solución numérica, el término $\exp$ son soluciones analíticas directas de muchas ecuaciones diferenciales de segundo orden, el factor de amortiguamiento se basa en la Ecuación 3 y finalmente todo depende en la región temporal.

4.2. Generación de formas de onda

Es posible generar Formas de Onda personalizadas con Matemáticas aplicadas a un script numérico, similar a la figura 1, donde las condiciones iniciales son esenciales para la resolución de la Forma de Onda que se denomina como $h(t)$ y el rango de tiempo en el que se propagó la Onda Gravitacional.

Inicialmente el script solo era capaz de generar una sola OG, era vital automatizar y generar muchas de ellas, la idea clave fue un bucle que solucionara ese error, y aplicando esto, se obtuvieron 1500 Formas de Onda. Como decimos, no se exige la información de la FO base, necesitamos los Espectrogramas de la misma. Usando la función $\mathrm{sg.spectrogram}()$ y proponiendo una frecuencia de muestreo de 16384 que es la asociada a la frecuencia de LIGO para sus detectores obtenemos unas gráficas que se muestran en la figura 1 del lado izquierdo.

Para el paso final en esta sección de trabajo, la matriz del Espectrograma fue re-dimensionada y adaptada a una matriz de sólo 28x28 píxeles, la razón de esta idea es que el análisis y el mantenimiento de los Datos podría ser más fácil para el estudio de toda la idea detrás del ML. Consecuentemente, ajustamos esta matriz a solo un vector con solo 784 componentes y concatenamos los 1500 OGs iniciales. Este re-dimensionamiento se puede observar en la figura 1 en la parte central.

A continuación, la implementación de la librería Pandas podría rescatar el redimensionamiento del Espectrograma de Datos en un excel, aplicando una clave y una columna de Datos a la Organización DataFrame. Como nota, para los diferentes Tipos de Onda, cada uno de ellos necesita tener la misma longitud de Datos. Y para cada tipo nombramos de la siguiente manera (Como es obvio, los datos se pusieron en un nuevo . csv):

$$\begin{aligned} 1 &\rightarrow \text{Lenta} \quad \forall \text{ Pendiente} \in [100-1000] \\ 2 &\rightarrow \text{Moderada} \quad \forall \text{ Pendiente} \in [1000-3000] \\ 3 &\rightarrow \text{Rápida} \quad \forall \text{ Pendiente} \in [3000-5000] \end{aligned}$$

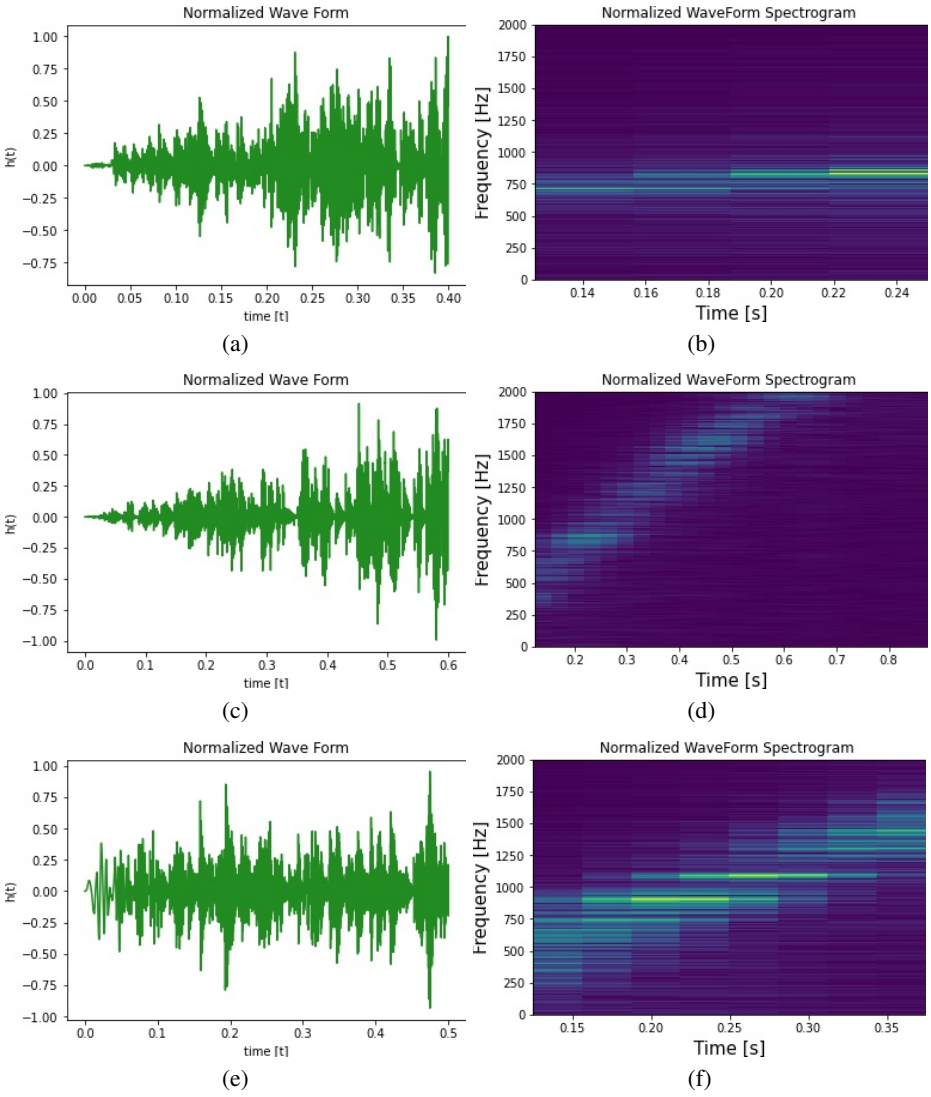


Fig. 2. Forma de onda de tipo lenta con espectrograma normal y redimensionado al valor su pendiente $\in$ $[840, 2550, 3660]$ respectivamente.

4.3. Aplicación de algoritmos de clasificación

El estudio de las clasificaciones se dividió en dos estudios, uno para los datos entrenados y otro para datos puestos a prueba. observado en el gráfico 5:

$$\text{Div. Algoritmo} = \begin{cases} \text{DE} \in 80\,\%, \\ \text{DP} \in 20\,\%. \end{cases} \tag{5}$$

Tabla 1. Valores de los espectrogramas y su tipo.

	0	1	2	···	781	782	783	Type
1	30	30	30	···	38	39	40	1
2	30	30	30	···	38	40	42	1
⋮	⋮	⋮	⋮	⋱	⋮	⋮	⋮	⋮
1499	37	36	35	···	36	35	34	2
1500	31	32	32	···	36	35	34	2

Siendo DE el conjunto de Datos de entrenamiento que permite un análisis basado en la cantidad de Datos, mientras que DP son los Datos de prueba que prueban las variables de entrenamiento para obtener un valor de predicción.

4.4. Modelos de clasificación

Utilizamos 2 tipos de estudios, uno para dos clases (Rápido y lento) y otro para tres clases (Rápido, Lento y moderado), entrenando esta información en 5 algoritmos de clasificación, ya que necesitamos comparar qué estudio es mejor que otro, ver la diferencia de cada uno de ellos y finalmente obtener una puntuación máxima en la puntuación de clasificación. Los modelos utilizados son los siguientes:

1. Logistic Regression (LR): Modelo estadístico usado para la clasificación y predicción de analíticas, ya que estima la probabilidad de que un evento ocurra dado un set de datos que mantiene las variables independientes en un rango entre [0:1].
2. Linear Discriminant Analysis (LDA): Modelo que permite la superposición de variables para la clasificación y reducción de la dimensionalidad de variables para la extracción de clasificación de patrones.
3. Support Vector Machine (SVM): Se basa en el encuentro de un hiperplano dentro de un espacio N-Dimensional que clasifica claramente los puntos de los datos encontrados dentro del hiperplano.
4. Decision Tree Classifier (DTC): Algoritmo usado para la clasificación y regresión de un conjunto de datos para crear un modelo que predice el valor de una variable que tenga como objetivo aprender decisiones simples sobre las características del modelo.
5. Random Forest Classifier (RFC): Algoritmo que combina la salida de múltiples árboles de decisión para alcanzar un resultado particular. Resolviendo los problemas de clasificación y regresión.

4.5. Métricas de clasificación

Mediante estos algoritmos la precisión era vital para saber realmente si el modelo funciona, como observamos en la ecuación 9, existen dos variables, siendo:

Tabla 2. Precisión de la clase Total, 1 y 2.

Precisión Total (2 clases)	**Media**	**STD**	**Max**	**Min**
LR	100.0	0.0	1.0	1.0
LDA	96.0	0.0	0.9	0.9
SVM	100.0	0.0	1.0	1.0
DTC	99.7	0.0	0.9	0.9
RFC	100.0	0.0	1.0	1.0
Previsión de la clase 1	**Media**	**STD**	**Max**	**Min**
LR	100.0	0.0	1.0	1.0
LDA	96.1	2.48	0.9	0.9
SVM	100.0	0.0	1.0	1.0
DTC	99.7	0.3	1.0	0.9
RFC	100.0	0.4	1.0	1.0
Previsión de la clase 2	**Media**	**STD**	**Max**	**Min**
LR	100.0	0.0	1.0	1.0
LDA	96.0	2.69	0.9	0.9
SVM	100.0	0.0	1.0	1.0
DTC	99.7	0.3	0.9	0.9
RFC	100.0	0.0	1.0	1.0

$$\mathrm{TP} = \text{Positivos Totales}, \tag{6}$$

$$\mathrm{TN} = \text{Negativos Totales}, \tag{7}$$

$$\mathrm{n} = \text{Iteraciones Totales}. \tag{8}$$

El análisis cambia para cada clase, para la precisión de una clase la ecuación 10, satisface este problema en base a la Matriz de Confusión, esta explica el rendimiento del algoritmo de una manera visual basado en la clasificación dada:

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{n}, \tag{9}$$

$$\mathrm{ACC} = \frac{\mathrm{TP}_i + \mathrm{TN}_i}{n_i}. \tag{10}$$

5. Resultados

5.1. Espectrogramas y formas de onda

En este apartado vamos a prestar especial atención a los resultados que se han generado. En primer lugar, se presentan tres ejemplos de todos los tipos de Onda-Forma que producimos, es necesario entender que la forma propia de cada onda va a ser variable con las otras dos, el comportamiento, como explicamos al principio es completamente estocástico, lo que significa que es realmente difícil de predecir de una manera exacta.

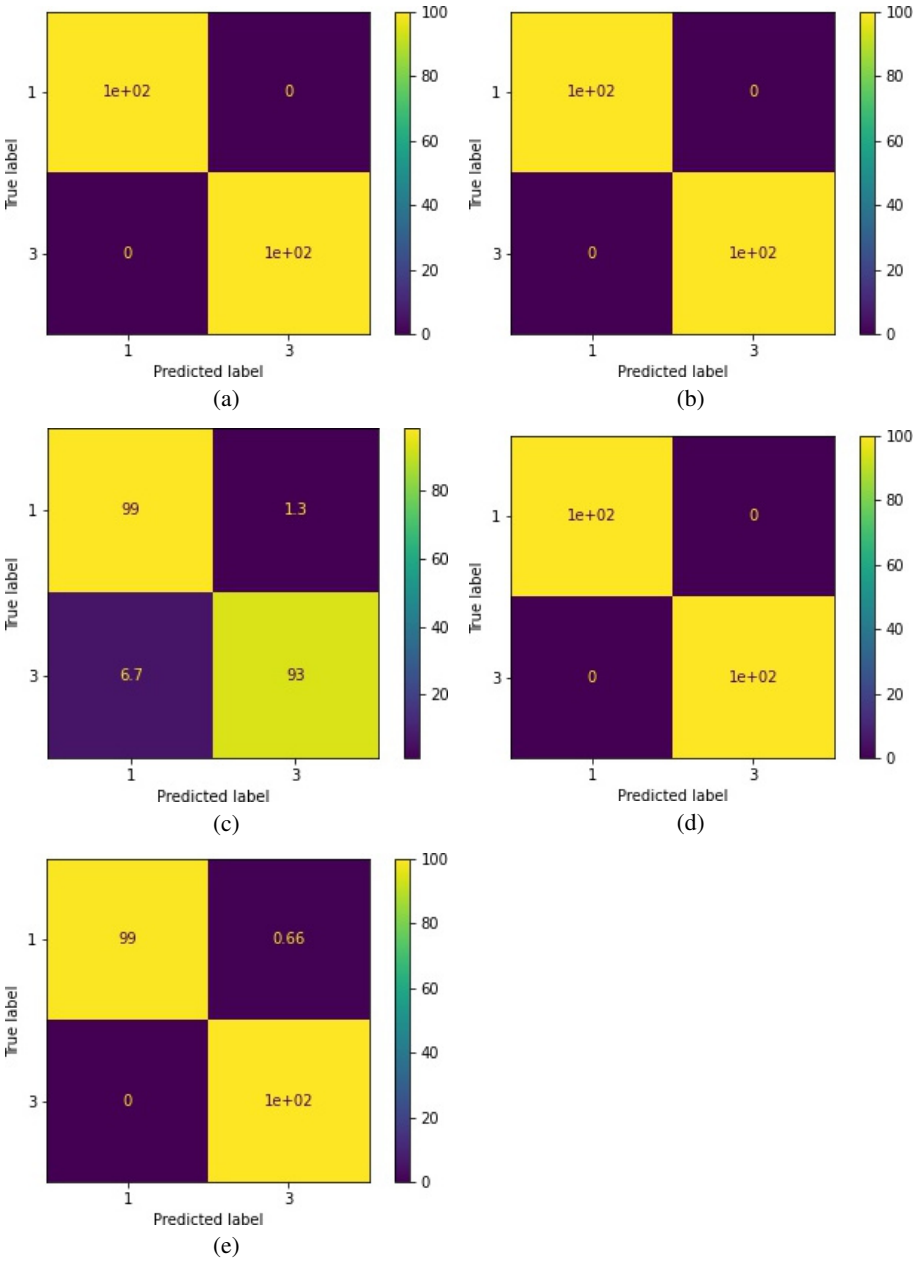


Fig. 3. Matriz de confusión para las métricas LR, LDA, SVC, DTC y RFC de izquierda a Derecha respectivamente para dos clases.

En la figura 2 presentamos un ejemplo del tipo lento donde la oscilación es más estocástica al final del recorrido, comienza lentamente y al final es completamente violenta (Imágenes superiores).

Tabla 3. Precisión de la clase Total, 1, 2 y 3.

Precisión Total (3 clases)	**Media**	**STD**	**Max**	**Min**
LR	100.0	0.0	1.0	1.0
LDA	98.7	1.0	0.9	0.9
SVM	100.0	0.0	1.0	1.0
DTC	100.0	0.0	1.0	1.0
RFC	100.0	0.0	1.0	1.0
Precisión de la clase 1	**Media**	**STD**	**Max**	**Min**
LR	100.0	0.0	1.0	1.0
LDA	98.6	1.0	0.9	0.9
SVM	100.0	0.0	1.0	1.0
DTC	100.0	0.0	1.0	1.0
RFC	100.0	0.0	1.0	1.0
Precisión clase 2	**Media**	**STD**	**Max**	**Min**
LR	100.0	0.0	1.0	1.0
LDA	98.7	1.0	0.9	0.9
SVM	100.0	0.0	1.0	1.0
DTC	100.0	0.0	1.0	1.0
RFC	100.0	0.0	1.0	1.0
Precisión clase 3	**Media**	**STD**	**Max**	**Min**
LR	100.0	0.0	1.0	1.0
LDA	98.7	1.0	0.9	0.9
SVM	100.0	0.0	1.0	1.0
DTC	100.0	0.0	1.0	1.0
RFC	100.0	0.0	1.0	1.0

Siguiendo por otro ejemplo tipo moderado en la figura 2, observamos que son más lentas y pasivas, de hecho sólo hay un comportamiento estocástico sólo unos instantes al final, pero la mayor parte de la longitud de la trayectoria es completamente suave (Figuras de en medio). Y para terminar el tipo rápido con la figura 2, es claramente el tiene el comportamiento más violento y complejo es impredecible y completamente estocástico, incluso al principio (Imágenes Inferiores).

A continuación, cómo habíamos dicho el cálculo de los espectrogramas es el siguiente paso, de nuevo, vamos a mostrar las gráficas de las mismas formas de onda que están presentes en la misma figura distribuido de la misma forma previamente dicha en la figura 2. Para el primer caso (El tipo Lento), se ve claramente que el valor de la pendiente es pequeño, debido al tipo de FO, y por eso la intensidad que se representa con la línea verde en todo el espectrograma no aumenta mucho con un valor de pendiente de tipo lento.

A continuación está el tipo moderado, en este caso la pendiente crece en comparación con el tipo anterior, y de nuevo es debido al valor de la pendiente, los valores son más altos y tienen más cambio pero respectivamente el tiempo tiene una disminución en la evaluación del mismo mostrada en la Figura 2.

Por último, el tipo Rápido está relacionado con el valor de la pendiente, en este caso tenemos un equilibrio con los parámetros que se están evaluando respecto al tiempo y tiene los mayores valores de frecuencias, la pendiente (Representada con la línea verde) se mantiene sin ningún cambio drástico.

5.2. Data set de los espectrogramas redimensionados

No necesitamos las gráficas de los Espectrogramas, la matriz de información tiene que ser la clave para el análisis final, en este caso con la herramienta $save$, se generó el DataSet, como vemos en la tabla 1, hay un total de 1500 filas con 784 columnas, el primer elemento se anticipa al total de CCSNe WF generados y el segundo elemento es adecuado para la longitud de una matriz de tamaño 28x28 en un solo vector.

5.3. Estudio de 2 clases (algoritmos de clasificación)

Finalmente con todos los Datos recompilados, es hora de utilizar las herramientas de Machine Learning y aplicar las métricas que mencionamos anteriormente a este contexto, pero ¿Por qué necesitamos el estudio para 2 y 3 clases? La primera es que 2 clases nos permiten el análisis y clasificación en sólo 1000 Formas de onda con el tipo de carácter de rápido y lento, nos traen un estudio más rápido que ampliar la precisión efectiva de la clasificación, existen algunos puntos que es importante saber, como:

– **Ventajas:**

1. Bajo coste computacional: Debido a la poca cantidad de datos que se tienen que además, soporta la computadora para que su análisis sea eficiente.
2. Análisis profundo en las 2 formas de onda: Como solo se están tomando dos tipos de onda, se da un análisis más enfocado para estas formas de onda que permite un muestreo más profundo para cada tipo.

– **Desventajas:**

1. Con sólo 1000 formas de onda la clasificación esta podría estar sesgada o tener un aumento de porcentaje de error al agregarle ruido:

   Ya que solo se están tomando 2 tipos de FO para su análisis, cuando se le añada ruido es posible que el porcentaje de error suba demasiado ya que no está considerando muchas variables de FO o diferentes condiciones iniciales que sí pueden afectar a un análisis más amplio, la solución a esto es tomar muchos más datos sobre tipos, formas de onda con una gran variedad de condiciones iniciales, peso de prueba y comportamiento de la onda.

Con estos puntos en mente, el análisis debe tener una buena puntuación en la clasificación, en la tabla 2, presentamos la Precisión total para las dos clases que estudiamos y de manera particular para cada una de ellas. Es importante notar que la puntuación media es realmente buena, casi cercana al 100 % en las cinco métricas, no tenemos desviación estándar y la información obtenida nos muestra que nuestro trabajo estuvo realmente cerca de un algoritmo eficiente.

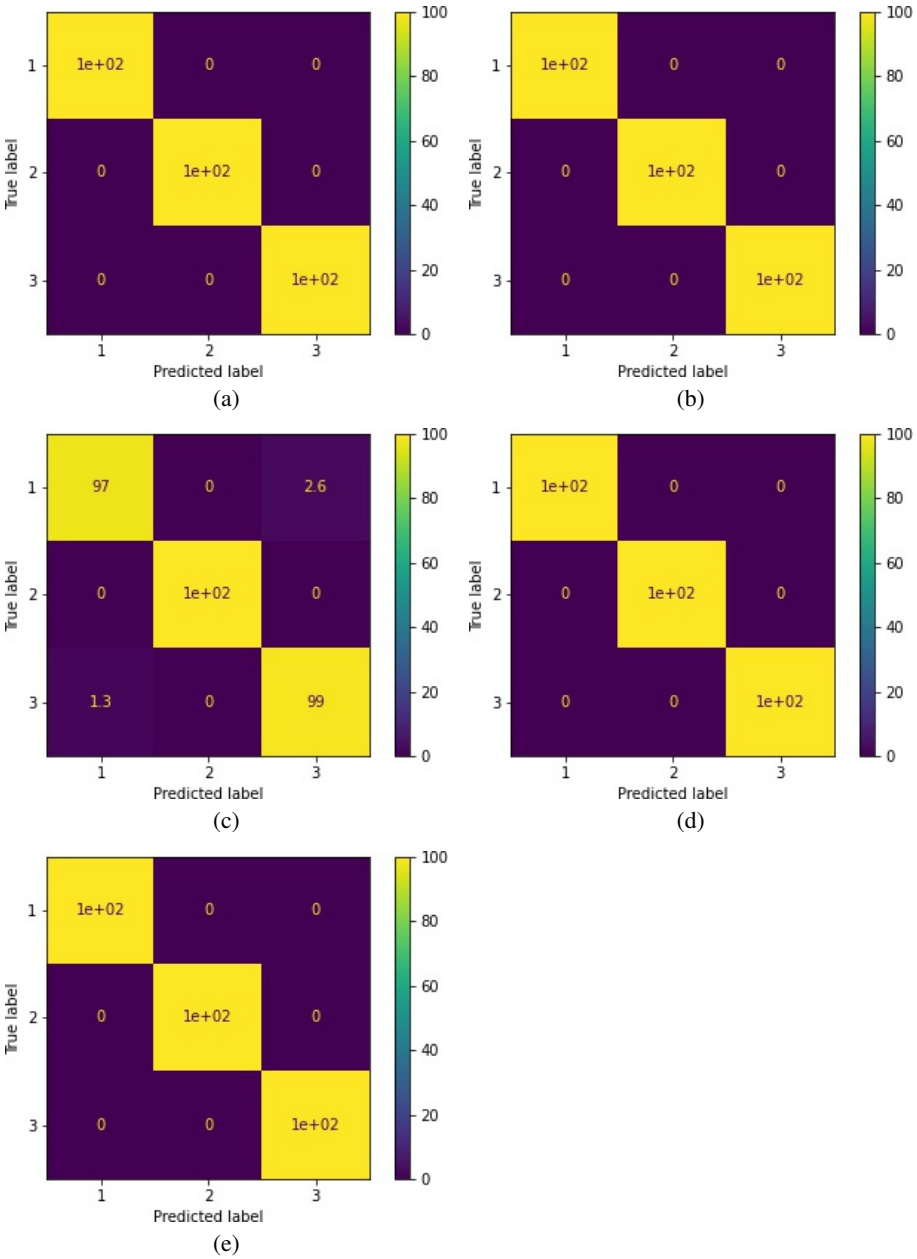


Fig. 4. Matriz de confusión para 3 clases con metricas LR, LDA, SVC, DTC y RFC.

La precisión total tiene que ser la media en la precisión de la clase 1 y clase 2, pero con la información separada podemos ver más profundo. Para la clase 1 (Véase el segundo cuadro de la tabla previa), se observa que para la mayoría de las métricas (Excepto las métricas DTC y LDA), los valores son relativamente más cercanos para

el 100 %, lo que significa que para la primera clase (Que depende del tipo Lento) su clasificación dará resultados mejores. Seguido por la Clase 2 (Tercer cuadro de la tabla previa), los resultados son inferiores a su contraparte excepto en la métrica DTC donde el STD tiene un valor más alto que la otra clase, al menos la clasificación se mantiene en una buena puntuación.

Ahora, necesitamos saber ¿Qué pasa con la matriz de confusión? Bueno, podemos anticipar que si las tablas muestran un buen puntaje de precisión, y están siendo evaluadas a partir de la matriz de confusión, el porcentaje de las mismas será el más cercano a este valor.

Y de hecho, el gráfico para la mayoría de la matriz de confusión es relativamente alta y tiene un montón de precisión, vemos en la figura que para la métrica LR y LDA (Visto en la figura 3), que el porcentaje de clases arroja valores iguales, esto representa que ambas métricas son lo suficientemente buenos para analizar los datos, mientras que para las métricas SVC, DTC y RFC los valores varían en un porcentaje relativamente bajo.

5.4. Estudio de 3 clases (algoritmos de clasificación)

Al finalizar el estudio para las dos clases nos preguntamos, ¿Podemos hacer un estudio de tres clases?, la diferencia será el trato con todos los tipos que generemos, incluyendo el tipo moderado, rápido y lento. De esta forma conseguimos un análisis más ambicioso, ahora los principales puntos a tener en cuenta son:

– **Ventajas:**

1. Un análisis más grande que toma todos las FOs: Al tener un análisis un poco más amplio, los algoritmos de clasificación tienen la posibilidad de realizar un muestreo más amplio para obtener mejores resultados al momento de su clasificación.
2. La espera a resultados similares: A partir del estudio de 2 clases, se obtuvieron buenos resultados, al agregar otro tipo se espera que los mismos algoritmos mejoren los resultados previos ya que tienen más datos que poner a prueba.

– **Desventajas:**

1. El mayor coste computacional: Por lo mismo que son muchos datos, el tiempo computacional aumenta, así mismo también las características de la computadora para que soporte la simulación, la solución a esto es primero encontrar una optimización del código base y obtener una computadora de mejor rendimiento en caso de no lograr una mejora.

Además de las otras puntuaciones Media y ETS añadidas páginas atrás, las estimamos pero para este estudio, con resultados similares, podemos ver que para la Precisión Total (Ver en la parte superior de la tabla 3), la media de las clases juntas obtiene un porcentaje mayor que para una sola clase. Ahora, en el análisis de cada clase podemos notar que los resultados son realmente similares o iguales a la precisión total, la clase 1 y la clase 2 (Ver en la parte media de la tabla 3) obtienen elevados de las 2 clases estudio.

Y la clase 3 (Ver en la parte inferior de la tabla 3) consiguen de igual forma un alto porcentaje, el único contraste se ve en las métricas LDA y STD que varían sólo en pocos valores. Con los cálculos de estimación se está calculando la matriz de confusión dando mejores resultados que el otro estudio, en este caso, la precisión para cada clase da valores más cercanos al 100 %, para las métricas LR y LDA los resultados se mantienen básicamente iguales al otro estudio (Visto en la figura 4). Pero para las métricas SVC, DTC y RFC, los porcentajes se elevan al 100 %, la mayoría de la matriz de confusión da este valor.

6. Conclusiones

Como se observó, el trabajo presentado da un algoritmo de clasificación que realiza el objetivo principal, el almacenamiento de los Datos del Espectrograma y la aplicación de herramientas de Machine Learning obtienen los resultados anteriores, de ellos podemos notar que la optimización del algoritmo fue un éxito, ya que puede generar más de 1500 Formas de Onda con el tiempo debido, este factor hace funcionar el código por aproximadamente 10 horas, si queremos hacer más, la paciencia será nuestra aliada. El proceso de entrenamiento es capaz de reconocer el tipo de Onda Gravitacional que estamos buscando. En este sistema, debido al bajo porcentaje de ruido, la clasificación es realmente buena, cabe mencionar que si trabajamos con más Datos el porcentaje puede disminuir con la adición de ruido.

Finalmente las Ondas Gravitacionales están alcanzando un auge nunca antes visto, la astronomía de por sí nos permite conocer su comportamiento y la evolución del universo. Para un trabajo futuro, este trabajo necesita optimizar y mejorar el modelo implementado en las simulaciones numéricas para obtener más OGs. Después es necesario añadir ruido real de LIGO+VIRGO a las señales entrenadas para probar la precisión del modelo. Esto debería guiarnos en un nuevo entendimiento de estos objetos.

Referencias

1. Antelis, J. M., Cavaglia, M., Hansen, T., Morales, M. D., Moreno, C., Mukherjee, S., Szczepańczyk, M. J., Zanolin, M.: Using supervised learning algorithms as a follow-up method in the search of gravitational waves from core-collapse supernovae. Physical Review D, vol. 105, no. 8 (2022) doi: 10.1103/physrevd.105.084054
2. Caride, S., Inta, R., Owen, B. J., Rajbhandari, B.: How to search for gravitational waves from r-modes of known pulsars. Physical Review D, vol. 100, no. 6 (2019) doi: 10.1103/physrevd.100.064013 https://doi.org/10.1103/physrevd.100.064013
3. Charles W. Misner, J. W., Kip S. Thorne: Gravitation (2017)
4. Corda, C.: Interferometric detection of gravitational waves: The definitive test for general relativity. International Journal of Modern Physics D, vol. 18, no. 14, pp. 2275–2282 (2009) doi: 10.1142/S0218271809015904
5. Cuoco, E., Powell, J., Cavaglià, M., Ackley, K., Bejger, M., Chatterjee, C., Coughlin, M., Coughlin, S., Easter, P., Essick, R., Gabbard, H., Gebhard, T., Ghosh, S., Haegel, L., Iess, A., Keitel, D., Márka, Z., Márka, S., Morawski, F., Nguyen, T., et al.: Enhancing gravitational-wave science with machine learning. Machine Learning: Science and Technology, vol. 2, no. 1, pp. 011002 (2020) doi: 10.1088/2632-2153/abb93a

6. Fryer, C. L., New, K. C. B.: Gravitational waves from gravitational collapse. Living Reviews in Relativity, vol. 6, no. 1 (2003) doi: 10.12942/lrr-2003-2
7. Gossan, S. E., Sutton, P., Stuver, A., Zanolin, M., Gill, K., Ott, C. D.: Observing gravitational waves from core-collapse supernovae in the advanced detector era. Physical Review D, vol. 93, no. 4 (2016) doi: 10.1103/physrevd.93.042002
8. Hajime Sotani, T. T.: Dimension dependence of numerical simulations on gravitational waves from protoneutron stars (2020)
9. Hartle, J. B., Dray, T.: Gravity: An introduction to Einstein's general relativity. American Journal of Physics, vol. 71, no. 10, pp. 1086–1087 (2003) doi: 10.1119/1.1604390
10. Janka, H. T., Langanke, K., Marek, A., Martinez-Pinedo, G., Muller, B.: Theory of core-collapse supernovae. Physics Reports, vol. 442, no. 1-6, pp. 38–74 (2007) doi: 10.1016/j.physrep.2007.02.002
11. Ju, L., Blair, D. G., Zhao, C.: Detection of gravitational waves. Reports on Progress in Physics, vol. 63, no. 9, pp. 1317–1427 (2000) doi: 10.1088/0034-4885/63/9/201
12. Kawahara, H., Kuroda, T., Takiwaki, T., Hayama, K., Kotake, K.: A linear and quadratic time–frequency analysis of gravitational waves from core-collapse supernovae. The Astrophysical Journal, vol. 867, no. 2, pp. 126 (2018) doi: 10.3847/1538-4357/aae57b
13. Lasky, P. D.: Gravitational waves from neutron stars: A review. Publications of the Astronomical Society of Australia, vol. 32 (2015) doi: 10.1017/pasa.2015.35
14. Manam, S. R., Kaligatla, R. B.: A mild-slope model for membrane-coupled gravity waves. Journal of Fluids and Structures, vol. 30, pp. 173–187 (2012) doi: 10.1016/j.jfluidstructs.2012.01.003
15. Martin Phillips, O.: Theoretical and experimental studies of gravity wave interactions. In: Proceedings of the Royal Society of London, Series A, Mathematical and Physical Sciences, vol. 299, pp. 104–119 (1967) doi: 10.1098/rspa.1967.0125
16. Mezzacappa, A., Marronetti, P., Landfield, R. E., Lentz, E. J., Yakunin, K. N., Bruenn, S. W., Hix, W. R., Messer, O. B., Endeve, E., Blondin, J. M., Harris, J. A.: Gravitational-wave signal of a core-collapse supernova explosion of a 15 $M_{\odot}$. Physical Review D, vol. 102, no. 2 (2020) doi: 10.1103/physrevd.102.023027 https://doi.org/10.1103/physrevd.102.023027
17. Morozova, V., Radice, D., Burrows, A., Vartanyan, D.: The gravitational wave signal from core-collapse supernovae. The Astrophysical Journal American Astronomical Society, vol. 861, no. 1, pp. 10 (2018) doi: 10.3847/1538-4357/aac5f1
18. Murphy, J. W., Ott, C. D., Burrows, A.: A model for gravitational wave emission from neutrino-driven core-collapse supernovae. The Astrophysical Journal, vol. 707, no. 2, pp. 1173–1190 (2009) doi: 10.1088/0004-637x/707/2/1173
19. Ormiston, R., Nguyen, T., Coughlin, M., Adhikari, R. X., Katsavounidis, E.: Noise reduction in gravitational-wave data via deep learning. Physical Review Research American Physical Society, vol. 2, no. 3, pp. 033066 (2020)
20. Powell, J., Müller, B.: Gravitational wave emission from 3d explosion models of core-collapse supernovae with low and normal explosion energies. Monthly Notices of the Royal Astronomical Society, vol. 487, no. 1, pp. 1178–1190 (2019) doi: 10.1093/mnras/stz1304
21. Radice, D., Morozova, V., Burrows, A., Vartanyan, D., Nagakura, H.: Characterizing the gravitational wave signal from core-collapse supernovae. The Astrophysical Journal American Astronomical Society, vol. 876, no. 1, pp. L9 (2019) doi: 10.3847/2041-8213/ab191a
22. Sciama, D. W.: The physical structure of general relativity. Reviews of Modern Physics, vol. 36, pp. 463–469 (1964) doi: 10.1103/RevModPhys.36.463
23. Srivastava, V., Ballmer, S., Brown, D. A., Afle, C., Burrows, A., Radice, D., Vartanyan, D.: Detection prospects of core-collapse supernovae with supernova-optimized third-generation

gravitational-wave detectors. Physical Review D, vol. 100, pp. 043026 (2019) doi: 10.1103/PhysRevD.100.043026
24. Thorne, K. S.: Gravitational waves (1995) doi: 10.48550/ARXIV.GR-QC/9506086
25. Vajente, G., Huang, Y., Isi, M., Driggers, J. C., Kissel, J. S., Szczepanczyk, M. J., Vitale, S.: Machine-learning nonstationary noise out of gravitational-wave detectors. Physical Review D, American Physical Society, vol. 101, pp. 042003 (2020) doi: 10.1103/PhysRevD.101.042003
26. Warren, M. L., Couch, S. M., O'Connor, E. P., Morozova, V.: Constraining properties of the next nearby core-collapse supernova with multimessenger signals. The Astrophysical Journal American Astronomical Society, vol. 898, no. 2, pp. 139 (2020) doi: 10.3847/1538-4357/ab97b7
27. Yakunin, K. N., Marronetti, P., Mezzacappa, A., Bruenn, S. W., Lee, C.-T., Chertkow, M. A., Hix, W. R., Blondin, J. M., Lentz, E. J., Messer, O. E. B., Yoshida, S.: Gravitational waves from core collapse supernovae. Classical and Quantum Gravity, vol. 27, no. 19, pp. 194005 (2010) doi: 10.1088/0264-9381/27/19/194005

ISSN 1870-4069

Análisis y clasificación de señales electroencefalográficas para el control de una órtesis robótica de mano con una interfaz cerebro computador basada en el paradigma de imaginación motora

Diego Sánchez González[1], Johann Barragán[2],
Omar Mendoza-Montoya[1], Javier M. Antelis[1]

[1] Tecnológico de Monterrey,
Escuela de Ingeniería y Ciencias,
México

[2] Universidad Autónoma de Bucaramanga,
Colombia

{A00573206, omendoza83, mauricio.antelis}@tec.mx,
jbarragan262@unab.edu.co

Resumen. El análisis de señales de electroencefalografía (EEG) en frecuencia se ha convertido en una herramienta poderosa para el estudio de la actividad cerebral durante diferentes tareas cognitivas y motoras. En este estudio, se utilizó esta técnica para analizar los datos de 10 participantes jóvenes que realizaron el paradigma de imaginación motora para el control de una órtesis robótica llamada Hand of Hope (HoH) para la rehabilitación de pacientes con enfermedades cardiovasculares e infartos cerebrales. La adquisición de la señal tuvo lugar en las etapas de entrenamiento y validación en línea. Los datos de entrenamiento fueron usados para calibrar un modelo de inteligencia artificial conocido como Filter Bank Common Spatial Pattern (FBCSP) para la extracción de características y un análisis discriminante lineal (LDA) como clasificador. Posteriormente, se realizó un análisis en frecuencia de la señal, con lo cual se estimó la desincronización de la potencia espectral entre las tareas de relajación e imaginación y la correlación entre estas señales en el dominio de la frecuencia. Los datos de validación en línea fueron usados para evaluar los rendimientos y capacidad de los participantes para mover la órtesis robótica, además del tiempo de detección de la señal. Los resultados indicaron una desincronización en la potencia espectral de los canales C3 y CP3 en las bandas alpha y theta durante la imaginación motora en comparación con el estado de relajación, lo que sugiere una activación de la corteza motora, sobre la que están posicionados los electrodos usados.

Palabras clave: Análisis frecuencial, inteligencia artificial, electroencefalografía, imaginación motora, interfaz cerebro computadora.

Diego Sánchez González, Johann Barragán, Omar Mendoza-Montoya, Javier M. Antelis

Analysis and Classification of Electroencephalographic Signals for the Control of a Robotic Hand Orthosis Using a Brain-Computer Interface Based on the Motor Imagery Paradigm

Abstract. The frequency analysis of electroencephalography (EEG) signals has emerged as a powerful tool for studying brain activity during different cognitive and motor tasks. In this study, we employed this technique to analyze data from 10 young participants who performed the motor imagery paradigm for the control of a robotic orthosis, known as "Hand of Hope" (HoH), used for the rehabilitation of patients with cardiovascular diseases and stroke. Signal acquisition took place during both the training and online validation stages. The training data was used to calibrate an artificial intelligence model, known as the Filter Bank Common Spatial Pattern (FBCSP), for feature extraction, and a linear discriminant analysis (LDA) as a classifier. This was followed by a frequency analysis of the signal, through which the desynchronization of spectral power between relaxation and imagination tasks was estimated, as well as the correlation between these signals in the frequency domain. The online validation data was used to evaluate the participants' performance and ability to move the robotic orthosis, as well as signal detection time. The results indicated a desynchronization in spectral power of the C3 and CP3 channels in the alpha and theta bands during motor imagery compared to the relaxed state, suggesting an activation of the motor cortex over which the used electrodes were positioned.

Keywords: Frequency analysis, artificial intelligence, electroencephalography, motor imagery, brain-computer interface.

1. Introducción

Los accidentes cerebrovasculares son una de las principales causas de discapacidad y muerte en todo el mundo. En México, el accidente cerebrovascular es la segunda causa principal de muerte después de las enfermedades del corazón. Según la Secretaría de Salud de México, cada año se producen alrededor de 130,000 casos de accidentes cerebrovasculares, y de estos, alrededor del 25 % resultan en la muerte del paciente [5].

Para abordar estos desafíos, la neuroingeniería ha surgido como una disciplina en rápido crecimiento que tiene como objetivo desarrollar tecnologías innovadoras para ayudar a los pacientes que han sufrido un accidente cerebrovascular a recuperarse y mejorar su calidad de vida.

Una de las áreas de enfoque en la neuroingeniería es la investigación y el desarrollo de interfaces cerebro-computadora o Brain Computer Interfaces (BCI), que permiten a los pacientes controlar dispositivos electrónicos y mecánicos mediante señales cerebrales. Las BCI también tienen el potencial de mejorar el proceso de rehabilitación al proporcionar una retroalimentación inmediata y precisa a los pacientes y los terapeutas.

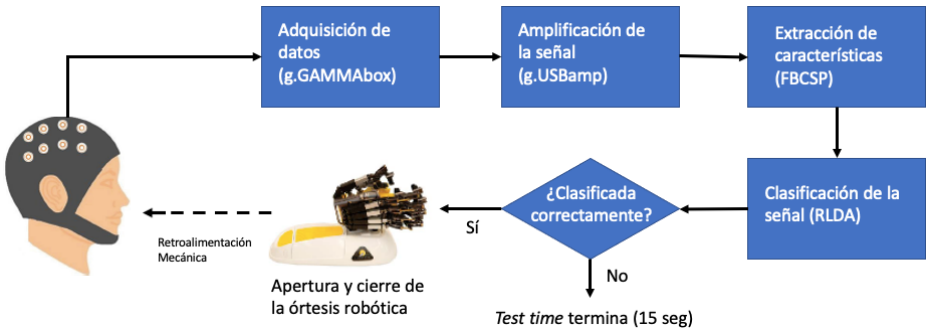


Fig. 1. Diagrama de bloques del sistema BCI+HoH.

Esto puede ayudar a los pacientes a adaptarse a los cambios en su capacidad cognitiva y física y mejorar su capacidad para realizar actividades cotidianas. Para lograrlo, las BCI pueden hacer uso del paradigma de Imaginación Motora, Motor Imagery (MI), para producir los cambios en la actividad cerebral. En este paradigma, el usuario imagina el movimiento sin realizarlo. Esto induce un cambio en la potencia de las bandas de frecuencia de las señales de EEG obtenidas de la corteza motora.

Aunado a que las BCI basadas en EEG son una opción no invasiva para controlar dispositivos, este sistema resulta idóneo para su implementación en pacientes con las patologías previamente descritas [3]. Este artículo no solo se centra en analizar señales de EEG en frecuencia, las cuales fueron adquiridas a partir de un experimento diseñado con base al paradigma de imaginación motora, sino que además estudia distintas métricas evaluadas a partir de la implementación del algoritmo de clasificación.

2. Métodos y materiales

2.1. Descripción del experimento

La BCI implementada es una interfaz basada en EEG (por tanto, no invasiva y con buena resolución temporal) con etapas de entrenamiento y validación en línea, donde la primera adquiere los datos para entrenar al modelo y la segunda utiliza al efector HoH para otorgar una retroalimentación mecánica al participante al ejecutar el movimiento de apertura y cierre de la mano derecha de forma automática. En la (Fig. 1) se muestra el diagrama de bloques de la BCI con todos sus componentes.

El experimento consistió en que el participante estuviera sentado en una posición cómoda frente a un computador que le mostrara la secuencia de imágenes y sonidos para guiarle a través del experimento. Durante la etapa de entrenamiento se le permitió tener sus extremidades superiores en cualquier posición que juzgara más cómoda, mientras que durante la etapa de validación en línea se le sujetó la mano derecha a la órtesis robótica para permitir la retroalimentación motora del sistema.

Cada etapa consistió de 24 pruebas, con una duración de 6 minutos con 24 segundos para el entrenamiento y 9 minutos con 36 segundos para la validación en línea (como máximo, debido a que el tiempo de las pruebas durante la segunda etapa variaba con respecto al tiempo de clasificación de la señal).

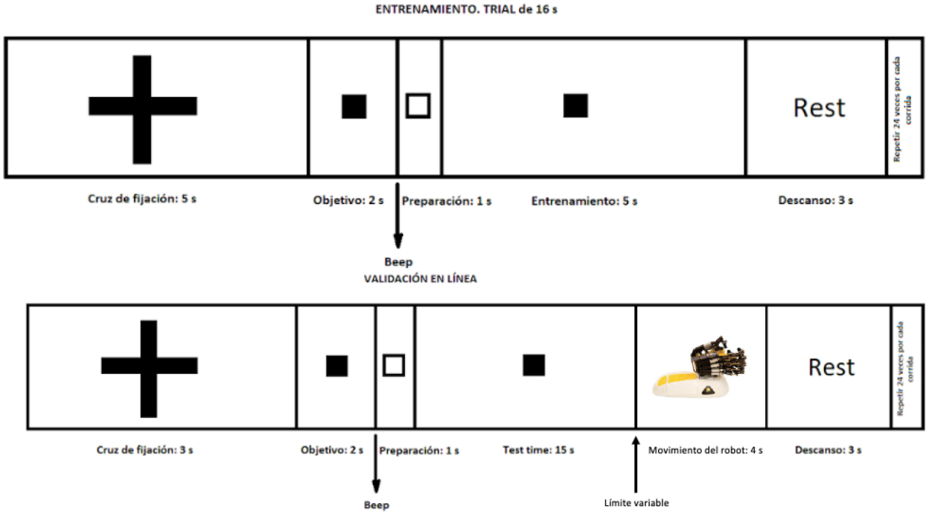


Fig. 2. (a) Línea temporal del entrenamiento. (b) Línea temporal de la validación en línea.

El experimento se realizó en un ambiente visual y acústicamente aislado para permitir la mayor concentración del participante y evitar distracciones que pudieran generar artefactos en la señal. En la figura (Fig. 2) se presenta la línea temporal del experimento para el entrenamiento y la validación en línea.

El experimento propuesto involucró a 10 participantes sanos de la Universidad Tecnológico de Monterrey (5 hombres, 5 mujeres; edad: 21 ($\pm$2) años; todos diestros) sin ningún historial clínico de enfermedades cardiovasculares como accidentes cerebrovasculares o trastornos musculoesqueléticos. Todos aceptaron proporcionar un consentimiento informado por escrito antes de su participación.

El experimento fue facilitado por el Laboratorio de Neurotecnología (NTLab) del Tecnológico de Monterrey Campus Guadalajara. Ambas etapas hicieron uso del mismo equipo, a excepción del dispositivo de terapia Hand of Hope de Rehab Robotics (Universidad Politécnica de Hong Kong) que fue usado y conectado inalámbricamente con el sistema de clasificación para permitir la retroalimentación mecánica durante la segunda etapa del experimento.

Este dispositivo cuenta con un amplio rango de movimiento y una alta precisión en los movimientos, lo que permite una amplia gama de ejercicios de terapia para la mano y el brazo. También cuenta con sensores de presión en la superficie de contacto y con un sistema de control de fuerza que permite ajustar la cantidad de resistencia en los movimientos.

La fase de adquisición de señales fue hecha con un dispositivo de adquisición (g.USBamp, g.tec, Austria), el cual cuenta con seguridad médica tipo II y puede adquirir señales de hasta 16 canales simultáneamente, permite una frecuencia de muestreo de hasta 1200 Hz, así como una ganancia de hasta 200.000 veces la amplitud original, además de estar diseñado para minimizar el ruido de las señales de EEG durante la adquisición, lo que garantiza una alta calidad de las señales adquiridas. También fueron usados 9 electrodos activos (g.SCARABEO) húmedos (g.GAMMA gel).

Los canales C3, C1, Cz, C2, C4, CP3, CPz y CP4 fueron colocados sobre la corteza motora y lóbulos parietales, y se usó AFz como tierra (GND) y el lóbulo derecho de la oreja como referencia (REF). Las señales de EEG fueron obtenidas a una frecuencia de muestreo de 256 Hz y digitalmente filtradas con un filtro pasa bandas con frecuencias de corte de 4 Hz a 60 Hz, usando un filtro Butterworth de octavo orden en cada electrodo.

Antes de comenzar la adquisición de datos se procuró que la señal fuera de buena calidad, para ello se revisó que la señal se mantuviera dentro del rango de -100 μV a 100 μV y que no existiera ningún artefacto presente como el latido del corazón. Adicionalmente, un software desarrollado internamente en C++ fue usado para manejar y controlar las señales de EEG, guardar los datos, y procesarlos tanto offline como online (Copyright @ 2018 Instituto Tecnológico y de Estudios Superiores de Monterrey).

2.2. Estudios con datos de entrenamiento

El análisis offline de los datos de entrenamiento comenzó con la importación de los archivos en formato Electronic Batch Report (EBR) a MATLAB (MATLAB. (2022b). Natick, Massachusetts: The MathWorks Inc.).

Densidad de potencia espectral (PSD). En general, la Transformada de Fourier (FT) es la operación matemática usada para llevar una señal del dominio del tiempo al de la frecuencia. Existe además la noción de la Transformada Discreta de Fourier (DFT) (eq. 1) que usa señales discretas. Ambas son sin embargo ineficientes temporalmente, por lo que se usa un algoritmo llamado la Transformada Rápida de Fourier (FFT), cuya versión estocástica es la PSD [8]:

$$X_k = \sum_{n=0}^{N-1} x_n e^{-\frac{2\pi i}{N} nk}, \quad k = 0, 1, \ldots, N-1. \tag{1}$$

La PSD es una medida utilizada para describir la distribución de energía de una señal en diferentes frecuencias. En el contexto del EEG, la PSD se refiere a la cantidad de energía de la señal que se encuentra en diferentes bandas de frecuencia del espectro de frecuencia de la señal.

Este parámetro es importante en el análisis de EEG porque proporciona información sobre la actividad eléctrica del cerebro en diferentes frecuencias. Diferentes patrones de actividad de EEG en diferentes frecuencias se han asociado con diferentes estados mentales y procesos cognitivos.

Por lo tanto, el análisis de la PSD puede proporcionar información valiosa sobre la función cerebral y puede ser útil en la investigación y el diagnóstico de trastornos neurológicos y psiquiátricos. Para calcular la PSD, se segmenta la señal discreta en ventanas solapadas cuyo valor de solapamiento es variable, entonces se extrae la FFT de cada ventana y se calcula su magnitud de acuerdo a (eq. 2):

$$\hat{S}_p(f_k) = \frac{1}{N} |X'(f_k)|^2, \quad k = 0, 1, \ldots, N-1, \tag{2}$$

donde $X'(f_k)$ es la DFT de la señal $x'(n) = x(n)w(n)$, y $w(n)$ es una función ventana de tamaño N.

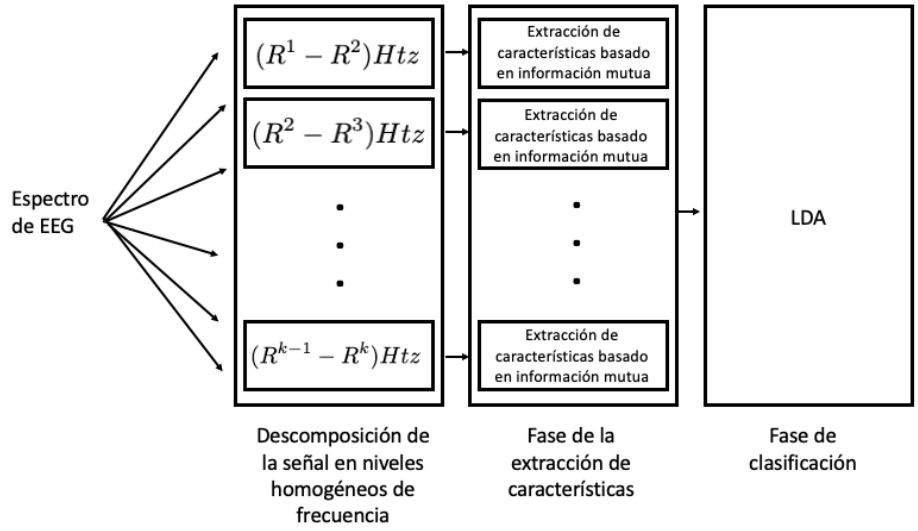


Fig. 3. Marco generalizado del FBCSP.

Cuando esta ventana es rectángular, el resultado es conocido como periodograma. A diferencia, cuando es no-rectángular, por ejemplo una ventana Hanning o Hamming, el resultado es un periodograma modificado. El efecto de una ventana no-rectángular es mitigar errores de estimación, como la ”fuga”, mejor conocido como leakage (cuando la potencia espectral de una banda de frecuencia se fuga a otras bandas, resultando en componentes incorrectos de frecuencia), los cuales son resultado de trabajar con señales finitas.

La mayor limitante de ambos métodos es que la PSD presenta alta variabilidad que no se reduce aún cuando su longitud aumenta. Para reducir el impacto de este inconveniente, se calcula el periodograma promediado, el cual resulta de promediar un número K de periodogramas de la señal estacionaria.

La idea del promediado se basa meramente en el hecho de que la varianza de la suma de un número K de señales independientes e idénticamente distribuidas a través de sus respectivas variables aleatorias, es $1/K$ veces la varianza de cada una de las variables aleatorias.

No obstante, en la práctica se tiene sólo un periodograma por señal adquirida, por lo que se utiliza el método de Welch-Barlett, el cual resulta de dividir el periodograma en segmentos independientes y luego promediarlos. Se sigue que el método para calcular el periodograma promediado está dado por (eq. 3):

$$\hat{S}(f_k) = \frac{1}{P} \sum_{p=0}^{P-1} \hat{S}_p(f_k), \quad k = 0, 1, \ldots, M-1, \tag{3}$$

donde M es la longitud de los segmentos P, y $\hat{S}_p(f_k)$ se calcula mediante (eq. 2). Este es justamente el tipo de método escogido para el análisis frecuencial presentado.

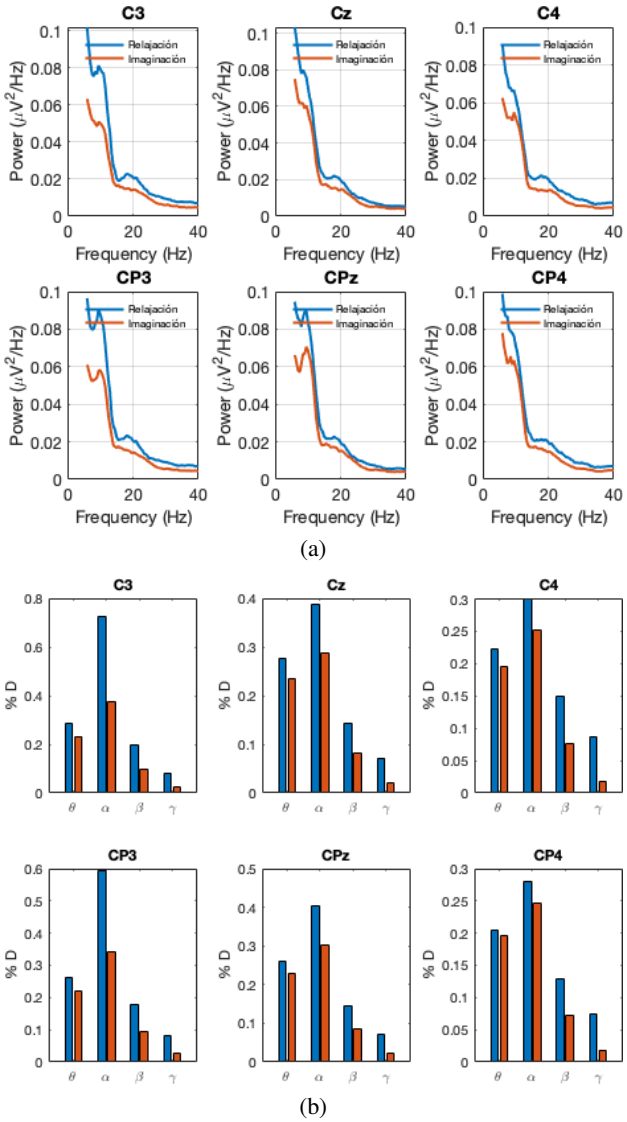


Fig. 4. (a) Desincronización espectral Promedio de participantes. (b) Desincronización espectral por bandas Promedio de Participantes.

Análisis de desincronización. Como se mencionó, el análisis de la Desincronización de la Potencia Espectral (D) es una técnica de EEG que permite estudiar la actividad cerebral asociada con la imaginación motora. Algunas de las ventajas de esta técnica incluyen su alta sensibilidad y especificidad para detectar cambios en la actividad cerebral relacionados con la imaginación motora, su capacidad para identificar patrones de actividad específicos para diferentes tipos de movimiento, y su utilidad en la rehabilitación de pacientes con discapacidades motoras [2].

Para obtener la desincronización se segmentaron las señales en ventanas de tiempo de 4 segundos en la etapa de entrenamiento, de forma que la ventana de relajación comprende los últimos 4 segundos del bloque de la cruz de fijación y la ventana de imaginación comprende 1 segundo después de haber comenzado el entrenamiento y hasta 4 segundos después.

Se obtuvo la PSD de estos tensores, cuyas medidas son $8 \times 1024 \times 24$ (canales, muestras, tareas), mediante FieldTrip (Donders Centre for Cognitive Neuroimaging) que segmenta las señales y las transforma al dominio frecuencial con la PSD, cuyo dominio va de 6 a 40 Hz, con una resolución frecuencial de 2 Hz/Hz.

A continuación, un promedio a lo largo de cada una de las 24 tareas fue realizado para reducir la estocasticidad de las señales y poder analizar mejor la disminución de la potencia espectral durante la ejecución de la imaginación en contraste con el periodo de relajación. De forma análoga, un promedio a lo largo de los 10 participantes fue hecho para concentrar la información.

Finalmente, se obtuvieron los porcentajes de desincronización en cada electrodo con (eq. 4) y por cada banda. Para obtener las bandas de frecuencia de donde esta disminución en la potencia espectral es más notable, se realiza una segmentación de la señal en ventanas de frecuencia:

$$\mathrm{D}_{\%} = \frac{\mathrm{PSD}_{\text{event}} - \mathrm{PSD}_{\text{baseline}}}{\mathrm{PSD}_{\text{baseline}}} \times 100. \tag{4}$$

Correlación entre relajación e imaginación. La correlación en el espectro de frecuencia de señales de EEG es un aspecto fundamental para el diseño de BCI especialmente en el caso de paradigmas de imaginación motora utilizados para el control de dispositivos robóticos [1].

Al utilizar la imaginación motora, se activan las mismas redes neuronales que se utilizan para la planificación y ejecución de movimientos físicos, lo que se traduce en cambios en la actividad cerebral detectados por la EEG. El estudio de la correlación en el espectro de frecuencia de estas señales puede proporcionar información valiosa sobre los patrones de actividad neuronal asociados con la imaginación de movimientos específicos.

Para su estudio, se utilizaron las librerías de análisis de correlación en el dominio frecuencial de FieldTrip. Esta librería calcula el coeficiente de correlación usado para examinar diferencias significativas en las características basadas en la potencia espectral de distintos eventos, y con ello seleccionar la potencia espectral de los canales con la tasa más alta de discriminación de clases. El coeficiente de correlación fue calculado independientemente para cada electrodo y frecuencia (eq. 5):

$$r_j = \frac{\sum_{i=1}^{N}(x_{j,i} - \bar{x_j})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_{j,i} - \bar{x_j})^2 \sum_{i=1}^{N}(y_i - \bar{y})^2}}, \tag{5}$$

donde $\mathrm{x_{j,i}}$ es la i-ésima muestra de la j-ésima característica, y_i es la clase de la etiqueta asociada a la muesta i-ésima y la notación de barra representa el promedio a lo largo de todos los participantes.

Tabla 1. Desincronización Porcentual por Canal por Banda.

Banda	C3	C1	Cz	C2	C4	CP3	CPz	CP4
Theta (θ)	51.65	42.67	36.65	39.88	39.29	52.53	43.72	31.08
Alpha (α)	58.68	46.22	25.10	22.49	20.79	52.08	25.09	20.30
Beta (β)	43.83	36.65	32.90	34.49	41.68	44.43	31.10	36.75
Gamma (γ)	57.18	32.33	31.68	33.93	55.02	59.50	41.33	47.28
Promedio	52.83	39.47	31.59	32.70	39.19	52.14	35.31	33.85

Las características seleccionadas fueron los valores de potencia espectral en α:[8, 12] Hz y θ:[4 8] Hz, las cuales son las bandas de imaginación motora que presentan mayor r-cuadrada. En el estudio, esta selección de características fue realizada individualmente para cada electrodo. Entonces el vector de características es $x \in \mathbb{R}^{n\times 1}$ con una etiqueta de clase $y \in [\text{Relajación}, \text{Imaginación}]$ para imaginación motora, donde n varía de acuerdo al número de electrodos usados.

2.3. Estudios con datos de validación

Extracción de características El algoritmo Filter Bank Common Spatial Pattern (FBCSP) es una técnica de extracción de características que se utiliza comúnmente en el análisis de señales de electroencefalograma (EEG) para el paradigma de imaginación motora. Este algoritmo es importante en la imaginación motora debido a que los componentes frecuenciales de las señales pueden variar entre los sujetos.

Por ejemplo, alguna frecuencia particular de los ritmos sensorimotores no es la misma para todos los usuarios. El FBCSP utiliza una técnica de filtrado de banco de filtros para separar la señal EEG en varias bandas de frecuencia. Luego, se aplica la técnica Common Spatial Pattern (CSP) a cada una de las bandas de frecuencia para extraer características discriminativas de la señal [1].

Una de las principales razones para elegir FBCSP es maximizar la varianza relativa entre el par de valores en los Estados de Información Mutua. Otro criterio para la selección de FBCSP es su capacidad para discriminar el espectro de la señal EEG específico del sujeto. En varias demostraciones FBCSP ha demostrado una notable mejora de rendimiento en comparación con los modelos más avanzados del estado del arte.

Estructuralmente, el algoritmo FBCSP (Fig. 3) consta de 4 pasos para seleccionar características espaciotemporales del espectro del EEG. En la fase 1, el canal de EEG se descompone en trozos de señal equidistantes mediante el filtro Chebyshev Tipo II. En la segunda fase los trozos de señal descompuestos se transforman linealmente en el vector de características (ecuación 6):

$$X = [cf_1, cf_2, ..., cf_k], \tag{6}$$

donde $cf_i \in \mathbb{R}^{2m}$ representa m pares de características CSP para mediciones de EEG filtradas por un pasa bandas. En la tercera fase, las características son seleccionadas basándose en Mutual Information of Best Individual Feature (MIBF).

Este algorítmo ordena las primeras k características en orden decreciente considerando información mútua de las características. Finalmente, se han clasificado las observaciones fusionando características débiles basadas en aprendizaje, en una única característica fuerte de forma iterativa [6].

Clasificador. Después de la extracción de características con FBCSP y CSP, se utiliza un clasificador para determinar la clase a la que pertenece la señal EEG. Uno de los clasificadores más utilizados para este propósito es el análisis discriminante lineal (LDA), que es un clasificador supervisado que busca una combinación lineal de características que maximice la separación entre las clases.

El LDA utiliza la información obtenida de la matriz de covarianza de cada clase y la matriz de covarianza combinada de todas las clases para calcular la combinación lineal óptima. Luego, se utiliza esta combinación lineal para clasificar la señal EEG en una de las dos clases [9]. Si x representa un vector real de un número n de características para una época de EEG, el modelo de clasificación evalua la función (eq.7):

$$f(x) = g\left(\sum_{i=1}^{n} b_i x_i + d\right), \tag{7}$$

donde b y d son los coeficientes del modelo lineal y $g(a)$ es una función escalar. Entonces el modelo de clasificación devuelve la etiqueta $l \in [1, -1]$ para la evaluación observada basada en la evaluación de $f(x)$. Un enfoque típico es usar algún umbral dado que los valores encima de ellos tienen la etiqueta $l = 1$ y valores por debajo $l = -1$.

LDA encuentra la clase de la etiqueta l que maximiza la condición probabilística $p(L = l|X = x)$. Asume que las funciones de densidad probabilísticas $p(X = x|L = -1)$ y $p(X = x|L = 1)$ tienen una distribución normal m_{-1}, m_1 y matrices de covarianza C_{-1}, C_1. Bajo estas suposiciones, la regla de decisión $p(L = 1|X = x) > p(L = -1|X = x)$ es expresada como un producto punto $b'x + d > 0$, donde:

$$b = 2C^{-1}(m_1 - m_{-1}), \tag{8}$$

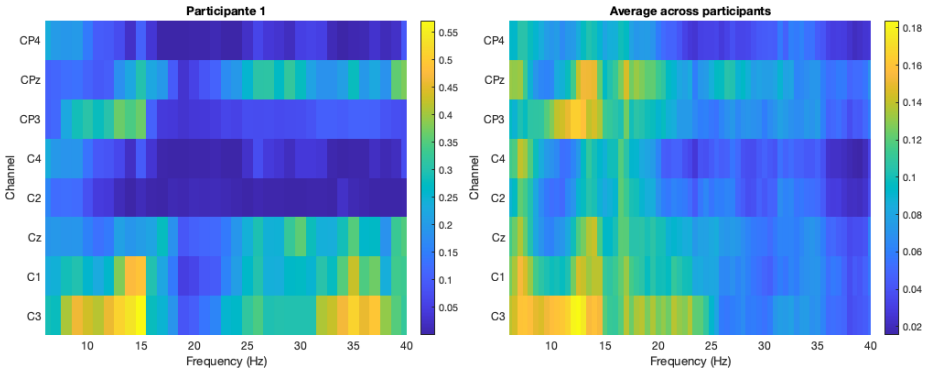
$$d = ln\left(\frac{P(L = -1)}{P(L = 1)}\right) + m'_{-1}C_{-1}^{-1}m_{-1} - m'_1C_1^{-1}m_1, \tag{9}$$

y $P(L = l)$ es la probabilidad de la clase etiquetada como l. El objetivo de estos dos modelos de clasificación es la discriminación de las clases para obtener una etiqueta $(l = 1)$ cuando se detecta la imaginación y $(l = -1)$ cuando no se detecta.

Metricas de rendimiento. Para evaluar el rendimiento del clasificador en la tarea de detección de imaginación motora se usaron las siguientes métricas:

– Precisión de la clase de detección de imaginación motora (C1): rendimiento en la clasificacion de la imaginación de apertura y cierre de la mano derecha.

– Precisión de la clase de no detección de imaginación motora (C2): rendimiento en la clasifición de la falta de imaginación de apertura y cierre de la mano derecha.



(a) (b)

Fig. 5. (a) Correlación en el dominio de la frecuencia entre ambos eventos Participante 1. (b) Correlación en el dominio de la frecuencia entre ambos eventos Promedio de participantes.

– Promedio de ambas clases (Ctotal): promedio de la clase de detección de imaginación motora (C1) y la clase de falta de imaginación motora (C2).
– Puntaje de clasificación correcta (Score): cantidad de veces que se detectó correctamente la imaginación de apertura y cierre de la mano derecha. Este puntaje varía de 0 (ninguna detección) a 24 (detección en todas las tareas).
– Tiempo de detección (TD): Tiempo transcurrido entre el inicio del test time y la detección de la imaginación motora. No hay un tiempo de detección determinado para participantes en cuyas tareas se haya terminado el tiempo disponible para realizar la imaginación y que no se haya clasificado.

Cabe mencionar que el tiempo de detección puede ser útil para evaluar la capacidad del participante para generar señales de EEG claras y distinguibles que se puedan utilizar para controlar una BCI. Además, el tiempo de clasificación de señales de EEG también puede estar relacionado con el rendimiento y el puntaje total del participante. Por ejemplo, si el tiempo de clasificación es largo, puede indicar que el participante tiene dificultades para generar señales de EEG claras y distinguibles, lo que puede afectar negativamente su rendimiento en la tarea [9].

3. Resultados

3.1. Datos de entrenamiento

En la figura (Fig. 4) se presentan los resultados de la densidad de potencia espectral y del análisis de desincronización a lo largo de un promedidado de los 10 participantes, con el fin de eliminar variabilidades producto de la estocasticidad de la señal. Estos resutados corresponden al promedio de la potencia espectal de las 24 pruebas en 6 de los 8 canales de EEG. Los canales C1 y C2 no se muestran para permitir mejor visualización de las gráficas. Finalmente, la banda delta no se muestra porque se le aplicó un filtro a la señal a partir de 4 Hz.

Tabla 2. Resultados del algoritmo de clasificación.

Participante	C1	C2	Ctotal	Score	Score (%)
1	0.97	0.96	0.97	21	87.50 %
2	0.97	0.95	0.96	14	58.33 %
3	0.86	0.89	0.87	21	87.50 %
4	0.87	0.86	0.87	11	45.83 %
5	0.91	0.94	0.93	10	41.67 %
6	0.90	0.86	0.88	24	100.00 %
7	0.87	0.94	0.91	11	45.83 %
8	0.87	0.93	0.90	6	25.00 %
9	0.86	0.93	0.89	21	87.50 %
10	0.90	0.99	0.94	23	95.83 %

Los cambios más relevantes aparecen en la figura 4(b), donde la desincronización porcentual es la métrica cuantitativa que permite concluir que son los canales C3 y CP3 los que exhiben mayor diferencia en este respecto. Es decir que la tarea de imaginación motora de la apertura y cierre de la mano derecha se encuentra con mayor prominencia en los canales C3 y CP3 sobre la corteza motora y en las banda alpha de frecuencia.

Haciendo uso de (eq.4), se obtiene en la tabla (1) la desincronización porcentual para todos los canales en cada una de las bandas estudiadas. Los valores se presentan en valor absoluto, debido a que estos son negativos por exhibir mayor potencia en la relajación (baseline) que en la imaginación. Nótese que este procedimiento reafirma que la desincronización de la PSD es más prominente en C3 y CP3 en la banda alpha.

Otro análisis que además tuvo lugar en los datos de entrenamiento fue la correlación entre la relajación e imaginación en el espectro de potencia aplicado a cada canal de forma independiente. Se muestran los resultados del participante 1 en la figura (Fig. 5)(a) y el promedio de todos ellos en la figura (Fig. 5)(b). Se discute además los resultados análogos entre ambos estudios, de desincronización espectral y correlación entre eventos, y cómo ambos llevan a conclusiones similares acerca de qué canales y qué bandas de frecuencia exhiben mayor importancia en la detección de imaginación.

La gráfica 5(a) muestra que existe una correlación significativa en el canal C3 de la banda alpha. Aunque esto es cierto, también muestra lo mismo en la banda gamma, lo cual no corresponde con la literatura de las BCI usadas con imaginación motora. En realidad no es sino hasta promediar los resultados a lo largo de todos los participantes que se disminuye la estocasticidad de los datos permitiendo reforzar las conclusiones hasta ahora mostradas y eliminando este fenómeno atípico de la banda gamma. Se muestra además en la figura 5(b) que otro canal representativo en la banda alpha es CP3, lo cual, corresponde con los resultados esperados.

3.2. Datos de validación

Se muestra en la tabla (2) los resultados de las métricas utilizadas para evaluar al rendimiento de los participantes al utilizar la BCI. Donde una precisión mayor al 75 % en las métricas de las clases indica que el clasificador es capaz de identificar correctamente la imaginación de apertura y cierre de la mano derecha por lo

Tabla 3. Resultados del tiempo de detección (tiempo en segundos).

Prueba	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Promedio	6.6	7.9	5.8	4.4	6.0	5.9	6.4	6.4	4.0	4.3
Desviación estándar	3.6	3.4	2.6	1.8	3.1	3.2	3.0	3.9	1.5	1.5
Máximo	13.3	14.2	12.0	8.5	11.7	13.2	11.6	13.8	9.4	8.3
Mínimo	3.2	3.4	3.4	3.2	3.4	3.2	3.2	3.4	3.2	3.2

menos tres cuartas partes de las veces, mucho mayor al 50 % que en teoría sería mínimamente necesario para discernir entre un modelo de toma de decisiones y la aleatoriedad cuando existen solo 2 opciones de clasificación. Sin embargo, 4 de los 10 participantes movieron la órtesis robótica un un número de veces tal que el score porcentual se muestra por debajo del umbral de la aleatoriedad del 50 % (participantes: 4, 5, 7, 8).

Estos mismos participantes no exhiben precisiones en las métricas de las clases necesariamente bajas (de hecho los participantes 5, 7 y 8 obtuvieron precisiones de clases promediadas mayores al 90 %). Por lo cual no se puede decir que existe una relación directa entre el rendimiento del clasificador y el score que finalmente el participante logra.

El segundo estudio realizado con los datos de validación en línea fue el tiempo de detección, el cual otorga información valiosa sobre qué tan difícil fue para el participante completar la tarea. La tabla (3) muestra un resumen de estos resultados, excluyendo aquellas pruebas donde no se clasificó la señal.

El promedio permite tener una comprensión más clara acerca de qué participante logró clasificar las señales de forma más rápida, siendo este el participante 9, quien con un puntaje del 87.5 % de aciertos, duró solo 4 segundos. Mientras que el participante que más tardó fue el 2, con un puntaje del 58.33 % de aciertos. De forma análoga, los extremos del número de aciertos muestran que el participante que obtuvo el 100 % duró 5.9 segundos, lo cual es justo la media del promedio de los tiempos.

Por su parte, el participante que obtuvo solo el 25 % de aciertos está tan solo un poco por encima de la media. Sin embargo, los resultados más concluyentes son aquellos que contrastan los extremos del tiempo de clasificación, puesto que se observa que, generalmente, los participantes que más tardan en clasificar la señal son también quienes menos aciertos tienden a hacer.

Finalmente, cabe destacar que los participantes 9 y 10 tienen una desviación estándar muy por debajo de la de los demás, al mismo tiempo que son justamente estos participantes quienes mayor número de aciertos obtuvieron.

4. Conclusiones

Los resultados obtenidos en este estudio indicaron una desincronización significativa en la potencia espectral de los canales C3 y CP3 en las bandas alpha y theta durante la tarea de imaginación motora de apertura y cierre de la mano derecha en comparación con la relajación. Esta desincronización en la potencia espectral sugiere una activación de la corteza motora, específicamente en la región sobre la que están posicionados los electrodos utilizados en este estudio.

Estos hallazgos son consistentes con investigaciones previas que han demostrado que la activación de la corteza motora durante la imaginación se puede detectar mediante el análisis de las señales de EEG en frecuencia.

Por otro lado, este estudio utilizó diferentes técnicas de evaluación de la señal de EEG de imaginación motora para los datos de entrenamiento, donde se calibró un modelo de inteligencia artificial para extraer las características de la señal y se realizó un análisis frecuencial para detectar los canales y bandas de potencia más representativos, y para los datos de validación en línea, donde se planteó hacer una correlación entre las métricas del clasificador. Sin embargo, en lo que respecta a dicha correlación, se concluye que no existe una significancia estadística lo suficientemente alta como para correlacionar directamente estas métricas entre sí.

En conjunto, estos hallazgos demuestran que el análisis de señales de EEG en el dominio de la frecuencia puede ser una herramienta útil para el estudio de la actividad cerebral durante la imaginación motora, y puede ser utilizado para la calibración de modelos de inteligencia artificial para el control de dispositivos robóticos.

Referencias

1. Antelis, J. M., Gudiño-Mendoza, B., Falcón, L. E., Sanchez-Ante, G., Sossa, H.: Dendrite morphological neural networks for motor task recognition from electroencephalographic signal. Biomedical Signal Processing and Control, vol. 44, pp. 12–24 (2018) doi: 10.1016/j.bspc.2018.03.010
2. Antelis, J. M., Montesano, L., Ramos-Murguialday, A., Birbaumer, N., Minguez, J.: On the usage of linear regression models to reconstruct limb kinematics from low frequency EEG signals. Public Library of Science ONE, vol. 8, no. 4, p. e61976 (2013) doi: 10.1371/journal.pone.0061976
3. Gilete-Tejero, I. J., Ippolito-Bastidas, H. Z., Bernal-García, L. M., Mata-Gómez, J., García-Moreno, R., Ortega-Martínez, M., Cabezudo-Artero, J. M.: Efecto de la edad en el pronóstico de pacientes con traumatismo craneoencefálico sometidos a craneotomía: análisis de una serie quirúrgica. Revista de Neurología, vol. 66, no. 4, pp. 113–120 (2018) doi: 10.33588/rn.6604.2017411
4. Hernandez-Rojas, L. G., Cantillo-Negrete, J., Mendoza-Montoya, O., Carino-Escobar, R. I., Leyva-Martinez, I., Aguirre-Guemez, A. V., Barrera-Ortiz, A., Carrillo-Mora, P., Antelis, J. M.: Brain-computer interface controlled functional electrical stimulation: Evaluation with healthy subjects and spinal cord injury patients. IEEE Access, vol. 10, pp. 46834–46852 (2022) doi: 10.1109/access.2022.3170906
5. Johnson, C. O., Nguyen, M., Roth, G. A., Nichols, E., Alam, T., Abate, D., Abd-Allah, F., Abdelalim, A., Abraha, H. N., Abu-Rmeileh, N. M., Adebayo, O. M., Adeoye, A. M., Agarwal, G., Agrawal, S., Aichour, A. N., Aichour, I., Aichour, M. T. E., Alahdab, F., Ali, R., Alvis-Guzman, N., et al: Global, regional, and national burden of stroke, 1990–2016: A systematic analysis for the Global Burden of Disease Study 2016. The Lancet Neurology, vol. 18, no. 5, pp. 439–458 (2019) doi: 10.1016/s1474-4422(19)30034-1
6. Kai-Keng, A., Zhang-Yang, C., Haihong, Z., Cuntai, G.: Filter bank common spatial pattern (FBCSP) in brain-computer interface. In: IEEE International Joint Conference on Neural Networks (2008) doi: 10.1109/ijcnn.2008.4634130
7. Klein, T. J., Lewis, M. A.: A physical model of sensorimotor interactions during locomotion. Journal of Neural Engineering, vol. 9, no. 4, pp. 046011 (2012) doi: 10.1088/1741-2560/9/4/046011

8. Ojeda, L. D., Pabon, J. J., Antelis, J. M.: Classification of hand movements in motor execution and motor imagery tasks from EEG signals recorded with a low-cost recording system. In: Memorias del Congreso Nacional de Ingeniería Biomédica, vol. 1, no. 1, pp. 187–190 (2014)
9. Triana-Guzman, N., Orjuela-Cañon, A. D., Jutinico, A. L., Mendoza-Montoya, O., Antelis, J. M.: Decoding EEG rhythms offline and online during motor imagery for standing and sitting based on a brain-computer interface. Frontiers in Neuroinformatics, vol. 16 (2022) doi: 10.3389/fninf.2022.961089

ISSN 1870-4069

Clasificación de rostros con aprendizaje hebbiano para bases de datos pequeñas

Fernando Aguilar-Canto, Alberto Espinosa-Juárez,
Juan Eduardo Luján-García, Hiram Calvo

Instituto Politécnico Nacional,
Centro de Investigación en Computación,
México

{faguilarc2021, aespinosaj2021, jlujang2020,
hcalvo}@cic.ipn.mx

Resumen. El aprendizaje hebbiano es un paradigma biológicamente plausible de optimización paramétrica en Redes Neuronales Artificiales. En años recientes, varias reglas hebbianas han sido incluidas en arquitecturas de Aprendizaje Profundo. No obstante, estas implementaciones están principalmente enfocadas en demostrar el potencial del aprendizaje hebbiano, y poca investigación se ha desarrollado en problemas específicos, como la clasificación de rostros. En este artículo, los autores presentan una aplicación del aprendizaje hebbiano en clasificación de rostros utilizando tanto una base de datos pública como privada. Asimismo, las reglas de aprendizaje hebbiano implementadas fueron la regla de Hebb Simple, la regla de Oja y la regla BCM. En general, el algoritmo de kNN mostró mejores resultados comparativos para esta tarea, mientras que las reglas de aprendizaje hebbiano fueron particularmente sensibles al desbalace de los datos. Para mitigar el problema, los autores introdujeron reglas de Hebb escaladas, las cuales lograron mejores resultados en comparación de las versiones no escaladas, teniendo un desempeño cercano al de optimizadores basados en el gradiente como el algoritmo de Adam.

Palabras clave: Aprendizaje hebbiano, clasificación de rostros, redes neuronales artificiales.

Face Classification with Hebbian Learning for Small Datasets

Abstract. Hebbian learning is a biologically plausible paradigm for parametric optimization in Artificial Neural Networks. In recent years, several Hebbian-based rules have been integrated into Deep Learning architectures. However, most of these implementations have primarily focused on demonstrating the potential of Hebbian Learning, with limited research conducted on specific problems, such as face classification. This paper presents an application of Hebbian Learning in face classification, utilizing both a public and a private dataset. The authors implemented three Hebbian learning rules: the Basic Hebb rule, the Oja rule, and the BCM rule. Although the kNN algorithm

yielded better comparative results for this task, the Hebbian-based rules exhibited sensitivity to imbalanced data. To address this issue, the authors introduced Scaled-Hebbian learning rules, which achieved improved results compared to the non-scaled versions. These rules performed at a similar level to gradient-based optimizers, such as the Adam algorithm.

Keywords: Hebbian learning, face classification, artificial neural networks.

1. Introducción

El aprendizaje hebbiano es un mecanismo de aprendizaje biológicamente plausible a nivel neuronal que modela los fenómenos de plasticidad sináptica (cambios en las fuerzas de conexión de las neuronas[1]) conocidos como Potenciación a Largo Plazo (Long-Term Potentiation, LTP) y Depresión a Largo Plazo (Long-Term Depression, LTD) [10, 29]. La plasticidad sináptica, en forma de LTP y LTD, proporciona las bases de numerosos modelos de aprendizaje y memoria, así como el desarrollo de los mapas corticales en el cerebro [1].

En el contexto de computación, el aprendizaje hebbiano está conformada de una familia de reglas de aprendizaje inspiradas en la neuroplasticidad, en donde el cambio de los pesos sinápticos incrementa proporcionalmente a los valores de las actividades (medidas en tasa de disparo o potenciales de acción) de las neuronas presinápticas y postsináptica [5].

En su versión general, una regla de aprendizaje hebbiana está dada por el cambio de los pesos $\Delta_w = H(x, y, w, h)$, donde H es una función que modela dicho cambio, x es el valor de actividad de la neurona presinática, x la actividad de la neurona postsináptica, w es el peso entre ambas neuronas y h es el "tercer factor" mediado por neuromoduladores [21]. El aprendizaje hebbiano ha sido empleado en el contexto del aprendizaje profundo (Deep Learning), en particular con el uso de redes neuronales convolucionales (Convolutional Neural Networks, ConvNets) [5, 22, 3, 27].

La clasificación de rostros es el proceso de clasificar caras en diferentes categorías [34], que pueden pertenecer a personas individuales (como este trabajo), género [17], etnicidad [12] u otras clases. En el contexto de aprendizaje de máquina, muchos clasificadores de rostros utilizan ConvNets, debido a que son robustos a transformaciones de imágenes [26], y como consecuencia, en general han sido efectivas en tareas relacionadas con imágenes [20].

Aunque la clasificación de rostros es un tópico principal en aprendizaje profundo, se ha implementado escasamente con aprendizaje hebbiano (véanse los Trabajos relacionados en la Sección 2). En este trabajo, se hace una implementación de clasificación de rostros utilizando aprendizaje hebbiano con modelos pre-entrenados. La intención de este trabajo es doble: por un lado expandir los límites del aprendizaje hebbiano profundo (tradicionalmente aplicado en bases de datos como MNIST o CIFAR) y evaluar el desempeño del mismo en contextos diferentes.

[1] Es equivalente al ajuste paramétrico de pesos en redes artificiales

De igual forma, nos hemos centrado en bases de datos pequeñas, debido a que la disponibilidad de datos para ciertos fines comerciales muchas veces se restringe a este tipo de problemáticas. Un ejemplo práctico donde este problema aparece, es en la clasificación de rostros de personas pertenecientes a un grupo pequeño, por ejemplo, los estudiantes de un aula o miembros de una junta directiva, con el fin de registrar asistencia de forma automática.

Adicionalmente, es ampliamente conveniente que el sistema aprenda en tiempo real (online) porque eso puede permitir que los usuarios ingresen únicamente las caras y las etiquetas (por voz) de las personas a clasificar, sin tener que utilizar un proceso de entrenamiento que puede demandar supervisión experta. Dado que las reglas hebbianas realizan aprendizaje en tiempo real [3], son candidatas ideales para este tipo de tareas.

Este trabajo se estructura de la siguiente manera: en la sección 2 presentamos los trabajos relacionados con el nuestro; en la sección 3 presentamos a la arquitectura para su evaluación; en la sección 4 mostramos los resultados y una comparación con otras técnicas, así como algunas observaciones relevantes; en la sección 5 recabamos una discusión sobre las observaciones alcanzadas; y finalmente las conclusiones se presentan en la sección 6.

2. Trabajos relacionados

El aprendizaje hebbiano profundo puede definirse como el conjunto de técnicas que utilizan tanto redes neuronales profundas como arquitectura, así como algunas reglas de aprendizaje hebbiano, ya sea en toda la arquictectura o parcialmente.

En el caso particular de la Clasificación de rostros, que es el problema presentado en este artículo, se han utilizado diferentes técnicas con redes neuronales, destacando las ConvNets [8, 12, 17, 33] y el hardware neuromórfico [34].

A pesar de la existencia de implementaciones en cómputo neuromórfico (donde se suele utilizar la regla de Hebb) para el problema de clasificación de rostros, existen pocos trabajos sobre aprendizaje hebbiano en este tópico y un caso similar se observa en el similar problema de reconocimiento de rostros (face recognition). Para el caso de clasificación de ciertas categorías de rostros, en [16] se utiliza un método de detección de coincidencia con redes pulsantes y la regla de Plasticidad Dependiente del Tiempo del Impulso (Spike-Timing Dependent Plasticity) [28].

En cuanto a la detección de rostros, se tienen más trabajos relacionados. En [24] se aplica a la Regla de Oja para implementar una versión del modelo jerárquico de la visión extendido a la Corteza Infratemporal [31], permitiendo reconocer si dos caras son idénticas con invariancia a rotaciones.

Previo al surgimiento del aprendizaje profundo, se exploraron ideas para la clasificación de rostros como en [11], con el uso del Algoritmo Hebbiano Generalizado. Por otro lado, en [6], se implementa el Aprendizaje Hebbiano Competitivo y una red atractiva para obtener una clasificación de rostros robusta e invariante a la pose.

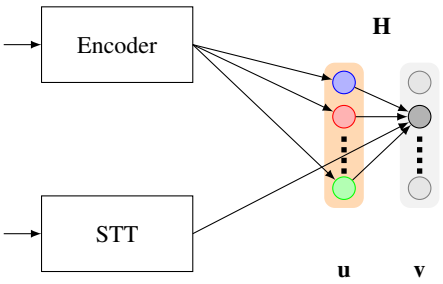


Fig. 1. Representación del esquema de [3] donde el vector **u** representa la salida del codificador visual (generalmente una red convolucional o transformador visual), **v** representa a la codificación one-hot de las etiquetas, **H** representa la matriz de pesos que relaciona ambos vectores de forma lineal.

3. Metodología

El objetivo principal que se persigue para este artículo es el aprendizaje en tiempo real para la clasificación de imágenes de rostros de personas individuales. Esto puede ser auxiliado con el uso de transferencia de aprendizaje con otras arquitecturas.

La metodología que se utiliza para este artículo consiste en utilizar el esquema presentado en el artículo de [3], representada en la figura 1, en la cual se dispone de un codificador visual que devuelve un vector de características **u**. Normalmente, este codificador es una red neuronal convolucional preentrenada o un transformador visual.

El vector **v** representa a la codificación one-hot de las etiquetas (que podrían ser introducidas por voz). Asimismo, el vector **v** puede ser aproximado mediante la operación **Hu**, de forma que la matriz de pesos **H** es una aproximación lineal del vector de características al vector de etiquetas. Esta matriz de pesos **H** se entrena utilizando aprendizaje hebbiano.

3.1. Reglas de Hebb

Existen varias formulaciones de la llamada "regla de Hebb", que conforman modelos de plasticidad sináptica. En este artículo consideraremos tres principales: la regla de Hebb Simple, la Regla de Oja (empleada en [24]), y la Regla BCM. Estas reglas se expresan en términos de ecuaciones diferenciales, pero se pueden discretizar utilizando el Método de Euler. Por lo general, estas reglas se pueden ejecutar en tiempo real, ya que reciben un dato a la vez y actualizan los parámetros de la red (véase la figura 2).

Regla de Hebb Simple. Consideremos un vector de tasa de disparo presináptico **u** y la neurona postsináptica v, con el vector de pesos **w**. La formulación más simple de la regla de Hebb está dada por la ecuación diferencial (Eq. 1) mostrada en [14]:

$$\tau_w \frac{d\mathbf{w}}{dt} = v\mathbf{u}, \tag{1}$$

donde $\tau_w > 0$ es una constante que controla la razón de cambio de los pesos. Sin embargo, esta regla de aprendizaje presenta el problema del crecimiento no acotado de los pesos sinápticos.

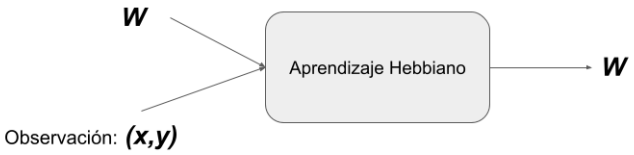


Fig. 2. Esquema de ajuste de parámetros con aprendizaje hebbiano. Un dato etiquetado (x, y) es recibido y la red ajusta sus pesos W. En este sentido, no es necesario enviar un lote de entrenamiento o ver todo el conjunto de datos, sino que con un solo dato se realizan los ajustes.

Regla de Oja. La regla de Oja [30] es una solución al problema del crecimiento no acotado de los pesos al introducir normalización. La regla de Oja (Eq. 4) se puede describir con la ecuación diferencial. En este caso se fijó $\beta = 0.01$:

$$\tau_w \frac{d\mathbf{w}}{dt} = v\mathbf{u} - \beta v^2 \mathbf{w}. \tag{2}$$

Regla BCM. Otra solución al problema del crecimiento no acotado de los pesos está dada por la regla Bienenstock-Cooper-Munro (BCM), la cual introduce un umbral dinámico para regular los pesos y considera tanto a los fenómenos de LTP y LTD [9]. Esta regla de aprendizaje ha recibido evidencia empírica [23, 13] y está dada por la Eq. 5 y Eq. 6:

$$\tau_w \frac{d\mathbf{w}}{dt} = v\mathbf{u}(v - \theta_v), \tag{3}$$

$$\tau_\theta \frac{d\theta_v}{dt} = v^2 - \theta_v, \tag{4}$$

donde $\tau_\theta > 0$ es la constante que controla la razón de cambio del umbral. Debido a que $\theta_v \neq \tau_w$, la discretización de θ_τ se expresa como $\Delta\theta_v = \gamma(v^2 - \theta_v)$. En este caso se fijó $\gamma = 0.5$.

Reglas de Hebb Escaladas. Utilizando el método de Euler, las reglas de Hebb se pueden discretizar de la forma siguiente:

$$\Delta\mathbf{w} = \alpha H(v, \mathbf{u}), \tag{5}$$

donde α es la tasa de aprendizaje y $H(v, \mathbf{u})$ es el modelo utilizado, por ejemplo $H(v, \mathbf{u}) = v\mathbf{u}$ para la regla de Hebb Simple. Una forma de combatir el desbalance en los datos consiste en actualizar dividiendo entre el número de ocurrencias de la clase (Eq. 3), es decir:

$$\Delta\mathbf{w} = \frac{1}{|\{v(t) = 1\}|} H(v, \mathbf{u}). \tag{6}$$

Normalizar a la regla de Hebb ha sido implementada en otros artículos como [25, 4], pero generalmente se normaliza sobre la suma de los pesos $\sum_i w_i$, y no sobre la suma de las ocurrencias de $v(t) = 1$. Para nuestro conocimiento, no se ha implementado este tipo de normalización / escalamiento.

Fig. 3. Ejemplo de muestras de 5 Celebrity Faces Dataset.

3.2. Codificador visual

Un codificador visual apropiado para esta tarea está dado por el módulo de `face recognition` de Python [15], el cual está basado en el módulo Dlib-ml [18]. Este codificador realiza tanto reconocimiento de rostros (devolviendo las coordenadas de ubicación del rostro) como el vector de características $\mathbf{u} \in \mathbb{R}^{128}$, las cuales pueden ser utilizadas para su clasificación.

El detector de rostros utilizado por el módulo Dlib-ml es MMOD CNN que utiliza una ConvNet entrenada con el método de detección de objetos de máximo margen (MMOD) para detectar rostros en la imagen. Este algoritmo es capaz de detectar rostros desde diferentes ángulos de visión, condiciones de iluminación y oclusión.

3.3. Algoritmos de comparación

Para comparar los resultados de las reglas online, se utilizaron el algoritmo de k-vecinos más cercanos (k-nearest neighbors, kNN) con $k = 3$ y una red neuronal simple (sin capas intermedias) con 100 épocas, optimizador Adam [19] sobre la función de costo de entropía cruzada categórica.

Fig. 4. Ejemplo de muestras de base de datos GPIC.

3.4. Bases de datos empleadas

Se emplearon dos bases de datos:

1. **5 Celebrity Faces Dataset**: Este conjunto de datos [7] fue creado con la finalidad de probar algoritmos en tareas de visión por computadora, cuenta con cinco clases pertenecientes a:
   a) Ben Afflek,
   b) Elton John,
   c) Jerry Seinfeld,
   d) Madonna,
   e) Mindy Kaling.

   Cada clase tiene entre 14 y 20 imágenes, sin embargo, 7 imágenes de esta base de datos fueron eliminadas porque no fueron identificadas como rostros por el codificador visual, debido a la presencia de dos rostros. La base de datos se restableció utilizando 5 imágenes de cada clase como parte del conjunto de prueba. Un ejemplo de cada clase puede ser observada en la figura 3.
2. **Base de datos GPIC**: Este conjunto de datos fue creado tomando como base a 36 trabajadores de la empresa de construcción mexicana GPIC. Se conforma de dos fotografías tomadas por el dispositivo móvil de cada persona que pertenece al conjunto de datos en posición frontal tipo selfie.

Tabla 1. Resultados de exactitud sobre el conjunto de prueba con la Regla de Hebb Simple, Regla de Oja, Regla BCM, Regla del Perceptrón (PLR), en comparación con el algoritmo de kNN y una red neuronal con optimizador Adam. $|C_i|$ representa la cardinalidad de la clase i utilizada para entrenar.

$\|C_0\|$	$\|C_1\|$	$\|C_2\|$	$\|C_3\|$	$\|C_4\|$	**Simple**	**Oja**	**BCM**	**PLR**	**Adam**	**kNN**
1	1	1	1	1	0.8	0.8	0.4	0.8	0.96	0.32
2	2	2	2	2	0.92	0.92	0.6	0.92	1	1
4	4	4	4	4	0.92	0.92	0.6	0.92	1	1
5	5	5	5	5	0.96	0.96	0.64	0.96	1	1
7	7	7	7	7	0.96	0.96	0.68	0.96	1	1
8	8	8	8	8	0.92	0.92	0.68	0.96	1	1
9	9	9	9	9	0.92	0.92	0.68	0.92	1	1
12	12	12	12	12	0.92	0.92	0.68	0.84	1	1
12	13	13	13	13	0.92	0.92	0.68	0.64	1	1
12	14	14	14	14	0.72	0.72	0.68	0.6	1	1
12	15	15	15	15	0.76	0.76	0.68	0.6	1	1
12	16	15	16	16	0.6	0.6	0.68	0.6	1	1
12	17	15	17	17	0.6	0.6	0.68	0.6	1	1
12	18	15	18	18	0.6	0.6	0.68	0.6	1	1
12	19	15	19	18	0.6	0.6	0.68	0.6	1	1
12	20	15	20	18	0.52	0.56	0.68	0.6	1	1
12	20	15	21	18	0.4	0.56	0.68	0.48	0.96	1

Las imágenes no fueron controladas, por lo que la variación de luz, exposición, tamaño e incluso posición del rostro son más naturales y por ende, agrega dificultad al problema de clasificación. La razón por la cual se eligieron estas personas fue por la disposición que tuvieron para formar parte de este proyecto. Se solicitó la autorización por escrito para el uso de las imágenes con fines académicos y de investigación. La figura 4 muestra un ejemplo de algunas clases.

4. Resultados

4.1. Base de datos 5 celebrity faces dataset

Para el caso de la primera base de datos, se controló el número de datos para el conjunto de entrenamiento, dejando fijo el de prueba, con el objetivo de medir la influencia del número de datos en la decisión final para reglas de clasificación hebbianas. Además de las reglas BCM, Oja y Simple, se utilizó la regla del perceptrón [32] (Perceptron Learning Rule, PLR, $\alpha = 0.01$).

La tabla 1 presenta los resultados en esta base de datos con las reglas sin escalar y en la tabla 2 se muestran los resultados con las reglas escaladas. En la figura 5 se muestra la evolución de la exactitud con respecto al número de datos máximo por clase controlados. Tal como se especificó previamente, también se utilizaron el algoritmo de aprendizaje de Adam en una red neuronal y el algoritmo de kNN.

Tabla 2. Resultados de exactitud sobre el conjunto de prueba con la Regla de Hebb Simple, Regla de Oja, Regla BCM, Regla del Perceptrón (PLR), todas las reglas en su forma escalada. $|C_i|$ representa la cardinalidad de la clase i utilizada para entrenar.

$\|C_0\|$	$\|C_1\|$	$\|C_2\|$	$\|C_3\|$	$\|C_4\|$	**Simple**	**Oja**	**BCM**	**PLR**
12	13	13	13	13	0.92	0.92	0.68	0.92
12	14	14	14	14	0.92	0.92	0.68	0.92
12	15	15	15	15	0.96	0.92	0.68	0.96
12	16	15	16	16	0.96	0.92	0.8	0.96
12	17	15	17	17	0.92	0.92	0.8	0.92
12	18	15	18	18	0.92	0.92	0.76	0.92
12	19	15	19	18	0.92	0.92	0.4	0.92
12	20	15	20	18	0.92	0.92	0.4	0.92
12	20	15	21	18	0.92	0.92	0.4	0.92

4.2. Base de datos GPIC

Para el caso de la base de datos GPIC, también se realizó control sobre el número de datos para entrenamiento, dejando fijo al conjunto de prueba. Las comparaciones que se realizaron fueron las mismas de la base de datos anterior y se presentan en la tabla 3.

5. Discusión

Para ambas bases de datos, el algoritmo de kNN el mejor para efectuar la clasificación, siendo únicamente inefectivo para conjuntos de entrenamiento extremadamente pequeños (una maestra por clase). En cuanto a las reglas de aprendizaje en tiempo real (hebbianas y PLR), tanto la Regla de Hebb Simple, Oja y Perceptrón obtuvieron desempeños muy similares, mientras que la Regla BCM observó un comportamiento menos predecible, siendo en ocasiones mejor que otras formas de aprendizaje y en otras ocasiones peor.

En la base de datos 5 Celebrity Faces Dataset, el optimizador Adam se mostró comparativamente mejor con respecto a las reglas online, aunque en algunos casos la diferencia no fue tan notoria.

Un aspecto destacable sobre estas reglas es que la exactitud se reduce con el aumento de los datos, lo cual parece ser un fenómeno contradictorio. Sin embargo, en su mayor parte, este fenómeno puede explicarse con la introducción de desbalance de datos, donde la regla de Hebb se muestra ampliamente sensible.

En ambas bases de datos, las reglas escaladas mostraron mejores resultados. Tanto la regla BCM como la regla de Oja evitan el crecimiento no acotado mediante técnicas distintas, sin embargo, como se observa en los resultados, las formas escaladas terminan siendo más apropiadas.

No obstante, incluso las reglas escaladas y hasta el optimizador Adam se mostraron sensibles a la introducción de ciertos datos, al grado de reducir la exactitud. Este fenómeno es más notorio en la segunda base de datos, donde observamos que únicamente el algoritmo kNN logra resolver el problema, pero incluso una red entrenada con el optimizador Adam tiende a presentar este problema.

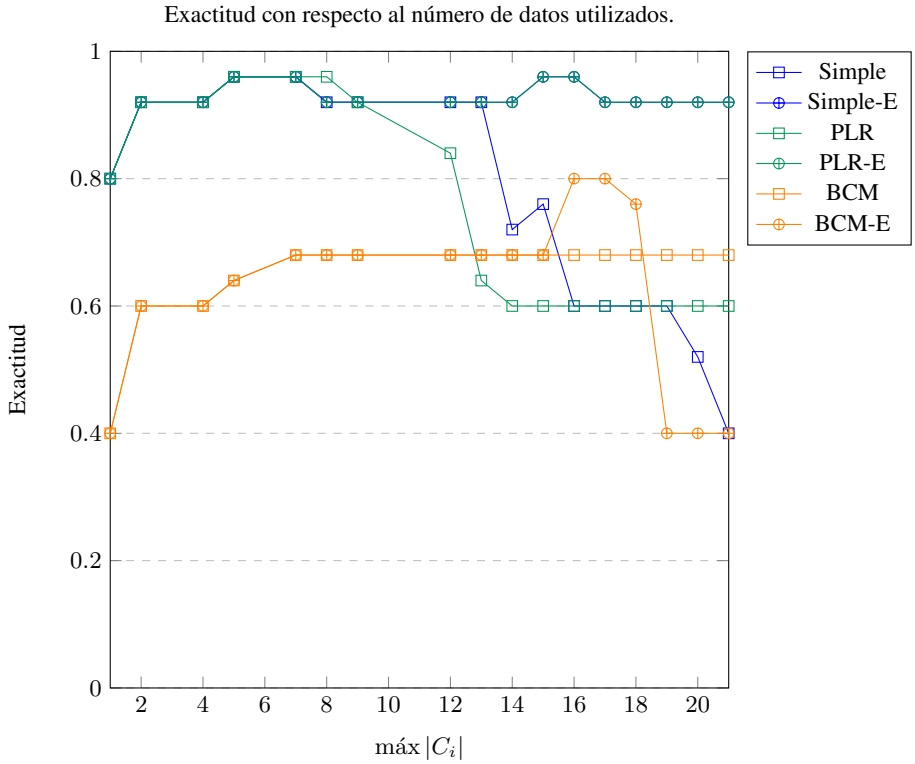


Fig. 5. Evolución de la exactitud para las Reglas de Hebb Simple sin escalar (Simple) y escalada (Simple-E); la Regla del Perceptrón (PLR) y su versión escalada (PLR-E), la Regla BCM y su versión escalada (BCM-E).

El hecho de que tanto la regla de Oja como BCM no logren obtener mejores resultados significa que el crecimiento no acotado de los pesos puede tener escaso o ningún papel. Esto parece entrar en conflicto con las conclusiones arrojadas en [2], pero al tratarse de un problema más generalizado puede tratarse de un ejemplo del principio trash in, trash out.

No obstante, el hecho de que ciertos clasificadores clásicos como kNN evadan este problema puede ser muestra de que no se trata del caso mencionado o el uso de las distancias más cortas permita al algoritmo de kNN ser robusto a este problema. En la base de datos GPIC, la Regla del Perceptrón mostró un desempeño considerablemente bajo en comparación con reglas hebbianas (de hasta un 30 %), y aunque Adam también presentó problemas similares, mostró mejores resultados.

Finalmente, es importante mencionar el hecho de que la Regla de Hebb Simple en su versión escalada tuviera los resultados más robustos en cuanto a las reglas online. A diferencia de las otras reglas de aprendizaje, la Regla de Hebb Simple no necesita hiperparámetros y se obtiene el mismo resultado si se utilizan diferentes tasas de aprendizaje positivas. El desempeño disimilar de reglas como BCM y PLR, o incluso Oja, parecen indicar una falta de ajuste de los hiperparámetros.

Tabla 3. Resultados de exactitud en cuatro experimentos sobre el conjunto de prueba con la Regla de Hebb Simple, Regla de Oja, Regla BCM, Regla del Perceptrón (PLR), así como sus versiones escaladas (-E), en comparación con el algoritmo de kNN y una red neuronal con optimizador Adam. $|C_i|$ representa la cardinalidad de la clase i utilizada para entrenar.

	Exp 1	**Exp 2**	**Exp 3**	**Exp 4**
$\|C_0\|$	1	2	2	2
$\|C_1\|$	1	2	3	3
$\|C_2\|$	1	2	2	2
$\|C_3\|$	1	2	2	2
$\|C_4\|$	1	2	2	2
$\|C_5\|$	1	2	2	2
$\|C_6\|$	1	2	3	5
$\|C_7\|$	1	2	2	2
$\|C_8\|$	1	2	2	2
$\|C_9\|$	1	2	2	2
$\|C_{10}\|$	1	2	2	2
$\|C_{11}\|$	1	2	2	2
$\|C_{12}\|$	1	2	2	2
$\|C_{13}\|$	1	2	3	3
$\|C_{14}\|$	1	2	3	5
$\|C_{15}\|$	1	2	2	2
$\|C_{16}\|$	1	2	3	5
$\|C_{17}\|$	1	2	2	2
$\|C_{18}\|$	1	2	2	2
$\|C_{19}\|$	1	2	2	2
$\|C_{20}\|$	1	2	2	2
$\|C_{21}\|$	1	2	3	3
$\|C_{22}\|$	1	2	3	3
$\|C_{23}\|$	1	2	3	3
$\|C_{24}\|$	1	2	3	3
$\|C_{25}\|$	1	2	2	2
$\|C_{26}\|$	1	2	2	2
$\|C_{27}\|$	1	2	2	2
$\|C_{28}\|$	1	2	2	2
$\|C_{29}\|$	1	2	2	2
$\|C_{30}\|$	1	2	2	2
$\|C_{31}\|$	1	2	2	2
$\|C_{32}\|$	1	2	3	3
$\|C_{33}\|$	1	2	2	2
$\|C_{34}\|$	1	2	2	2
$\|C_{35}\|$	1	2	3	5
Simple	0.75	0.83	0.31	0.11
Simple-E	0.75	0.83	0.39	0.11
Oja	0.75	0.83	0.31	0.11
Oja-E	0.75	0.83	0.44	0.11
BCM	0.64	0.72	0.28	0.11
BCM-E	0.64	0.72	0.47	0.47
PLR	0.75	0.53	0.25	0.11
PLR-E	0.75	0.53	0.36	0.11
Adam	0.56	0.92	0.64	0.22
kNN	0.31	1	1	1

6. Conclusiones

En este trabajo, se presentó el problema de clasificación de rostros para bases de datos pequeñas, con la restricción de que el aprendizaje se efectuara en tiempo real. Se utilizaron dos bases de datos pequeñas para tal fin, con 5 y 36 clases respectivamente, que corresponden a prototipos de situaciones reales de clasificación de rostros para entornos sociales reducidos como aulas, donde se tienen pocas muestras por persona.

De manera general, las reglas de aprendizaje hebbianas tuvieron un desempeño similar con respecto a la Regla del Perceptrón (basada en el gradiente y online) y en general, aceptable con respecto al optimizador Adam. Sin embargo, en general, el algoritmo de kNN parece más adecuado para esta tarea, exceptuando cuando se tiene solamente un dato por clase.

Las reglas implementadas en este artículo se mostraron ampliamente sensibles al desbalance de los datos, razón por la cual se propusieron las reglas escaladas, que tienen la desventaja de requerir dos iteraciones sobre los datos (para contar el número de elementos por clase) por lo que no serían completamente online.

7. Disponibilidad, privacidad y acceso a los datos

Los datos de la base de datos 5 Celebrity Faces Dataset es pública y puede consultarse en la referencia [7]. La base de datos GPIC es privada pero puede solicitarse directamente a los autores, siempre que se garantice respetar la privacidad de los participantes, los cuales firmaron de conformidad para que su imagen sea utilizada para publicación e investigación, análisis biométrico y desarrollo de software.

Disponibilidad de código

El código fuente puede ser consultado de forma libre en colab.research.google.com/drive/17AMrl5Pwe7K4dqSrj2w_aNV0ubMbJmSf?usp=sharing.

Agradecimientos. Agradecemos al Instituto Politécnico Nacional por su apoyo para la realización de este trabajo; de igual manera, al apoyo del gobierno mexicano a través del Consejo Nacional de Ciencia y Tecnología (CONACYT).

Referencias

1. Abbott, L. F., Nelson, S. B.: Synaptic plasticity: taming the beast. Nature neuroscience, vol. 3, no. 11, pp. 1178–1183 (2000) doi: 10.1038/81453
2. Aguilar-Canto, F., Calvo, H.: The role of the number of examples in convolutional neural networks with hebbian learning. In: Advances in Computational Intelligence: 21st Mexican International Conference on Artificial Intelligence, MICAI 2022, Monterrey, Mexico, October 24–29, 2022, Proceedings, Part I, pp. 225–238 (2022) doi: 10.1007/978-3-031-19493-1_19

3. Aguilar Canto, F. J.: Convolutional neural networks with hebbian-based rules in online transfer learning. In: Advances in Soft Computing: 19th Mexican International Conference on Artificial Intelligence, pp. 35–49 (2020) doi: 10.1007/978-3-030-60884-2_3
4. Aguilar Canto, F. J.: Eficacia de diferentes reglas hebbianas en el aprendizaje supervisado: Efficacy of different hebbian rules in supervised learning. Tecnología Educativa Revista CONAIC, vol. 7, no. 1, pp. 92–97 (2020) doi: 10.32671/terc.v7i1.22
5. Amato, G., Carrara, F., Falchi, F., Gennaro, C., Lagani, G.: Hebbian learning meets deep convolutional neural networks. In: Image Analysis and Processing–ICIAP 2019: 20th International Conference, pp. 324–334 (2019) doi: 10.1007/978-3-030-30642-7_29
6. Bartlett, M., Sejnowski, T. J.: Viewpoint invariant face recognition using independent component analysis and attractor networks. Advances in Neural Information Processing Systems, vol. 9 (1996)
7. Becker, D.: 5 celebrity faces dataset (2018)
8. Belcar, D., Grd, P., Tomičić, I.: Automatic ethnicity classification from middle part of the face using convolutional neural networks. Informatics, vol. 9, no. 1, pp. 18 (2022) doi: 10.3390/informatics9010018
9. Bienenstock, E. L., Cooper, L. N., Munro, P. W.: Theory for the development of neuron selectivity: orientation specificity and binocular interaction in visual cortex. Journal of Neuroscience, vol. 2, no. 1, pp. 32–48 (1982) doi: 10.1523/jneurosci.02-01-00032.1982
10. Bliss, T. V., Lømo, T.: Long-lasting potentiation of synaptic transmission in the dentate area of the anaesthetized rabbit following stimulation of the perforant path. The Journal of physiology, vol. 232, no. 2, pp. 331–356 (1973) doi: 10.1113/jphysiol.1973.sp010273
11. Brennan, V., Principe, J.: Face classification using a multiresolution principal component analysis: Neural networks for signal processing VIII, In: Proceedings of the 1998 IEEE Signal Processing Society Workshop, pp. 506–515 (1998) doi: 10.1109/nnsp.1998.710681
12. Chiu, C. T., Ding, Y. C., Lin, W. C., Chen, W. J., Wu, S. Y., Huang, C. T., Lin, C. Y., Chang, C. Y., Lee, M. J., Tatsunori, S., Chen, T., Lin, F. Y., Huang, Y. H.: Chaos LiDAR based RGB-D face classification system with embedded CNN accelerator on FPGAs. IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 69, no. 12, pp. 4847–4859 (2022) doi: 10.1109/tcsi.2022.3190430
13. Cooper, L. N., Bear, M. F.: The BCM theory of synapse modification at 30: Interaction of theory with experiment. Nature Reviews Neuroscience, vol. 13, no. 11, pp. 798–810 (2012) doi: 10.1038/nrn3353
14. Dayan, P., Abbott, L. F.: Theoretical neuroscience: computational and mathematical modeling of neural systems, MIT press (2005)
15. Geitgey, A.: Face recognition (2022)
16. Kamaruzaman, F., Shafie, A. A., Mustafah, Y. M.: Coincidence detection using spiking neurons with application to face recognition. Journal of Applied Mathematics, vol. 2015 (2015) doi: 10.1155/2015/534198
17. Khan, K., Attique, M., Syed, I., Gul, A.: Automatic gender classification through face segmentation. Symmetry, vol. 11, no. 6, pp. 770 (2019) doi: 10.3390/sym11060770
18. King, D. E.: Dlib-ml: A machine learning toolkit. The Journal of Machine Learning Research, vol. 10, pp. 1755–1758 (2009)
19. Kingma, D. P., Ba, J.: Adam: A method for stochastic optimization (2014) doi: 10.48550/ARXIV.1412.6980
20. Krizhevsky, A., Sutskever, I., Hinton, G. E.: ImageNet classification with deep convolutional neural networks. Communications of the ACM, vol. 60, no. 6, pp. 84–90 (2017) doi: 10.1145/3065386
21. Kuśmierz, Ł., Isomura, T., Toyoizumi, T.: Learning with three factors: modulating hebbian plasticity with errors. Current opinion in neurobiology, vol. 46, pp. 170–177 (2017) doi: 10.1016/j.conb.2017.08.020

22. Lagani, G., Falchi, F., Gennaro, C., Amato, G.: Comparing the performance of hebbian against backpropagation learning using convolutional neural networks. Neural Computing and Applications, vol. 34, no. 8, pp. 6503–6519 (2022) doi: 10.1007/s00521-021-06701-4
23. Law, C. C., Cooper, L. N.: Formation of receptive fields in realistic visual environments according to the Bienenstock, Cooper, and Munro (BCM) theory. In: Proceedings of the National Academy of Sciences, National Acad Sciences, vol. 91, pp. 7797–7801 (1994) doi: 10.1073/pnas.91.16.7797
24. Leibo, J. Z., Liao, Q., Anselmi, F., Freiwald, W. A., Poggio, T.: View-tolerant face recognition and hebbian learning imply mirror-symmetric neural tuning to head orientation. Current Biology, vol. 27, no. 1, pp. 62–67 (2017) doi: 10.1016/j.cub.2016.10.015
25. Lindsay, G. W., Rigotti, M., Warden, M. R., Miller, E. K., Fusi, S.: Hebbian learning in a random network captures selectivity properties of the prefrontal cortex. Journal of Neuroscience, vol. 37, no. 45, pp. 11021–11036 (2017) doi: 10.1523/jneurosci.1222-17.2017
26. Lu, P., Song, B., Xu, L.: Human face recognition based on convolutional neural network and augmented dataset. Systems Science and Control Engineering, vol. 9, no. sup2, pp. 29–37 (2021)
27. Magotra, A., Kim, J.: Improvement of heterogeneous transfer learning efficiency by using hebbian learning principle. Applied Sciences, vol. 10, no. 16, pp. 5631 (2020) doi: 10.3390/app10165631
28. Markram, H., Lübke, J., Frotscher, M., Sakmann, B.: Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs. Science, vol. 275, no. 5297, pp. 213–215 (1997) doi: 10.1126/science.275.5297.213
29. Munakata, Y., Pfaffly, J.: Hebbian learning and development. Developmental science, vol. 7, no. 2, pp. 141–148 (2004) doi: 10.1111/j.1467-7687.2004.00331.x
30. Oja, E.: Simplified neuron model as a principal component analyzer. Journal of mathematical biology, vol. 15, no. 3, pp. 267–273 (1982) doi: 10.1007/bf00275687
31. Riesenhuber, M., Poggio, T.: Models of object recognition. Nature neuroscience, vol. 3, no. 11, pp. 1199–1204 (2000) doi: 10.1038/81479
32. Rosenblatt, F.: The perceptron: A probabilistic model for information storage and organization in the brain. Psychological Review, vol. 65, no. 6, pp. 386 (1958) doi: 10.1037/h0042519
33. William, F., Aygun, R.: Convoforest classification of new and familiar faces using EEG. In: IEEE 16th International Conference on Semantic Computing (ICSC), pp. 274–279 (2022) doi: 10.1109/icsc52841.2022.00052
34. Yao, P., Wu, H., Gao, B., Eryilmaz, S. B., Huang, X., Zhang, W., Zhang, Q., Deng, N., Shi, L., Wong, H. S. P., Qian, H.: Face classification using electronic synapses. Nature communications, vol. 8, no. 1, pp. 15199 (2017) doi: 10.1038/ncomms15199

ISSN 1870-4069

Evaluación de métodos de aprendizaje supervisado para la clasificación de palabras utilizando señales de electroencefalografía

Denise Alonso-Vázquez[1], Tonatiuh Hernández-del-Toro[2], Omar Mendoza-Montoya[1], Ricardo Caraza[3], Hector R. Martinez[3], Carlos A. Reyes-García[2], Javier M. Antelis[1]

Tecnológico de Monterrey, Escuela de Ingeniería y Ciencias, México

Instituto Nacional de Astrofísica Óptica y Electrónica, Departamento de Ciencias Computacionales, México

Tecnológico de Monterrey, Escuela de Medicina y Ciencias de la Salud, México

denise.alonso.v@tec.mx

Resumen. En este trabajo evaluamos diferentes métodos para la clasificación de palabras pronunciadas a partir de señales de electroencefalografía (EEG). Se utilizó la red neuronal convolucional EEGNet y se compararon los resultados obtenidos con el uso de características en el dominio de la frecuencia y clasificadores tradicionales: LDA, SVM y RF. Se utilizó una base de datos de cinco palabras en español: "si", "no", "agua", "comida" y "dormir", adquirida mediante 32 canales de electroencefalografía distribuidos uniformemente sobre el cuero cabelludo en participantes sanos. La clasificación se realizó intra-sujeto y en todos los casos se obtuvieron porcentajes de exactitud superiores al azar (20 %). Utilizando la EEGNet se obtuvo un mejor desempeño respecto a los otros métodos, obteniendo una exactitud promedio de 75.05±7.30 % entre todos los participantes. El participante con mayor exactitud obtuvo 84.00±4.54 % y el de menor desempeño 60.00±8.48 %. También se encontró que la palabra que mejor decodifica este método es "agua" y la peor es "dormir". Este trabajo es un estudio preliminar para la decodificación del intento del habla en pacientes con esclerosis lateral amiotrófica bulbar, utilizando un método de adquisición de señales no-invasivo como el EEG.

Palabras clave: Electroencefalografía, decodificación del habla, EEGNet.

Denise Alonso-Vázquez, Tonatiuh Hernández-del-Toro, Hector R. Martinez, et al.

Evaluation of Supervised Learning Methods for Words Classification using Electroencephalography Signals

Abstract. In this work, we evaluate different methods for the classification of pronounced words from electroencephalography (EEG) signals. The EEGNet convolutional neural network was used, and the results obtained were compared with models based on the traditional classifiers LDA, SVM, and RF; that used characteristics in the frequency domain. A database of five Spanish words was used: "si", "no", "water", "food" and "sleep". EEG recordings were acquired from 32 uniformly distributed electroencephalography channels on the scalp of healthy participants. The classification was carried out intra-subject, and for all methods, higher than chance accuracy percentages were obtained (20%). The best classification performance was obtained by EEGNet, in comparison to the other methods, obtaining an average accuracy of 75.05±7.30% among all the participants. The participant with the highest accuracy obtained 84.00±4.54 %, and the one with the lowest performance 60.00±8.48 %. It was also found that the word that was best decoded by this method was "water' and the worst was "sleep". This work is a preliminary study for the decoding of attempted speech in patients with bulbar amyotrophic lateral sclerosis using EEG as non-invasive signal acquisition method.

Keywords: Electroencephalography, speech decoding, EEGNet.

1. Introducción

Existen diferentes enfermedades en las que las neuronas motoras mueren progresivamente, y como consecuencia, se pierde la capacidad de hablar. Un ejemplo es el caso de la esclerosis lateral amiotrófica (ELA), enfermedad en la que aproximadamente el 85 % de los pacientes experimentan síntomas de disfunción bulbar, como disminución de la comunicación verbal y función de deglución, afectando significativamente su calidad de vida [14].

A nivel global más de 5,000 personas son diagnósticadas por año, con una prevalencia de 1 en 20,000 personas [1]. En México al menos hasta el 2018 se registraron más de 6,000 casos diagnosticados [7]. Actualmente, existen diversos dispositivos de asistencia para pacientes con limitaciones del habla.

En pacientes con ELA y como consecuencia tetraplegia y anartria (trastorno en la expresión del lenguaje que consiste únicamente en la imposibilidad de articular los sonidos) el dispositivo más utilizado es el dispositivo de comunicación por seguimiento ocular, en el que el usuario necesita señalar y mantener la mirada en los comandos que se muestran en el monitor, lo cual es detectado con una cámara infrarroja, sin embargo, son dispositivos de alto costo [6].

Otra alternativa que se ha estudiado ampliamente es el uso de las interfaces cerebro-computadora (BCIs por las siglas en inglés de brain-computer interface), las cuales detectan y cuantifican las características de las señales cerebrales que indican la intención del usuario, traducen estas mediciones en tiempo real en comandos para el dispositivo y proporcionan retroalimentación simultánea al usuario [22].

Existen diferentes mecanismos para medir la actividad cerebral, utilizando técnicas invasivas y no-invasivas, la electroencefalografía (EEG) es considerada como el método más común en la medición de señales cerebrales ya que tiene una alta resolución temporal, es fácil de usar, segura y asequible [15], por lo tanto, la mayoría de las BCIs utilizan señales adquiridas con EEG.

Algunas BCIs trabajan con potenciales evocados, como el potencial P300 o el potencial evocado visual de estado estacionario (SSVEP por las siglas en inglés de Steady-state visually evoked potential), y otras con tareas cognitivas, como el movimiento imaginado [19].

En el P300, por lo general, el participante mira la pantalla donde parpadean los caracteres y selecciona uno de ellos prestándole atención. En deletreadores que funcionan con SSVEP se utilizan estímulos visuales o recuadros que parpadean a distintas frecuencias. Cuando el usuario se concentra en el elemento que desea seleccionar, se genera un potencial en la corteza visual con la misma frecuencia de parpadeo que la imagen.

En el movimiento imaginado, el participante imagina que está moviendo una de sus extremidades sin realizar ningún movimiento, y con la decodificación de estas señales se controla una BCI. Estos paradigmas han sido ampliamente utilizados, sin embargo, no decodifican directamente la respuesta relacionada con el habla.

Se han realizado diferentes estudios decodificando el habla, ya sea pronunciada, susurrada, silenciosa (solamente se gesticula sin emitir sonidos) o imaginada (pronunciación interna de la palabra, sin gesticular ningún movimiento y sin emitir ningún sonido). En [16], se decodifica el intento del habla en un participante con anartria como consecuencia de un accidente cerebrovascular.

Se implantó una matriz de 128 electrodos en la corteza sensoriomotora del habla y se utilizó un modelo de lenguaje natural que calculó la probabilidad de la siguiente palabra dadas las palabras anteriores en una secuencia. En [2] se presentó un enfoque para sintetizar el habla audible a partir del habla imaginada y el habla susurrada, utilizando un diccionario de 100 palabras holandesas y electrodos estereotácticos profundos implantados en un participante.

A pesar de que estos trabajos decodifican el habla en sus diferentes paradigmas, utilizan métodos invasivos, lo que aporta un riesgo a la salud del paciente, al igual que incrementa los costos respecto al EEG. Algunos trabajos a partir de señales de EEG clasifican clases gramaticales, en [10] con habla imaginada (sustantivos y verbos) y en [3] utilizando habla pronunciada (adverbios de decisión y sustantivos).

También se han decodificado vocales en español [20] y palabras cortas y largas en inglés [17]. En [9] evalúan diferentes métodos de clasificación tradicionales y tres redes neuronales convolucionales distintas, con el objetivo de encontrar una óptima combinación de hiperparámetros para la clasificación de palabras en español.

La mayoría de los métodos que utilizan señales de EEG como método de adquisión, implementan sus protocolos con participantes sanos, es decir, sin ningún padecimiento neurológico diagnósticado y ningún transtorno del habla. Por lo tanto, en este trabajo se muestra un estudio preliminar de decodificación de palabras pronunciadas en una base de datos propia, que posteriormente será implementado en pacientes con ALS bulbar mientras realizan la tarea de intento del habla.

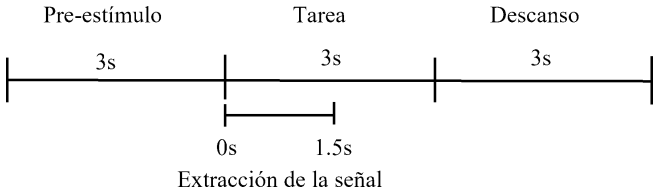


Fig. 1. Distribución de los estímulos en un ensayo.

Se evaluaron algoritmos de clasificación tradicionales ampliamente utilizados en BCIs (LDA, SVM y RF), utilizando una extracción de características en el dominio de la frecuencia, al igual que una red neuronal convolucional (EEGNet) que sigue la metodología comúnmente utilizada en las BCIs.

Se realizó una clasificación intra-sujeto de cinco clases correspondientes a las siguientes palabras: “si”, “no”, “agua”, “comida” y “dormir”. Los resultados obtenidos reflejan que la EEGNet tiene un mejor desempeño que los demás métodos de clasificación evaluados en este estudio, obteniendo una exactitud promedio entre todos los participantes de $75.05 \pm 7.30\,\%$ (nivel de azar del $20\,\%$).

2. Métodos

En esta sección se presenta la descripción de la base de datos, el preprocesamiento de las señales, los modelos de clasificación utilizados, el procedimiento de evaluación y las métricas de desempeño.

2.1. Descripción de la base de datos

La base de datos contiene el registro de señales de electroencefalografía (EEG) de diez participantes sanos, 6 hombres y 4 mujeres con edad promedio de 24.8 años (std=8.4 años), diestros y hablantes nativos del Español durante la tarea de habla pronunciada. Se utilizó un grupo de 5 palabras en Español que consideramos útiles para una persona con limitaciones en la comunicación: “si”, “no”, “agua”, “comida” y “dormir”. La duración de cada ensayo (ver Figura 1) fue de 9s divididos en estímulos de 3s. Los primeros 3s corresponden a una cruz de fijación donde la instrucción fue poner atención y evitar movimientos.

Posteriormente se muestra de forma aleatoria una de las cinco palabras, donde la tarea dada al participante fue pronunciar (una sola vez) la palabra de manera natural, es decir, la forma en la que normalmente habla, finalmente el bloque de descanso fue representado por una palmera con 3s de duración.

Se grabaron 4 bloques de 50 ensayos cada uno por participante, por lo tanto se obtuvieron 40 ensayos por cada palabra. Las señales fueron registradas mediante 32 electrodos activos (Ag/AgCl) distribuidos uniformemente sobre el cuero cabelludo de acuerdo al sistema 10-20.

El equipo utilizado fue el amplificador de bioseñales de alto rendimiento g.HIAMP 256 de g.tec. Los datos fueron adquiridos a una frecuencia de muestreo de 1200Hz,

se aplicó un filtro pasabanda Butterworth de 0.5Hz a 500Hz y un filtro Notch en 60Hz, colocando la referencia en el lóbulo de la oreja derecha y el electrodo de tierra en la posición AFz. Previo a la sesión todos los participantes declararon no tener ningún trastorno del habla o desorden neurológico diagnosticado, además de firmar el consentimiento informado y otorgar el permiso para el uso de sus datos.

2.2. Preprocesamiento

La señal fue submuestreada a 256 Hz, el segmento de tiempo utilizado fue de 0 s a 1.5 s, donde el cero corresponde al instante en el que la palabra aparece en la pantalla. De acuerdo a [11], el mayor pico de voltaje correspondiente a la señal de EEG contaminada por actividad muscular en la gesticulación de movimientos, se encuentra alrededor de los 30Hz, por lo que, se aplicó un filtro Butterworth pasabanda de 1 Hz a 20 Hz.

2.3. Modelos de clasificación

Para estudiar la discriminación entre las cinco palabras, por medio de la clasificación multiclase utilizando señales electroencefalográficas, se evaluaron dos alternativas, la primera basada en extracción de características y clasificadores convencionales, y la segunda basada en un modelo de aprendizaje profundo.

Extracción de características y clasificadores

– **Extracción de características utilizando potencia espectral:** La densidad de potencia espectral (PSD por las siglas en inglés de power spectral density) ha sido ampliamente utilizada en señales de EEG para proporcionar información de la distribución de la potencia en las distintas bandas de frecuencia que conforman la señal.

Utilizamos la transformación de frecuencia multitaper para estimar la potencia espectral usando ventanas de Hanning. Se calculó el promedio de la potencia espectral para cada canal, en cinco bandas de frecuencia con una resolución de 0.5Hz: delta (1-4Hz), theta (4-7 Hz), alfa (8-13 Hz), beta-baja (12-15) y beta-media (15 Hz - 20 Hz).

Por lo tanto, el vector de características resultante es $\mathbf{x} \in \mathbb{R}^{(N_{\text{Bandas}} \cdot N_{\text{Canales}}) \times 1}$, donde $N_{\text{Bandas}} = 5$ es el número de bandas de frecuencia y $N_{\text{Canales}} = 32$ es el número de canales de EEG, por consiguiente, la dimensión del vector es $N_{\text{Bandas}} \cdot N_{\text{Canales}} = 160$, con una etiqueta asociada $\mathbf{y} \in (\text{si, no, agua, comida, dormir})$.

Implementamos la selección de características con el objetivo de obtener una representación de baja dimensión del conjunto de datos original, pero con un alto poder de discriminación. Utilizamos un método de selección de características basado en el valor F de ANOVA, para seleccionar las 50 características mejor clasificadas. Por lo tanto, la dimensión final del vector es (1,50).

Para la clasificación, utilizamos algoritmos de aprendizaje máquina ampliamente utilizados en señales de EEG [18, 12], los cuales son análisis discriminante lineal, máquinas de soporte vectorial y árboles aleatorios, también conocidos como: LDA, SVM y RF respectivamente por sus siglas en inglés.

– **Análisis discriminante lineal (LDA):** es un algoritmo que puede ser utilizado para aprendizaje supervisado y no supervisado. Consiste en encontrar la proyección del hiperplano, definida por el vector discriminante $w*$, que maximice la distancia entre las medias de las clases al mismo tiempo que minimiza su varianza.

La función del discriminante toma la forma $f(x) = w^T x$, donde w es un vector aprendido de pesos y x representa el (n+1)-dimensional vector de características de la instancia a clasificar. El problema de optimización a resolver, es el siguiente:

$$w^* = \frac{w^T S_B w}{w^T S_W w}, \tag{1}$$

donde S_B es la varianza entre clases y S_W es la covarianza dentro de la clase[5]. La solución se puede obtener resolviendo un sistema de valores propios generalizado. El hiperplano obtenido puede ser utilizado para clasificación, reducción de dimensionalidad e interpretación de la importancia de las características dadas[23].

– **Máquina de soporte vectorial (SVM):** es un método utilizado para clasificación, regresión y estimación de densidad. Consiste en encontrar un hiperplano que discrimine entre las clases de manera que maximice los márgenes de separación entre él y los datos más cercanos en cada clase (llamados vectores de soporte). El problema de optimización a resolver, es el siguiente:

$$w^*, b^*, \zeta^* = \text{argmin}_{w,b,\zeta} \left(\frac{1}{2}\|w\|^2 + C \sum_{i=1}^{m} \zeta_i \right), \tag{2}$$

donde w es el vector de pesos, b el término de sesgo, ζ_i las variables de holgura, C un parámetro de regularización que determina el compromiso entre el ancho del margen y el error de entrenamiento y w^*, b^*, ζ^* los valores óptimos para el modelo [21]. También es posible que el límite de decisión sea no lineal a través de una función kernel, ya sea polinomial, radial o sigmoideo.

Para este trabajo utilizamos un kernel lineal con un parámetro de regularización de 1. El SVM es un clasificador binario, sin embargo, se puede ampliar fusionando varios de su tipo en un clasificador multiclase implementando el enfoque de "uno contra uno" [12].

– **Bosques aleatorios (RF):** son una modificación del bagging (empaquetado) que crea un grupo de árboles descorrelacionados y luego los promedia, con el objetivo de disminuir la varianza del modelo. En algunos problemas, el rendimiento de RF es similar a boosting (ayuda a disminuir el sesgo del modelo), y son más sencillos de entrenar y ajustar.

Un promedio $\bar{x}$ de variables aleatorias B cada una con varianza σ^2 tiene varianza σ^2/B. Si las variables son identicamente distribuidas, pero no necesariamente independientes con correlación positiva por pares ρ, la varianza del promedio es:

$$\text{var}(\bar{x}) = \rho\sigma^2 + \frac{1-\rho}{B}\sigma^2. \tag{3}$$

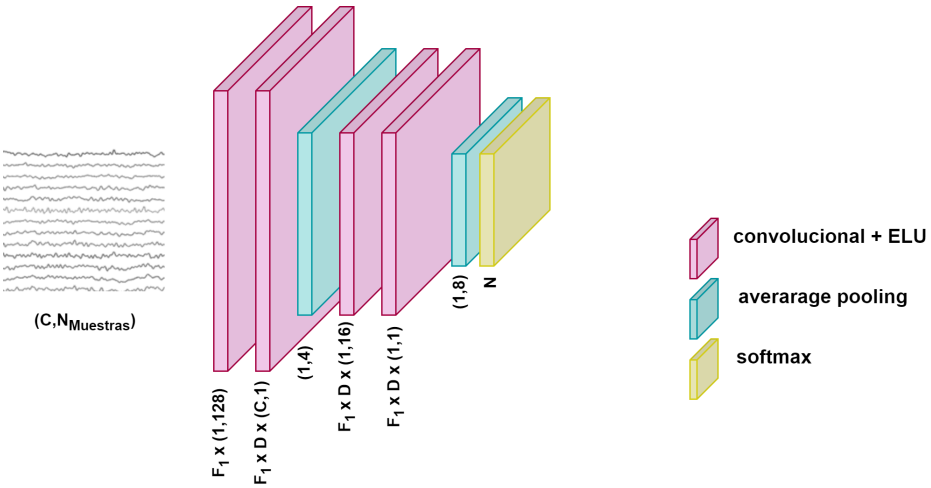


Fig. 2. Arquitectura de la red Neuronal EEGNet [8].

A medida que B aumenta, el segundo término desaparece, sin embargo, el primero permanece y el tamaño de la correlación de pares de árboles empaquetados, limita los beneficios del promedio. El objetivo de RF es mejorar la reducción de la varianza del bagging al reducir la correlación entre los árboles, sin aumentar demasiado la varianza. Esto se logra en el proceso de crecimiento de árboles a través de la selección aleatoria de las variables de entrada [4]. En este trabajo utilizamos un conjunto de 100 árboles.

Modelo de aprendizaje profundo

– **EEGNet:** La EEGNet es una red neuronal convolucional compacta para BCIs basadas en EEG, la cual se puede utilizar en diferentes paradigmas de BCI y entrenarse con una cantidad de datos muy limitados [8]. La arquitectura de la EEGNet (ver Figura 2) está compuesta por tres bloques: el primero se compone de una secuencia convolucional de dos pasos.

Comienza con una capa de convolución temporal para aprender filtros de frecuencia. Utilizamos $F1 = 8$, dónde $F1$ es el número de filtros temporales con un tamaño de kernel de (1,128), es decir, la mitad de la frecuencia de muestreo.

El segundo paso es una convolución profunda para aprender filtros espaciales para cada filtro temporal, con el objetivo de obtener una extracción eficiente de filtros espaciales específicos de frecuencia. El tamaño es (C, 1) con C definido como el número de canales, por lo tanto, C=32. El parámetro de profundidad D es es el número de filtros espaciales para aprender dentro de cada convolución temporal, D=2.

Este primer bloque está inspirado en el algoritmo Filter-Bank Common Spatial Pattern (FBCSP) [13]. Posteriormente se aplica Batch normalization entre la dimensión de los feature maps antes de aplicar el exponential linear unit (ELU) nonlinearity y la técnica Dropout para regularizar el modelo, con una probabilidad establecida en 0.85.

El bloque termina con una capa average pooling de tamaño (1, 4) que reduce la frecuencia de muestreo de la señal a 64Hz. El segundo bloque está compuesto por una

Tabla 1. Media aritmética ± desviación estándar de la exactitud total obtenida por participante en la clasificación de las cinco palabras (nivel de azar=20 %), para todos los métodos evaluados.

Participante	PSD+LDA	PSD+SVM	PSD+RF	EEGNet
1	38.00±9.09 %	43.00±12.3 %	39.00±5.75 %	81.00±2.23 %
2	37.00±7.37 %	34.00±5.18 %	29.00±5.18 %	79.00±7.41 %
3	59.50±8.73 %	62.50±3.53 %	44.00±8.22 %	75.50±9.25 %
4	29.00±5.75 %	32.00±7.58 %	30.00±6.85 %	68.00±7.79 %
5	41.50±2.23 %	36.50±7.42 %	38.00±5.12 %	73.50±9.11 %
6	32.00±4.80 %	38.00±5.97 %	31.00±2.85 %	79.00±4.54 %
7	53.00±7.34 %	50.50±7.37 %	48.00±4.11 %	70.00±8.10 %
8	50.00±10.5 %	54.00±12.3 %	49.50±4.81 %	84.00±4.54 %
9	43.50±9.94 %	45.00±9.84 %	37.50±9.01 %	60.00±8.48 %
10	31.00±3.35 %	32.00±8.73 %	31.00±2.24 %	80.50±6.71 %
Total	**41.45±10.1 %**	**42.75±10.3 %**	**37.70±7.54 %**	**75.05±7.30 %**

convolución separable formada por la combinación entre una convolución profunda (tamaño $(1, 16)$) y una convolución puntual, con $F2 = F1 * D$ dónde F2 es el número de filtros puntuales para aprender.

Al igual que en el bloque anterior se aplica Batch normalization entre la dimensión de los feature maps, el exponential linear unit (ELU) nonlinearity y la técnica Dropout para regularizar el modelo, con una probabilidad establecida en 0.85. El average pooling se establece en un tamaño de (1,8) para reducir las dimensiones.

En el bloque de clasificación se utiliza una softmax classification con N unidades, dónde N es el número de clases, por lo tanto N=5. La matriz que ingresa a la red neuronal tiene dimensión $\mathbf{x} \in \mathbb{R}^{N_{\text{Canales}} \times N_{\text{Muestras}}}$ con una etiqueta asociada $\mathbf{y} \in$ (si, no, agua, comida, dormir), donde $N_{\text{Canales}} = 32$ es el número de canales de EEG, y $N_{\text{Muestras}} = 385$ es el número de muestras contenidas en 1.5s de la señal.

Utilizamos los GPUs de Google Colab para el entrenamiento de la red desarrollada en Tensorflow utilizando Keras API. Se realizaron curvas de aprendizaje y basados en lo obtenido, se utilizaron 300 épocas. Para el ajuste del modelo se utilizó el optimizador Adam y la función de pérdida categorical cross-entropy.

2.4. Procedimiento de evaluación y métricas de desempeño

Los modelos se entrenaron intra-sujeto, es decir, el modelo se ajustó a cada uno de los participantes de forma independiente. Se realizó una clasificación multiclase de $N_{\text{Clases}} = 5$, donde cada clase corresponde a la pronunciación de cada una de las palabras. Para evaluar los modelos se utilizó validación cruzada de 5 iteraciones, es decir, el conjunto de características de cada participante se dividió en 5 grupos.

Las características de cuatro grupos se tomaron para ajustar los modelos de clasificación mientras que las características del grupo restante se utilizaron como un conjunto de datos de prueba para calcular las métricas de rendimiento.

Este procedimiento se repitió cinco veces, tomando un grupo diferente como conjunto de datos de entrenamiento y prueba en cada iteración, para garantizar que

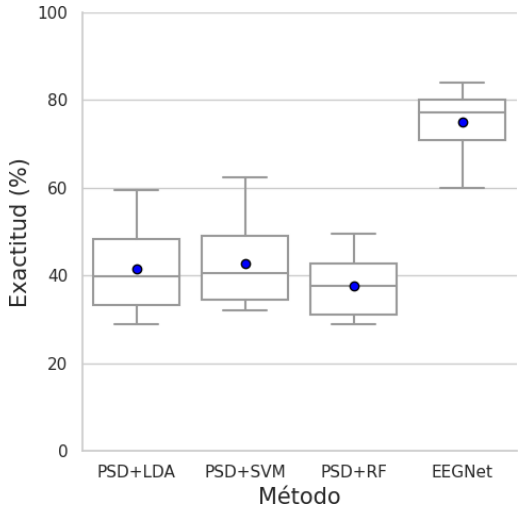


Fig. 3. Distribución de la exactitud total obtenida por cada método en la clasificación de las cinco palabras (nivel de azar=20 %). El punto corresponde a la media aritmética obtenida entre todos los participantes en cada uno de los métodos.

estos dos conjuntos siempre se excluyen mutuamente. Para evaluar el desempeño de los modelos, utilizamos las siguientes métricas:

– Exactitud de clasificación total, que representa el porcentaje total de valores correctamente clasificados.

– Precisión de clasificación por clase, indica qué tan confiable es nuestro modelo para predecir una clase específica.

– Sensibilidad por clase, se refiere al porcentaje de los elementos de la clase que fueron detectados correctamente entre todos los elementos de esa clase, es decir, el desempeño del modelo al detectar esa clase.

– Matriz de confusión para visualizar el número de predicciones de cada clase, respecto a las instancias en la clase real.

3. Resultados

La Tabla 1 contiene los resultados de la exactitud total obtenida por participante en la clasificación de las cinco palabras. Tomando en cuenta que el nivel de azar es del 20 % todos los participantes alcanzaron una exactitud promedio por encima del azar.

Utilizando la red neuronal convolucional EEGNet, se obtuvieron los porcentajes de exactitud más altos, donde el participante 8 logró el mejor desempeño con 84.00 ± 4.54 % de exactitud, mientras el peor caso con este método se obtuvo en el participante 9 con 60.00 ± 8.48 %.

El promedio de los resultados obtenidos utilizando la EEGNet fue de 75.05 ± 7.30 %, aproximadamente el doble del total obtenido en el método que combina

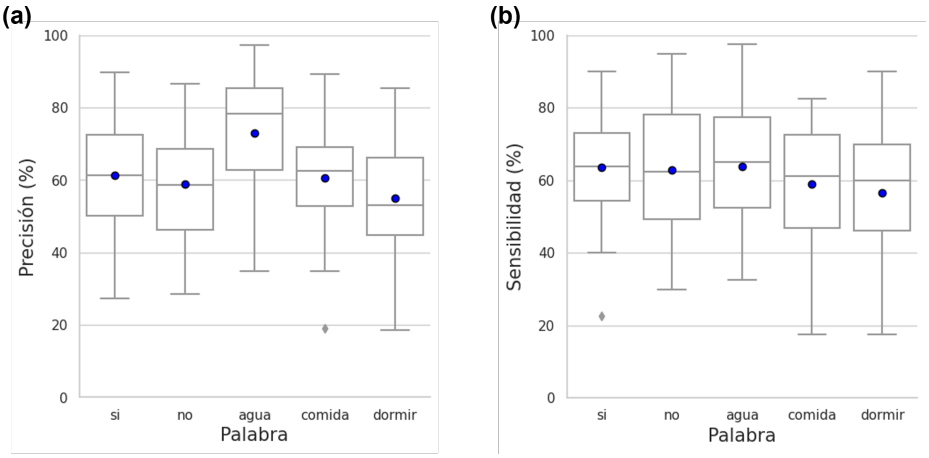


Fig. 4. Distribución de la precisión y sensibilidad por palabra, utilizando la EEGNet (nivel de azar=20 %). El punto corresponde a la media aritmética obtenida entre todos los participantes en cada una de las palabras. a) Porcentaje de precisión, b) porcentaje de sensibilidad.

PSD y RF donde se obtuvo el menor porcentaje de exactitud (37.70±7.54 %). Con PSD+RF el peor caso ocurre en el participante 2 (29.00±5.18 %) y su mejor caso al igual que con la EEGNet, en el participante 8 (49.50±4.81 %).

En los métodos que combinan PSD+LDA y PSD+SVM se obtuvieron resultados similares entre sí; 41.45±10.1 % y 42.75±10.3 % respectivamente. En estos dos casos los participantes que obtienen los porcentajes de exactitud más altos (participante 3 con 59.50±8.73 % y 62.50±3.53 %) y los mas bajos (participante 4 con 29.00±5.75 % 32.00±7.58 %) coinciden.

La Figura 3 muestra la distribución de los porcentajes de exactitud total alcanzados para cada uno de los métodos. Utilizando la EEGNet la mediana de los valores de exactitud total se encuentra en 77.25 % con una asimetría negativa, es decir, el 50 % de los valores están mayormente concentrados por encima de la mediana, mientras que, por debajo de ella el 50 % restante está más disperso.

Las combinaciones que se basan en la extracción de características de potencia junto con un algoritmo de clasificación, tienen la mediana de sus valores alrededor de 40 %, es decir, el doble del nivel de azar (20 %), además de una asimetría positiva.

Tomando en cuenta únicamente estos tres métodos, el promedio más alto se logra con la combinación PSD+SVM y el más bajo en PSD+RF. Con el objetivo de evaluar si la EEGNet tiene significativamente mejor desempeño que los métodos restantes, se realizó la prueba estadística Wilcoxon de una cola para cada uno de los métodos respecto a EEGNet.

Con un nivel de significancia $\alpha = 0.01$ y un p–valor=0.0009766 para cada una de las pruebas, se comprueba que estadísticamente los resultados obtenidos con la EEGNet son superiores respecto a cada uno de los 3 métodos restantes. Se calculó la distribución de los porcentajes de precisión y sensibilidad por clase obtenidos entre todos los participantes utilizando la EEGNet (ver Figura 4).

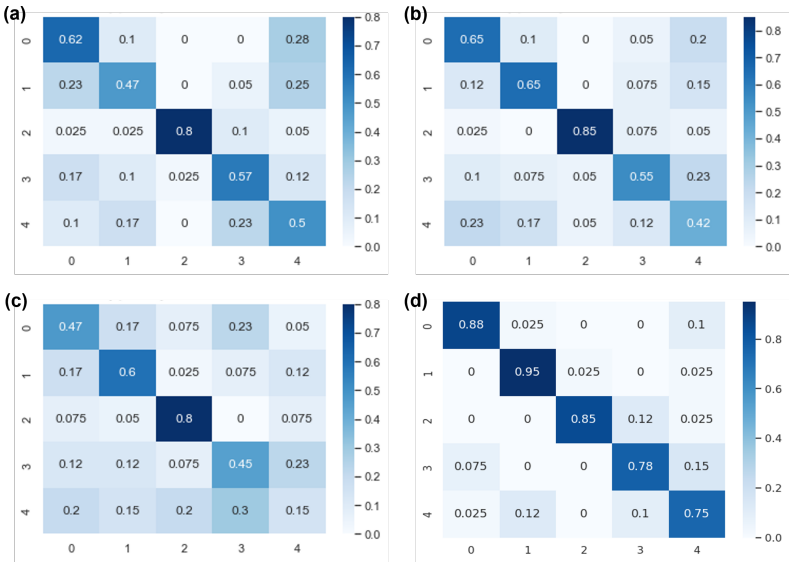


Fig. 5. Matrices de confusión obtenidas en los participantes con mejor porcentaje de exactitud. Método: a) PSD+LDA, b) PSD+SVM, c) PSD+RF y d)EEGNet. Cada valor numérico corresponde a cada una de las palabras de la siguiente manera: 0=“si”, 1=“no”, 2=“agua”, 3=“comida” y 4=“dormir”.

Con una precisión promedio por encima del 70 % se observa que la palabra que el modelo clasifica con mayor confiabilidad es “agua”. Mientras que, en la palabra “dormir” se obtiene la menor precisión promedio. La sensibilidad nos indica la relación de los elementos de la clase que fueron detectados correctamente entre todos los elementos de esa clase. Los resultados muestran que la sensibilidad del modelo es similar en todas las palabras (alrededor del 60 %).

La Figura 5 contiene las matrices de confusión de cada uno de los participantes con mejor desempeño en cada uno de los métodos. Se observa que para las combinaciones PSD+LDA, PSD+SVM y PSD+RF la palabra que mejor se reconoce es “agua”, resaltando esa clase entre las restantes de la diagonal con una exactitud igual o por encima de 80 %, mientras que el resto de los valores de la diagonal oscilan entre el 15 % y 65 %.

En contraste con lo anterior, utilizando la EEGNet la distribución de la exactitud en la diagonal es uniforme, es decir, solamente toma valores entre el 75 % y 95 %. La palabra que mejor se reconoce con este método es “no”, seguida de “si” y “agua”. En general, utilizando los cuatro métodos, la palabra que más se confunde con las demás es “dormir”, seguida de “comida”.

4. Conclusiones

En este trabajo se evaluó el desempeño de la red neuronal convolucional EEGNet y de tres métodos de clasificación tradicionales (LDA, SVM con kernel lineal y RF) en la

decodificación de palabras pronunciadas a través de señales adquiridas mediante EEG. Se utilizó un conjunto de 5 palabras en español "si", "no", "agua", "comida" y "dormir" pronunciadas por participantes sanos.

Para los clasificadores tradicionales se realizó una extracción de características basada en la densidad de potencia espectral de cinco bandas de frecuencia. La clasificación multiclase (nivel de azar del 20 %) se llevo a cabo por participante. El método con el mejor desempeño fue la EEGNet, obteniendo un promedio de 75.05 ± 7.30 % calculado con todos los participantes, donde el valor máximo fue 84.00 ± 4.54 % y el mínimo 60.00 ± 8.48 %. Utilizando este método, se encontró que la palabra que mejor clasifica el modelo es "agua" y la que menos reconoce es "dormir".

Los resultados indican que una red neuronal convolucional como la EEGNet tiene mejor desempeño que clasificadores tradicionales en la decodificación de palabras pronunciadas. Este trabajo presenta resultados prometedores en el uso de redes neuronales convolucionales como la EEGNet, para la decodificación de palabras a partir de señales de EEG. Este estudio preliminar nos otorga fundamentos para el desarrollo de un estudio posterior donde se realice la decodificación del intento del habla en pacientes con ALS utilizando un método no-invasivo como el EEG.

Referencias

1. ALS news today, how common is ALS? (2019) alsnewstoday.com/how-common-is-als/
2. Angrick, M., Ottenhoff, M. C., Diener, L., Ivucic, D., Ivucic, G., Goulis, S., Saal, J., Colon, A. J., Wagner, L., Krusienski, D. J., Kubben, P. L., Schultz, T., Herff, C.: Real-time synthesis of imagined speech processes from minimally invasive recordings of neural activity. Communications Biology, vol. 4, no. 1 (2021) doi: 10.1038/s42003-021-02578-0
3. Barrientos Rojas, S. J., Ramirez-Valencia, R., Alonso-Vazquez, D., Caraza, R., Martinez, H. R., Mendoza-Montoya, O., Antelis, J. M.: Recognition of grammatical classes of overt speech using electrophysiological signals and machine learning. In: IEEE 4th International Conference on BioInspired Processing (2022) doi: 10.1109/bip56202.2022.10032476
4. Bickel, P., Diggle, P., Fienberg, S., Gather, U.: Springer Series in Statistics (2005)
5. Bishop, C. M., Nasrabadi, N. M.: Pattern recognition and machine learning. vol. 4, no. 4, pp. 738 (2006)
6. Caligari, M., Godi, M., Guglielmetti, S., Franchignoni, F., Nardone, A.: Eye tracking communication devices in amyotrophic lateral sclerosis: Impact on disability and quality of life. Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration, vol. 14, no. 7–8, pp. 546–552 (2013) doi: 10.3109/21678421.2013.803576
7. Consejo nacional para el desarrollo y la inclusión de las personas con discapacidad. La Esclerosis Lateral Amiotrófica ELA (2018) www.gob.mx/conadis/articulos/la-esclerosis-lateral-amiotrofica-ela?idiom=es
8. Cooney, C., Korik, A., Folli, R., Coyle, D.: Evaluation of hyperparameter optimization in machine and deep learning methods for decoding imagined speech EEG. Sensors, vol. 20, no. 16, pp. 4629 (2020) doi: 10.3390/s20164629
9. Cooney, C., Korik, A., Folli, R., Coyle, D.: Evaluation of hyperparameter optimization in machine and deep learning methods for decoding imagined speech EEG. Sensors, vol. 20, no. 16, pp. 4629 (2020) doi: 10.3390/s20164629
10. Datta, S., Boulgouris, N. V.: Recognition of grammatical class of imagined words from EEG signals using convolutional neural network. Neurocomputing, vol. 465, pp. 301–309 (2021) doi: 10.1016/j.neucom.2021.08.035

11. Goncharova, I. I., McFarland, D. J., Vaughan, T. M., Wolpaw, J. R.: EMG contamination of EEG: spectral and topographical characteristics. Clinical Neurophysiology, vol. 114, no. 9, pp. 1580–1593 (2003) doi: 10.1016/s1388-2457(03)00093-2
12. Guler, I., Ubeyli, E. D.: Multiclass support vector machines for EEG-signals classification. IEEE Transactions on Information Technology in Biomedicine, vol. 11, no. 2, pp. 117–126 (2007) doi: 10.1109/titb.2006.879600
13. Keng-Ang, K., Yang Chin, Z., Zhang, H., Guan, C.: Filter bank common spatial pattern (FBCSP) in brain-computer interface. In: IEEE International Joint Conference on Neural Networks (2008) doi: 10.1109/ijcnn.2008.4634130
14. Lee, J., Madhavan, A., Krajewski, E., Lingenfelter, S.: Assessment of dysarthria and dysphagia in patients with amyotrophic lateral sclerosis: Review of the current evidence. Muscle and Nerve, vol. 64, no. 5, pp. 520–531 (2021) doi: 10.1002/mus.27361
15. Lotte, F., Congedo, M., Lécuyer, A., Lamarche, F., Arnaldi, B.: A review of classification algorithms for EEG-based brain–computer interfaces. Journal of Neural Engineering, vol. 4, no. 2, pp. R1–R13 (2007) doi: 10.1088/1741-2560/4/2/r01
16. Moses, D. A., Metzger, S. L., Liu, J. R., Anumanchipalli, G. K., Makin, J. G., Sun, P. F., Chartier, J., Dougherty, M. E., Liu, P. M., Abrams, G. M., Tu-Chan, A., Ganguly, K., Chang, E. F.: Neuroprosthesis for decoding speech in a paralyzed person with anarthria. New England Journal of Medicine, vol. 385, no. 3, pp. 217–227 (2021) doi: 10.1056/nejmoa2027540
17. Nguyen, C. H., Karavas, G. K., Artemiadis, P.: Inferring imagined speech using EEG signals: a new approach using Riemannian manifold features. Journal of Neural Engineering, vol. 15, no. 1, pp. 016002 (2017) doi: 10.1088/1741-2552/aa8235
18. Panachakel, J. T., Ramakrishnan, A. G.: Decoding covert speech from EEG-A comprehensive review. Frontiers in Neuroscience, vol. 15 (2021) doi: 10.3389/fnins.2021.642251
19. Rezeika, A., Benda, M., Stawicki, P., Gembler, F., Saboor, A., Volosyak, I.: Brain–uellers: A review. Brain Sciences, vol. 8, no. 4, pp. 57 (2018) doi: 10.3390/brainsci8040057
20. Sarmiento, L. C., Villamizar, S., López, O., Collazos, A. C., Sarmiento, J., Rodríguez, J. B.: Recognition of EEG Signals from Imagined Vowels Using Deep Learning Methods. Sensors, vol. 21, no. 19, pp. 6503 (2021) doi: 10.3390/s21196503
21. Schölkopf, B., Smola, A. J.: Learning with kernels: Support vector machines, regularization, optimization, and beyond. MIT press (2002) doi: 10.7551/mitpress/4175.001.0001
22. Wolpaw, J. R.: Brain–computer interfaces. Handbook of clinical neurology, pp. 67–74 (2013)
23. Xanthopoulos, P., Pardalos, P. M., Trafalis, T. B.: Linear discriminant analysis. Robust Data Mining, pp. 27–33 (2013) doi: 10.1007/978-1-4419-9878-1

ISSN 1870-4069

Estimación de mapas de sequías de EU mediante redes de Convolución-LSTM

Manuel Medrano[1], Héctor Rodríguez[1],
Rodrigo Lopez-Farias[2], Juan Flores[3],
Carlos Lara[4], Vicenç Puig[5]

[1] Tecnológico Nacional de México Campus Culiacán,
División de estudios de posgrado,
México

[2] Consejo Nacional de Ciencia y Tecnología,
Centro de Investigación en Ciencias de Información Geoespacial,
México

[3] Universidad de Oregón,
Departamento de Ciencias de la Computación e Informática,
Estados Unidos

[4] Centro de Investigación en Matemáticas,
Ciudad del Conocimiento Zacatecas,
México

[5] Universidad Politécnica de Cataluña,
Barcelona,
España

manuel.md@culiacan.tecnm.mx

Resumen. Los mapas de sequías son representaciones gráficas que identifican la severidad de la sequía en diferentes regiones. En este estudio, se propone una metodología para estimar mapas de sequías utilizando redes de Convolución-LSTM, las cuales permiten identificar patrones temporales y espaciales de las condiciones de sequías. Los resultados muestran que esta metodología es efectiva en la generación de estimaciones precisas de mapas de sequías, lo que la convierte en una herramienta valiosa para la gestión de recursos hídricos, la toma de decisiones informadas, y la prevención y mitigación de los impactos negativos de las sequías en diferentes regiones del mundo.

Palabras clave: Sequías, estimación, series de tiempo, convolución-lstm, procesamiento de imágenes.

Estimation of EU Drought Maps Using Convolution-LSTM Network

Abstract. Drought maps are graphical representations that identify the severity of drought in different regions. In this study, a methodology is proposed

to estimate drought maps using Convolution-LSTM networks, which allow identifying temporal and spatial patterns of drought conditions. The results show that this methodology is effective in generating accurate estimates of drought maps, which makes it a valuable tool for water resources management, informed decision making, and prevention and mitigation of the negative impacts of droughts in different regions of the world.

Keywords: Droughts, estimation, time series, convolution-LSTM, image processing.

1. Introducción

La sequía es un periodo seco prolongado en el ciclo climático natural que puede ocurrir en cualquier parte del mundo. Es un desastre de aparición lenta que se caracteriza por la falta de precipitaciones, lo que provoca una escasez de agua que puede afectar tanto a los seres humanos como a los ecosistemas naturales [10].

Existen diferentes tipos de sequía, que se clasifican según su impacto en la comunidad y la región: Sequía meteorológica, relacionada al déficit de precipitación sobre un período en una región; sequía hidrológica, cuando los ríos, acuíferos, estanques y reservas de agua están debajo de los niveles normales; sequía agrícola, sobre el déficit de humedad de suelo en largos términos; sequía socio-cultural, relacionado al suministro y demanda de agua para bienes económicos [3].

Entre los problemas más significativos asociados a las sequías se encuentran la falta de disponibilidad de agua potable, la escasez de alimentos debido al impacto en la agricultura y la ganadería, y la disminución de la productividad en la industria debido a la reducción en la disponibilidad de agua.

Estos impactos afectan la calidad de vida de las personas, especialmente en comunidades rurales y en países en desarrollo donde la agricultura es una parte fundamental de la economía y la subsistencia.

Por lo tanto, es importante tomar medidas preventivas y de gestión adecuadas para minimizar el impacto de la sequía en estas áreas. La educación sobre la gestión del agua y la implementación de tecnologías más eficientes en el uso del agua son algunas de las estrategias que se pueden implementar para enfrentar el problema de la sequía y sus consecuencias.

La información sobre las sequías se puede representar de diversas maneras, como los índices de sequía, la precipitación, la temperatura, la humedad del suelo, el índice de vegetación y los caudales y niveles de agua, incluso en forma de mapas de sequía [3].

En este estudio, se hace hincapié en el uso de los mapas de sequía como herramientas para visualizar la intensidad y extensión de la sequía en una determinada área.

Los mapas se generan a partir de datos recopilados por satélites, estaciones meteorológicas y otros sensores que miden la cantidad de lluvia y humedad del suelo [7]. Una área poco explorada consiste en utilizar mapas de sequías previos para estimar el próximo mapa.

Sin embargo, el uso de esta técnica puede resultar en un aumento exponencial en la complejidad de la estimación, ya que se debe considerar tanto la información espacial para tener en cuenta las condiciones del entorno, como la información temporal para considerar la evolución de la sequía en ese entorno a lo largo del tiempo.

La estimación de la sequía permite a las autoridades, investigadores, planificadores y a la población en general estar mejor preparados para hacer frente a las consecuencias de este fenómeno [1]. Entre los beneficios principales, se encuentran:

– Anticipación: Una estimación anticipada permite a las autoridades y comunidades prepararse con antelación y tomar medidas para minimizar el impacto negativo.
– Eficiencia: La utilización de las redes Convolución-LSTM permiten utilizar la información espacial y temporal de los mapas de sequías, generando mapas con mayor precisión y eficiencia, ayudando a identificar áreas criticas y a asignar recursos de manera eficiente.
– Mejora en la planificación: Los mapas de sequías ayudan a los planificadores a identificar áreas de mayor riesgo y tomar decisiones informadas en cuanto al uso de tierra, suministro de agua y otros recursos.
– Reducción de impactos negativos: La estimación y monitoreo de sequías ayudan a reducir impactos negativos en la sociedad, incluyendo perdida de cosechas, disminución de suministro de agua, perdida de biodiversidad y otros impactos ambientales y económicos.

Para predecir la evolución futura de un conjunto de datos, comúnmente se utilizan las series de tiempo. Estas representan una secuencia de datos correlacionados, tomados en intervalos fijos de tiempo. Una rama particularmente interesante en esta área son las Series de Tiempo de Imágenes (STI), que se componen de un conjunto de imágenes ordenadas cronológicamente y tomadas en intervalos fijos de tiempo.

Al igual que las series de tiempo, las STI presentan características de temporalidad, tendencia, estacionalidad, ciclos y auto-correlación. Esto las convierte en una herramienta poderosa para estimar la evolución temporal de variables relacionadas con el espacio, como las regiones sequía, la vegetación, los cambios en la cobertura del suelo o los desastres naturales.

En este trabajo, se utilizó un conjunto de mapas de sequías obtenidos del sitio web Drought Monitor en Estados Unidos, los cuales se organizaron en una Serie de Tiempo de Imágenes (STI). Esta STI permitió extraer información espacial y temporal mediante el uso de una arquitectura de red tipo Convolución-LSTM, la cual permite la estimación y generación del siguiente mapa de sequía.

En total, se obtuvieron 1,183 mapas de sequía que abarcan todas las regiones de Estados Unidos, ordenados cronológicamente en intervalos semanales desde el 4 de enero de 2000 hasta el 30 de agosto de 2022 [7].

El siguiente trabajo se divide en las siguientes secciones:

– Introducción: se presenta el contexto y la motivación para el trabajo, se explican los objetivos y se establece la estructura general del mismo.
– Trabajo relacionado: se describen diferentes trabajos que aplican técnicas diversas para la estimación y pronóstico de sequías en distintos formatos.

– Mapas de sequías: se detallan las características del conjunto de datos de mapas de sequías que se utilizarán.
– Metodología: se explica el conjunto de procedimientos realizados para la estimación y generación del siguiente mapa de sequías.
– Experimentación y resultados: se destacan los resultados obtenidos al poner en práctica los procedimientos especificados.
– Conclusiones: se presenta un resumen sobre la metodología planteada y se analizan los resultados obtenidos.

Con esta estructura, el lector puede tener una idea clara de la organización y el contenido del trabajo. Además, se utilizan frases más precisas y se mejora la redacción para hacer que el texto sea más legible y fácil de entender.

2. Trabajo relacionado

Existen numerosos trabajos que tratan sobre sequías y manipulación de imágenes, los cuales utilizan diversas técnicas y metodologías para abordar problemas particulares. A continuación, se describirán algunos de ellos que se centran en la problemática de las sequías.

En el trabajo de [8], se aborda el pronóstico del Índice Estandarizado de Precipitación (SPI) y el SPI de Evapotranspiración (SPEI) mediante la aplicación de la regresión logística con datos tomados de cinco ciudades europeas. De igual manera, los trabajos de [9] y [6] aplicaron la regresión logística para datos tomados de Turquía y del este de China, respectivamente.

En la actualidad, existen diversos enfoques para mejorar la precisión en el pronóstico de sequías. Uno de ellos es la combinación de modelos estocásticos ARIMA con otras arquitecturas. Un ejemplo de ello es el trabajo de [11], que utiliza un modelo híbrido basado en Transformación de Wavelet (WT), ARIMA y LSTM para pronosticar sequías a partir de información sobre la precipitación mensual y anual.

Por otro lado, [12] aplican una combinación de métodos ARIMA y LSTM para el pronóstico del índice SPEI, obteniendo una alta precisión en las estimaciones. Otro trabajo relevante es el de [2], donde se crea un modelo híbrido de ARIMA, Algoritmos Genéticos y Redes Neuronales, obteniendo resultados muy favorables para el pronóstico de SPI.

Entre los enfoques más utilizados para mejorar el pronóstico de sequías se encuentran las Redes Neuronales Artificiales (ANN). Por ejemplo,[4] evalúan el SPI Anual (SIAP) y el índice Estandarizado de Almacenamiento de Agua (SWSI) utilizando una combinación de WT y ANN (WANN), logrando coeficientes de correlación altos en la mayoría de los escenarios.

En [14] realizan una comparativa de diferentes métodos para el pronóstico de sequías en la cuenca del río Haihe, demostrando que las WANN obtienen mejores resultados de pronóstico en los valores SPI-6 y SPI-12.

Trabajar con mapas de sequías tiene su propia complejidad. En el trabajo [13], se crearon estos mapas a partir de un análisis de sequía, para realizar análisis más complejos sobre las características espacio-temporales de ocurrencia y patrones de propagación de sequía regional.

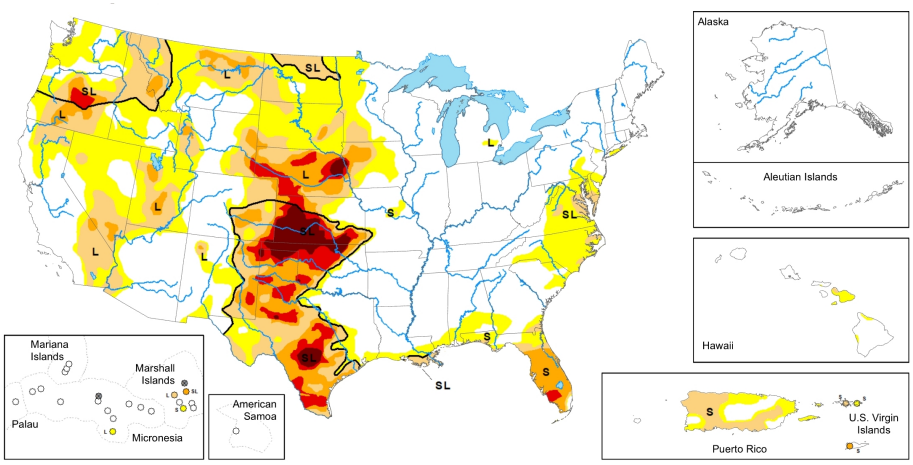


Fig. 1. Representación visual del mapa de sequías con delimitaciones. Tomado de [7].

En cambio, el trabajo [5] utilizó el SPI sobre mapas de sequías, concluyendo que las sequías ocurren con frecuencia durante la primavera con una tendencia de breves periodos de sequía frecuente.

3. Mapas de sequías

En este trabajo se utilizan datos obtenidos del monitor de sequías de Estados Unidos [7]. Donde se muestra un mapa de sequías tomado periódicamente semana tras semana. El mapa es creado a partir de un grupo de expertos que toma la información en crudo y la transforma a un nivel de sequía, dependiendo de las características de la información obtenida.

El nivel de sequía es asignado a cada una de las regiones del mapa y es transformado a un mapa de sequía. Estos niveles de sequías se dividen en cinco categorías, cada una de ellas mostrando su grado de impacto en las áreas afectadas, y la ausencia de sequía en color blanco. En el monitor de sequías se especifica cada una de las categorías como:

- **Anormalmente seco (D0)** : Se trata de una sequía de corta duración que ralentiza la siembra-crecimiento de los cultivos/pastos.
- **Sequía moderada (D1)** : Algunos cultivos o pastizales han resultado dañados, y los arroyos, embalses o pozos están bajando.
- **Sequía severa (D2)** : La posibilidad de pérdidas en los cultivos/pastizales y la escasez de agua son habituales.
- **Sequía extrema (D3)** : Pérdidas significativas en cultivos/pastizales y restricciones/escasez de agua generalizadas.
- **Sequía excepcional (D4)** : Pérdidas excepcionales en cultivos/pastos y emergencias hídricas debidas a la escasez de agua en embalses, arroyos y muros.

De esta manera se puede crear un mapa de sequía que refleje la región afectada y el grado de intensidad de la sequía. Se muestra un ejemplo de estos mapas en la Figura 1.

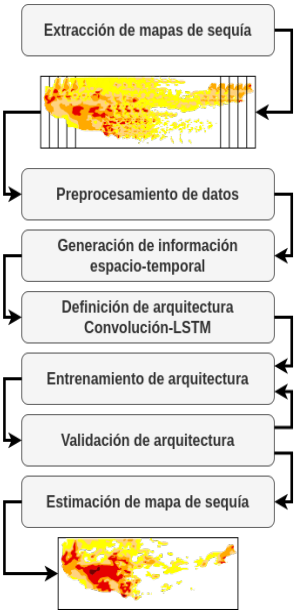


Fig. 2. Procedimiento para la estimación de mapas de sequías mediante Convolución-LSTM.

4. Metodología

En esta sección se presentan detalladamente los procedimientos y técnicas utilizadas para llevar a cabo la estimación de mapas de sequías mediante redes de Convolución-LSTM. se describirá la fuente y forma de extraer los mapas de sequías.

El preprocesamiento de los mapas para adaptarlo a las necesidades del problema. La generación de información espacio-temporal donde se transforman las series de tiempo en un conjunto de datos para estimación. Pasar a definir la arquitectura mediante el uso de redes de Convolución-LSTM, el entrenamiento y validación de la arquitectura. La Figura 2 muestra el orden de los procedimientos.

4.1. Extracción de mapas de sequía

El sitio *U.S. Drought monitor* ofrece un apartado de descarga de información. El formato original es de tipo *Geographic Information System* (GIS) obtenido a través de un servicio web *Web Map Service* (WMS). Este servicio provee las imágenes en formato ".PNG"de mapas georreferenciados que contienen la información libre de delimitaciones. Cada archivo WMS contiene un mapa de sequía basado en la forma de la proyección WGS84 [7].

El procedimiento de adquisición del mapa fue un procedimiento largo, se realizó de manera manual descargando mapa de sequía por mapa. La pestaña de la página de *GIS Data* contiene una tabla con la información de cada semana. Uno de los formatos que ofrece es el servicio WMS, que permite obtener una imagen PNG con el mapa de sequía.

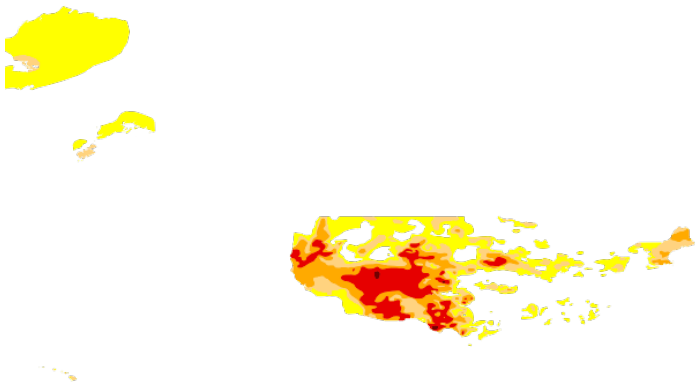


Fig. 3. Representación visual del mapa de sequías sin delimitaciones. Tomado de [7].

Para guardar el mapa se utilizó la función interna del navegador Google Chrome ”Guardar imagen como”, almacenando la imagen en una carpeta, donde la imagen fue etiquetada con su respectiva fecha de muestra. Al final se obtuvieron un total de 1,183 mapas de sequías, creando un conjunto de datos que va desde el 4 de Enero del 2000 hasta el 30 de Agosto del 2022.

Estos mapas de sequías originalmente tienen una resolución de (640, 480, 3), donde se presentan las dimensiones de anchura, altura y mapa de colores. La Figura 3 muestra el mapa de sequías completo.

Se decidió limitar el análisis de los mapas de sequía a la región peninsular de Estados Unidos. Esto permitió reducir la cantidad de características a procesar, pasando de una resolución original de (640, 480) a (122,360), y a la vez, enfocarse en la región de interés.

Los mapas de sequía cubren todas las regiones de Estados Unidos, incluyendo Alaska y Puerto Rico, pero para este estudio se optó por trabajar solo con la región peninsular. La Figura 4 muestra la zona resultante del mapa de sequías limitado a la región peninsular de Estados Unidos.

4.2. Preprocesamiento

En este trabajo se plantea el uso de sequías en escala de grises y monocromática. La escala de color original de los mapas de sequías es con formato ”RGB”. Por lo que, es necesario hacer una transformación de esas imágenes a dos escalas de colores. La Figura 5 muestra las imágenes en estas escalas de color.

Como se puede apreciar, al utilizar la escala de grises se mantiene la interpretación original sobre los grados de intensidad de sequía en varias regiones.

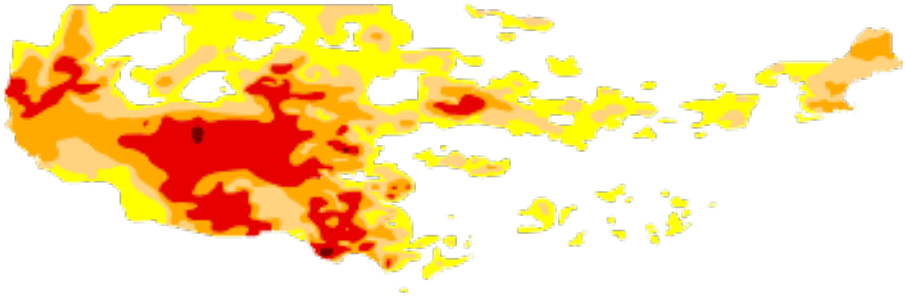


Fig. 4. Representación visual del mapa de sequías con área de interés.

Sin embargo, al aplicar una escala de color monocromática se modifica la interpretación de la información. Los mapas monocromáticos destacan únicamente los lugares donde hay sequía o no, sin tener en cuenta el grado de intensidad de la misma.

Después de transformar la escala de color de las imágenes, se mantiene el rango de valores original de 0 a 255. Sin embargo, en la escala monocromática, los valores se limitan a 0 o 255. Para ajustar los valores a un rango específico, se aplica un proceso de normalización. En este caso, se realiza una normalización que convierte los valores de los píxeles de las imágenes, que inicialmente oscilan entre 0 y 255, a un rango de 0 a 1.

4.3. Generación de información espacio-temporal

Las redes de convolución-LSTM son un tipo de red neuronal especial, esta requiere un formato especifico en los datos de entrada para su correcto funcionamiento. Se define una ventana temporal, la cual especifica la cantidad de datos simultáneos para formar subconjuntos de datos a los cuales se les asociará un objetivo. El objetivo es aquel subconjunto de datos de referencia que toma la red para intentar estimarlo.

Por lo anterior, para adecuar la serie temporal de mapas de sequías a una entrada para la red Convolución-LSTM, se transforma en un cubo de datos que representa información espacio-temporal. La información ”x” será utilizada como entrada para la red, mientras que la información ”y” será el objetivo de la misma. El formato de los subconjuntos de datos se define con 4 dimensiones, de esta manera: (Ventana temporal, Ancho, Alto, Color). La Figura 6 muestra esta transformación de la información.

4.4. Definición de arquitectura convolución-LSTM

Al definir la arquitectura de la red Convolución-LSTM, es importante considerar varios factores, como los tipos de capas a utilizar, las funciones de activación, la cantidad y forma de los kernels, y, lo más importante, la forma de los datos de entrada y salida. Para ello, se deben utilizar capas de Convolución-LSTM que requieren al menos la definición de un conjunto de parámetros: la cantidad y tamaño de los kernels y la función de activación.

Fig. 5. Transformación de color a escala de grises y monocromática.

Otra capa importante a utilizar es la capa de *BatchNormalization*. Esta capa permite mantener la normalización de los datos entre las capas intermedias de la red profunda, lo que ayuda a reducir el efecto del cambio en la distribución de los datos en cada capa y a acelerar el proceso de entrenamiento.

Debido a que el objetivo de las muestras y la arquitectura es estimar y generar un cubo de datos, la última capa debe ser de Convolución 3D, en la cual se define un solo kernel de 3 dimensiones. De esta manera, la entrada y salida de la arquitectura tendrán 4 dimensiones (Ventana temporal, Ancho, Alto, Color).

Una vez que las capas de la red han sido definidas, se procede a compilar el modelo, lo que implica la definición de ciertos aspectos importantes como la función de pérdida y el optimizador. En el caso de este trabajo, se utiliza la función de pérdida *Binary Crossentropy* ya que esta permite comparar la predicción del modelo con el objetivo verdadero y calcular la diferencia entre ellos.

En cuanto al optimizador, se utiliza el algoritmo *Adam*, el cual es comúnmente utilizado en problemas que requieren la optimización de funciones de pérdida no lineales.

4.5. Entrenamiento de la arquitectura

Entrenar una arquitectura consisten en alimentarla con un conjunto de datos de entrenamiento y validación. Al utilizar la función ”fit” serán necesarios diferentes parámetros a configurar. Entre los parámetros a utilizar, se encuentran los siguientes:

- Datos de entrenamiento: datos utilizados para alimentar el modelo en todo el entrenamiento.
- Objetivos de entrenamiento: datos utilizados como objetivos para que la red pueda estimar el futuro.
- Épocas: número de veces que el modelo se entrena con los datos de entrenamiento.
- *Batch_size*: número de muestras que se utilizan en cada actualización del modelo. Una actualización se refiere a un paso hacia adelante y hacia atrás a través de la red neuronal para actualizar los pesos del modelo.
- Datos de validación: datos utilizados para evaluar el rendimiento del modelo en cada época durante el entrenamiento. El objetivo es medir la capacidad del modelo para generalizar en nuevos datos y evitar el sobre ajuste.
- *Callbacks*: lista de objetos que se llaman al final de cada época para realizar acciones específicas, como guardar el modelo o detener el entrenamiento temprano si el rendimiento no mejora.

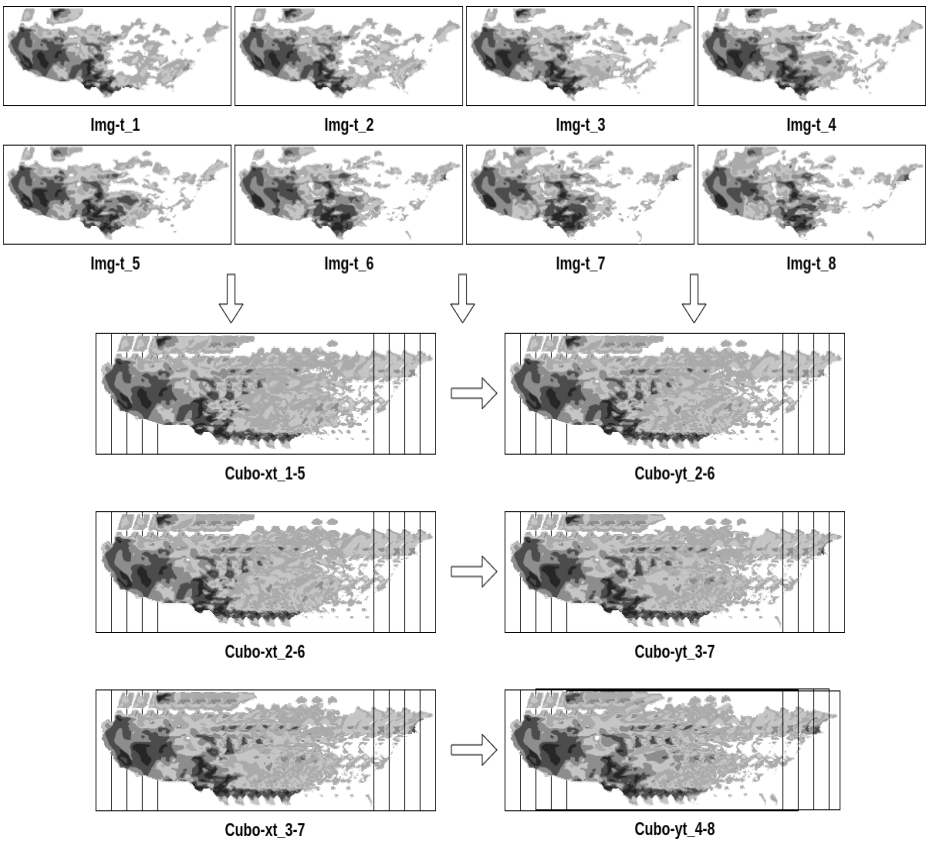


Fig. 6. Representación visual de los cubos de información espacio temporal.

4.6. Validación de arquitectura

Una vez que la arquitectura haya sido entrenada, se realiza un proceso de validación. En este procedimiento se evalúa el rendimiento del modelo en datos que no se utilizaron para entrenamiento.

Se utiliza un tercer conjunto de datos llamado conjunto de pruebas, el cual contiene un conjunto de muestras que el modelo nunca ha visto. El objetivo de la validación es medir la capacidad del modelo para generalizar en nuevos datos y evitar el sobre ajuste.

Dado que no se encontraron trabajos similares en la literatura, no es posible comparar los resultados obtenidos con otras metodologías. Para establecer un punto de referencia y evaluar la precisión de la metodología planteada, se emplea el modelo Naive.

Este modelo es una técnica simple que consiste en predecir que el valor futuro de una serie de tiempo será igual al valor actual, es decir, que no habrá cambios en la serie de tiempo.

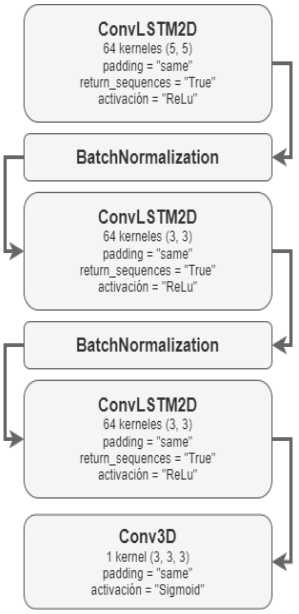


Fig. 7. Arquitectura Convolución-LSTM.

Para comparar las arquitecturas, se utiliza la medida del Error Cuadrático Medio (MSE por sus síglas en Ingles). Este error indica cuánto difieren las estimaciones de la arquitectura de los valores reales, siendo un valor más bajo indicativo de una mayor precisión del modelo.

El MSE se calcula sumando los errores cuadráticos para cada punto del tiempo y dividiendo el resultado entre el número total de puntos de la serie de tiempo. La Ecuación 1 muestra el calculo de esta función de error:

$$MSE = \frac{1}{N} * \sum_{t=1}^{N} (y_t - f_t)^2. \tag{1}$$

5. Experimentación y resultados

En este trabajo, se implementó una arquitectura basada en redes de Convolución-LSTM para la estimación de mapas de sequías. En esencia, la arquitectura se compone de tres capas de Convolución-LSTM, dos capas tipo *BatchNormalization* y una capa de salida de Convolución 3D. La Figura 7 muestra las especificaciones de cada capa de la arquitectura.

Se realizaron todos los experimentos en una única máquina con las siguientes especificaciones: CPU Intel Core i7-8700, 32 GB de memoria RAM y 2 NVIDIA GeForce RTX 2080 Ti. Es importante destacar que durante el entrenamiento de la arquitectura, ambas GPU fueron utilizadas simultáneamente mediante el uso de la estrategia *Mirrored Strategy* de la biblioteca TensorFlow.

Tabla 1. MSE promedio de los experimentos en escala monocromática.

Estimación	Naive
0.0229649	0.0254891

Tabla 2. MSE promedio de los experimentos en escala de grises.

Estimación	Naive
0.0178309	0.0199593

Se aplicó la arquitectura mencionada en dos conjuntos de datos diferentes. En el primer experimento, se utilizaron muestras en formato monocromático, donde el mapa de sequías solo indicaba si había o no sequía en una región determinada. En el segundo experimento, se usaron muestras en escala de grises, donde el mapa de sequías representaba el grado de sequía en distintas regiones.

Se utilizó una ventana temporal de 9 elementos por muestra para el conjunto de datos en escala monocromática. Después de pasar por las fases de preparación de datos, se obtuvieron 656 datos para entrenamiento, 165 datos para validación y 353 datos para pruebas.

Esto resultó en un subconjunto de cubos de datos con las dimensiones (9, 122, 360, 1). En cambio, para el conjunto de datos en escala de grises se aplicó una ventana temporal de 6 elementos por muestra. Luego de la preparación de datos, se obtuvieron 658 datos para entrenamiento, 165 datos para validación y 354 datos para pruebas. Esto dio como resultado un subconjunto de cubos de datos con las dimensiones (6, 122, 360, 1) para su estimación.

Los valores promedio del MSE se calcularon para cada uno de los conjuntos de datos de prueba, y se utilizó el modelo Naive como punto de referencia en ambos casos: tanto en la escala monocromática como en la escala de grises. Los resultados se presentan en la Tabla 1 para la escala monocromática y en la Tabla 2 para la escala de grises, junto con la correlación de los valores de MSE promedio correspondientes.

La metodología planteada en este trabajo es capaz de generar estimaciones de mapas de sequía con el mismo tamaño que los mapas originales. Para facilitar la interpretación visual de las estimaciones, se agregaron las delimitaciones geográficas de Estados Unidos. En las Figuras 8 y 9 se presentan ejemplos de imágenes originales, estimadas y del modelo Naive para las escalas monocromática y de grises, respectivamente.

6. Conclusiones

Los resultados presentados en este trabajo muestran la efectividad de la metodología propuesta basada en redes de Convolución-LSTM para la estimación de mapas de sequías. La generación precisa y actualizada de mapas de sequías puede ser útil para la toma de decisiones informadas y la gestión adecuada de los recursos hídricos.

En comparación con el modelo Naive, la metodología propuesta presentó un menor valor de MSE en la estimación, lo que sugiere que esta técnica puede ser una herramienta efectiva en la prevención y mitigación de los impactos negativos de las sequías en diferentes regiones del mundo.

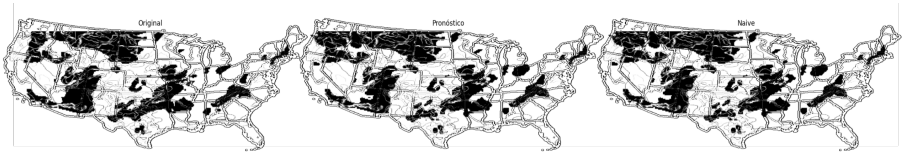


Fig. 8. Estimación del mapa de sequías en escala monocromática.

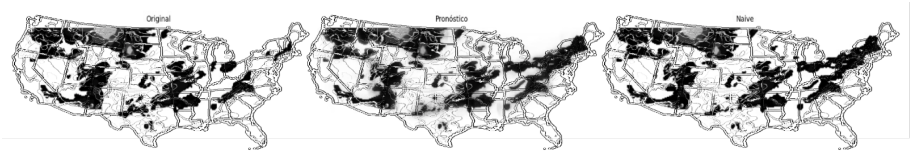


Fig. 9. Estimación del mapa de sequías en escala de grises.

Es fundamental considerar la complejidad computacional y la necesidad de memoria que implica el uso de redes Convolución-LSTM para la estimación de mapas de sequías. Durante los experimentos, se requirió el máximo rendimiento del equipo de hardware. Estas redes demandan una gran capacidad de procesamiento y una cantidad significativa de memoria tanto para el entrenamiento como para la estimación. Por lo tanto, es esencial asegurarse de contar con el hardware y almacenamiento adecuados antes de implementar esta metodología en diferentes contextos.

A pesar de lo anterior, los resultados obtenidos sugieren que la metodología propuesta obtiene estimaciones favorables. Por lo tanto, se espera que la investigación en esta área siga avanzando y que se encuentren formas de mejorar la eficiencia computacional de las redes de Convolución-LSTM para facilitar su implementación en diferentes contextos.

Sin embargo, es importante destacar que la metodología propuesta no es una solución única para la estimación de mapas de sequía, ya que se necesitaría adaptarla a las condiciones específicas de cada región y a los datos disponibles. Aun así, los resultados sugieren que la metodología propuesta es una herramienta valiosa para mejorar las estimaciones de mapas de sequías, lo que tendrá un impacto significativo en la gestión de recursos y la prevención de los impactos negativos de las sequías en la sociedad y el medio ambiente.

Referencias

1. Alawsi, M. A., Zubaidi, S. L., Al-Bdairi, N. S., Al-Ansari, N., Hashim, K.: Drought forecasting: A review and assessment of the hybrid techniques and data pre-processing. Hydrology, vol. 9, no. 7 (2022) doi: 10.3390/hydrology9070115
2. Alquraish, M., Abuhasel, K. A., Alqahtani, A. S., Khadr, M.: SPI-based hybrid hidden Markov–GA, ARIMA–GA, and ARIMA–GA–ANN models for meteorological drought forecasting. Sustainability, vol. 13, no. 22 (2021) doi: 10.3390/su132212576
3. Balti, H., Abbes, A. B., Mellouli, N., Farah, I. R., Sang, Y., Lamolle, M.: A review of drought monitoring with big data: Issues, methods, challenges and research directions. Ecological Informatics, vol. 60 (2020) doi: 10.1016/j.ecoinf.2020.101136

4. Khan, M. M., Muhammad, N. S., El-Shafie, A.: Wavelet-ANN versus ANN-based model for hydrometeorological drought forecasting. Water, vol. 10, no. 8 (2018) doi: 10.3390/w10080998
5. Lee, J. H., Cho, K. J., Kim, C. J., Park, M. J.: Analysis on the spatio-temporal distribution of drought using potential drought hazard map. Journal of Korea Water Resources Association, vol. 45, no. 10, pp. 983–995 (2012) doi: 10.3741/JKWRA.2012.45.10.983
6. Meng, L., Ford, T., Guo, Y.: Logistic regression analysis of drought persistence in east China. International Journal of Climatology, vol. 37, no. 3, pp. 1444–1455 (2017) doi: 10.1002/joc.4789
7. National Drought Mitigation Center: United states drought monitor. (2022) https://droughtmonitor.unl.edu/CurrentMap/StateDroughtMonitor.aspx#:~:text=Drought%20Mitigation%20Center-,National%20Drought%20Mitigation%20Center,the%20practice%20of%20drought%20planning
8. Stagge, J. H., Kohn, I., Tallaksen, L. M., Stahl, K.: Modeling drought impact occurrence based on meteorological drought indices in Europe. Journal of Hydrology, vol. 530, pp. 37–50 (2015) doi: 10.1016/j.jhydrol.2015.09.039
9. Tatli, H.: Downscaling standardized precipitation index via model output statistics. Atmósfera, vol. 28, no. 2, pp. 83–98 (2015) doi: 10.1016/S0187-6236(15)30002-3
10. World Health Organization: Drought. (2022) https://www.who.int/health-topics/drought#tab=tab_1
11. Wu, X., Zhou, J., Yu, H., Liu, D., Xie, K., Chen, Y., Hu, J., Sun, H., Xing, F.: The development of a hybrid wavelet-ARIMA-LSTM model for precipitation amounts and drought analysis. Atmosphere, vol. 12, no. 1 (2021) doi: 10.3390/atmos12010074
12. Xu, D., Zhang, Q., Ding, Y., Zhang, D.: Application of a hybrid ARIMA-LSTM model based on the SPEI for drought forecasting. Environmental Science and Pollution Research, vol. 29, no. 3, pp. 4128–4144 (2022) doi: 10.1007/s11356-021-15325-z
13. Yoo, J., So, B.-J., Kwon, H.-H., Kim, T.-W.: Development of drought map based on three-dimensional spatio-temporal analysis of drought. KSCE Journal of Civil and Environmental Engineering Research, vol. 40, no. 1, pp. 25–33 (2020) doi: 10.12652/Ksce.2020.40.1.0025
14. Zhang, Y., Li, W., Chen, Q., Pu, X., Xiang, L.: Multi-models for SPI drought forecasting in the north of Haihe river Basin, China. Stochastic Environmental Research and Risk Assessment, vol. 31, pp. 2471–2481 (2017) doi: 10.1007/s00477-017-1437-5

ISSN 1870-4069

Segmentación en color para el reconocimiento de leucemia mieloide aguda

José D. Sánchez-Chamorro[1], Rocio Ochoa-Montiel[1],
José Federico Ramírez-Cruz[2], Miguel A. Carrasco-Aguilar[1]

[1] Universidad Autónoma de Tlaxcala,
Facultad de Ciencias Básicas, Ingeniería y Tecnología,
México

[2] Instituto Tecnológico de Apizaco,
Tecnológico Nacional de México,
México

{demetriosanchezchamorro, ma.rocio.ochoa,
federico.ramirez, macarras2010}@gmail.com

Resumen. En el presente trabajo se propone un método de segmentación en color aplicado al reconocimiento de cinco tipos de leucemia mieloide aguda. Se utilizan técnicas clásicas de procesamiento digital de imágenes que derivan en un método de segmentación competitivo con respecto a otros métodos usados para el reconocimiento de leucemia. Los resultados muestran que la segmentación propuesta es adecuada para imágenes de distinta resolución, así como para imágenes con características distintas. La precisión de clasificación en la mayoría de los casos, supera el $90\,\%$ utilizando como clasificadores una red neuronal tipo Perceptron multicapa (MLP), una Máquina de vectores de soporte (SVM), Bosques aleatorios (RF) y Bayes.

Palabras clave: Segmentación en color, aprendizaje supervisado, leucemia.

Color Segmentation for Acute Myeloid Leukemia Recognition

Abstract. In this work, a color segmentation method applied to the recognition of five types of acute myeloid leukemia is proposed. Classical digital image processing techniques are used, resulting in a competitive segmentation method with respect to other methods used for leukemia recognition. The results show that the proposed segmentation is suitable for images of different resolution, as well as for images with different features. The classification accuracy in most cases exceeds 90% using a multilayer Perceptron (MLP) neural network, Support Vector Machine (SVM), Random Forests (RF) and Bayes as classifiers.

Keywords: Color segmentation, supervised learning, leukemia.

1. Introducción

La leucemia es un tipo de cáncer que se encuentra dentro de los diez que provocan mas muertes por cáncer mundialmente [7]. El diagnóstico de este padecimiento es importante para el tratamiento oportuno de la enfermedad, sin embargo su costo es elevado y su disponibilidad limitada principalmente a zonas urbanas [10].

La leucemia se clasifica como aguda o crónica en función de la rapidez con que evoluciona, también se clasifica en Linfocítica y Mielocítica de acuerdo al tipo de células que son afectadas [2]. Una de las técnicas para el diagnóstico de leucemias agudas consiste en el análisis visual de las imágenes de frotis de médula ósea [16].

El análisis morfológico estas imágenes requiere generalmente de la observación de los frotis por personal especializado, lo cual es propenso a errores que pueden culminar en un mal diagnóstico.

El problema del reconocimiento de leucemia aguda a partir de imágenes ha sido tratado con diversas técnicas de visión por computadora y aprendizaje automático. Sin embargo, los principales inconvenientes de la mayoría de estas propuestas son (1) solo diferenciar imágenes de células sanas y de células leucémicas [4], (2) uso de datasets de imágenes pequeños (menos de 110 imágenes) [12] u obtenidos para uso exclusivo del método propuesto [3], (3) datasets que contienen solo una célula por imagen [1], entre otras. Además, en lo que refiere al estudio de subtipos de leucemia, existen mas propuestas dirigidas hacia las leucemias linfocíticas [12, 3, 1, 11].

Aunque paradigmas tradicionales [13, 4, 6] y automáticos [18, 12] para el análisis y reconocimiento de imágenes son usados para tratar el problema del reconocimiento de la leucemia, los enfoques automáticos poseen algunas limitaciones como la gran cantidad de imágenes para su entrenamiento y los recursos computacionales requeridos para un desempeño adecuado. Esto implica el uso adicional de técnicas de aumentación de datos, o de modelos de generativos para dar soporte al problema de la escasez de datos.

Con este marco de referencia, en este trabajo se propone un método de segmentación en color para el reconocimiento de cinco tipos de leucemia mieloide aguda utilizando técnicas clásicas de procesamiento digital de imágenes. En la propuesta se considera un conjunto de 1085 imágenes RGB que contienen una o más células en cada imagen, a diferencia de otras propuestas donde las imágenes contienen una sóla célula por imagen con un fondo libre de ruido significativo.

La siguiente sección describe los Materiales y Métodos. La sección 3 presenta la Metodología propuesta. La sección 4 muestra los Experimentos y Resultados. Finalmente, en la sección 5 se encuentran las conclusiones.

2. Materiales y métodos

2.1. Leucemia

La leucemia es un tipo de cáncer de la sangre que comienza en la médula ósea, el tejido blando que se encuentra en el centro de los huesos, donde se forman las células sanguíneas. Las leucemias agudas (LA) suponen la proliferación desordenada de una clona de células hematopoyéticas y son de avance rápido.

Tabla 1. Clasificación Franco-Americana-Británica (FAB1976).

Leucemias Agudas Mieloblásticas (LMA)	
M0	**Leucemia mieloblástica aguda indiferenciada**
M1	Leucemia mieloblástica aguda con maduración mínima
M2	**Leucemia mieloblástica aguda con maduración**
M3	**Leucemia promielocítica aguda (LPA)**
M4	Leucemia mielomonocítica aguda
M4 eos	**Leucemia mielomonocítica aguda con eosinofilia**
M5	**Leucemia monocítica aguda**
M6	Leucemia eritroide aguda
M7	Leucemia megacarioblástica aguda
Leucemias Agudas Linfoblásticas (LAL)	
L1	Leucemia linfoblástica típica
L2	Leucemia linfoblástica atípica
L3	Leucemia similar al linfoma de Burkit

La letalidad media anual de las LA es de tres a cinco casos por cada 100 000 habitantes y hay un notable aumento del padecimiento [17]. La clasificación de las LA están sujetas a una clasificación Franco-Americana-Británica (FAB), en la Tabla 1 se muestra dicha clasificación.

La clasificación FAB está basada en el análisis morfológico de cada uno de los subtipos de leucemia. Esta clasificación divide a las LA en dos tipos: Leucemias Agudas Mieloblásticas (LMA) y Leucemias Agudas Linfoblásticas (LAL) en función del tipo de célula de la cual proceden. En este trabajo se consideran cinco tipos de LMA, los cuales se denotan en negrita en la primera sección de la Tabla 1.

2.2. Segmentación y extracción de rasgos

La segmentación de imágenes es el proceso de seleccionar y agrupar los pixeles que poseen características visuales y numéricas similares entre sí. Existen diversos criterios de selección y agrupación donde características como el color, la textura o la forma son comúnmente usados.

La segmentación puede realizarse en escala de grises o en color, y ser binivel o multinivel de acuerdo a la cantidad de regiones obtenidas de dicha segmentación. Por otra parte, las técnicas para agrupar los pixeles pueden ser basadas en el histograma, por crecimiento y mezcla de regiones, basadas en bordes, entre otras [8].

La segmentación mediante el agrupamiento de pixeles en algún espacio de color de acuerdo se considera como una segmentación basada en pixel. La función de un espacio de color es proporcionar la especificación de color de una manera estandarizada y generalmente aceptada, es decir, un espacio de color es la especificación del sistema de coordenadas tridimensional y un subespacio del sistema, donde cada color está determinado por un punto. Algunos espacios de color usados con frecuencia son: RGB, YIQ, CMY, YCbCr y HSI.

Por otro lado, las características para describir un objeto o sus atributos pueden ser representadas por límites o propiedades externas y por métodos de representación estructural o propiedades internas.

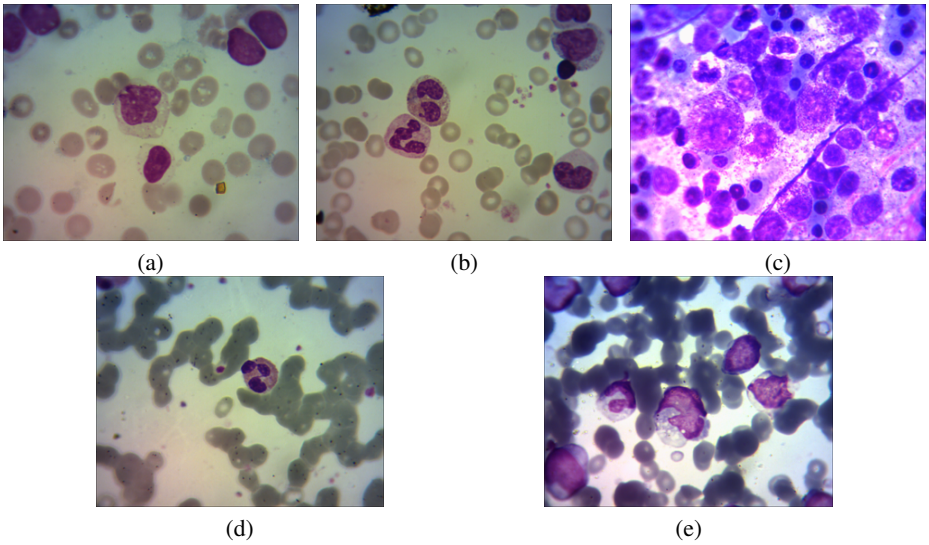


Fig. 1. Imágenes representativas de las clases (a) M0, (b) M2, (c) M3, (d) M4, (e) M5.

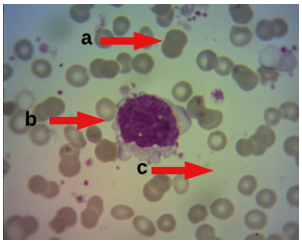


Fig. 2. Imagen de leucemia tipo M2. (a) Célula de no interés (b) Núcleo (c) Fondo.

También es deseable que las características sean invariantes a escala, rotación, y traslación. Particularmente, en el caso de las imágenes de frotis de leucemia propiedades como el color y la textura del núcleo de las células son relevantes para su reconocimiento.

La textura está relacionada con la distribución espacial de los tonos de gris también definida por la uniformidad, densidad, grosor, rugosidad, regularidad, intensidad y direccionalidad de medidas discretas del tono y de sus relaciones espaciales. Es posible definir la textura como un arreglo de píxeles cuya relación es la variación espacial de los tonos de grises.

Un enfoque clásico para el manejo de la textura es el basado en la matriz de coocurrencia de niveles de gris (GLCM), la cual es una matriz de frecuencias con la que un píxel con un nivel de gris (i) aparece en una relación de espacio específica con otro píxel de nivel de gris (j).

Las matrices de concurrencia son medidas de segundo orden porque consideran parejas de píxeles vecinos, separados una distancia δ y en un determinado ángulo θ.

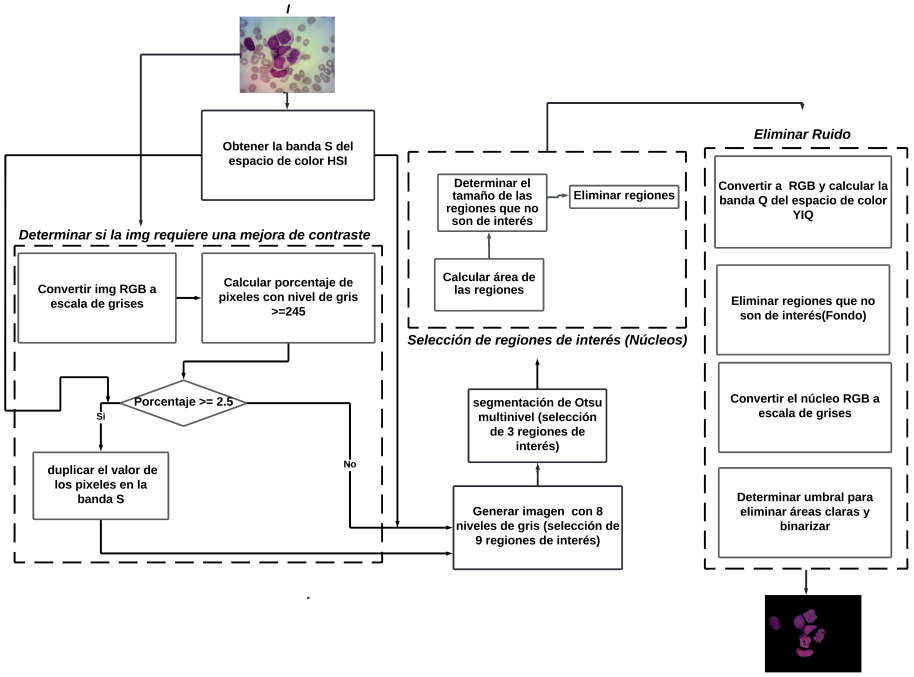


Fig. 3. Método de segmentación propuesto.

Por tanto, las matrices de coocurrencia pueden revelar ciertas propiedades sobre la distribución espacial de los grises en la textura de la imagen. La matriz de coocurrencia C_{ij} se calcula utilizando la Ecuación 1:

$$C_{ij} = \frac{P_{ij}}{\sum_{i,j=1}^{G} P_{ij}}, \tag{1}$$

donde P_{ij} representa el número de ocurrencias de los niveles de gris i y j dentro de una ventana, dado un cierto par(δ,θ); y G es el número cuantizado de niveles de gris.

Aunque existen varios descriptores de textura obtenidos a partir de la GLCM [9], algunos de ellos son invariantes a diversos cambios, como: uniformidad, entropía, disimilitud, contraste, diferencia inversa, momento de diferencia inversa y correlación [5].

3. Metodología

En esta sección se describe la metodología de esta propuesta. Primero se presenta el diseño del esquema de segmentación propuesto para el reconocimiento de los tipos de leucemia M0, M2, M3, M4 y M5. También se describe la extracción de características extraídas y la clasificación utilizando una red neuronal de tipo perceptron multicapa, una máquina de vectores de soporte, bosques aleatorios y Bayes, respectivamente.

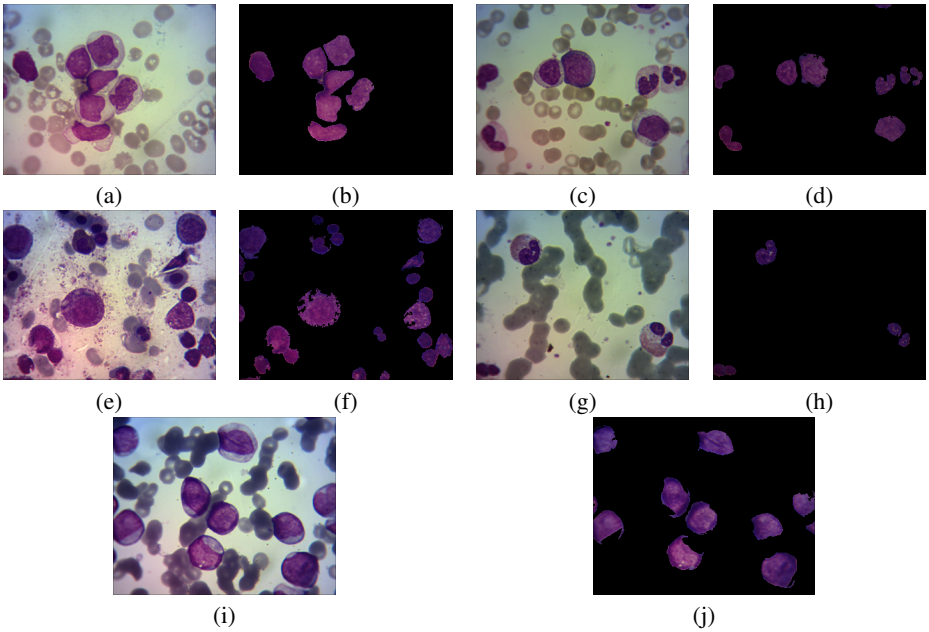


Fig. 4. Ejemplos de la base de datos: (a), (c), (e), (g), and (i) son imágenes de las clases M0, M2, M3, M4, and M5, respectivamente. Las imágenes en (b), (d), (f), (h), y (j) son de los núcleos segmentados de cada imagen.

3.1. Segmentación por color

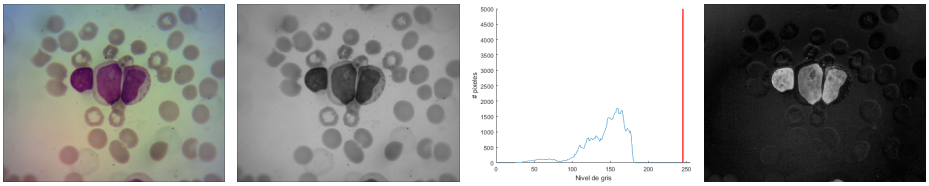
El método de segmentación propuesto se enfoca en extraer el núcleo de las células dado que presenta características relevantes para el reconocimiento de las clases de leucemia: M0, M2, M3, M4 y M5.

El método consta de 3 fases. En la primera se recibe como entrada una imagen a color y se evalúa si esta imagen requiere un preprocesamiento para eliminar el exceso de brillo. Cada una de las clases de leucemia presenta distintos rasgos, como se puede ver en la Figura 1.

También es importante notar que la iluminación es variable en las imágenes de algunas clases de leucemia, por ejemplo, en la Figura 1 (e) se muestra una imagen de la clase M5 donde se aprecia una mayor cantidad de brillo con relación al resto de las imágenes.

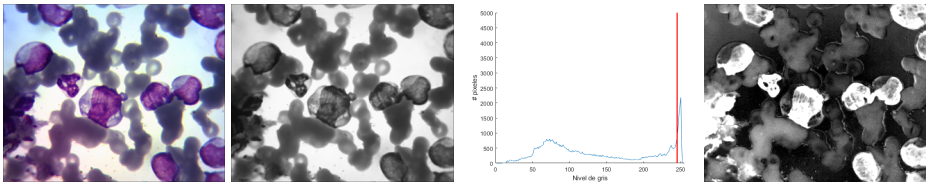
En la segunda fase se utiliza la imagen de la banda S del espacio de color HSI y se realiza una segmentación en dos pasos, primero generando una imagen con 8 niveles de gris y posteriormente aplicando una segmentación multinivel (usando 2 umbrales) con el método de Otsu [15]. Como resultado se obtienen 3 regiones de interés: el núcleo, células que no son de interés y el fondo. Estas regiones se muestran en la Figura 2.

En la tercera fase se aplica un post-procesamiento para eliminar regiones de la segmentación que no corresponden al núcleo de la célula. Esta fase se compone de dos etapas, en la primera se eliminan regiones pequeñas que tienen características similares al núcleo, pero que no forman parte de este.



(a) Img RGB de clase M0 (b) Escala de grises. (c) Histograma. (d) Banda S de clase M0.

Fig. 5. Imagen de la clase M0 que no necesita la mejora de contraste.



(a) Img RGB de clase M5. (b) Escala de grises. (c) Histograma. (d) Banda S de clase M5.

Fig. 6. Imagen de la clase M5 que necesita mejora de contraste.

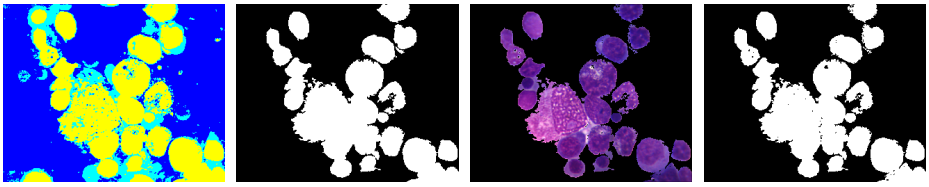
En la siguiente etapa se utiliza el color como criterio para la eliminación de regiones que no cumplen con las características de color de un núcleo, el cual típicamente tiene una coloración en tonos morados. En la Figura 3 se presenta el diagrama del método de segmentación propuesto.

3.2. Extracción de características

A partir de la imagen segmentada, se extraen características de textura y color para el reconocimiento de las clases de leucemia M0, M2, M3, M4 y M5. En este trabajo se extraen los descriptores de 7 característicasde textura basadas en la GLCM: uniformidad, entropía, disimilitud, contraste, diferencia inversa, momento de diferencia inversa, correlación.

Por otra parte, el color es útil para identificar a los núcleos celulares debido a que estos sobresalen en la imagen a causa de su coloración en distintos tonos morados. En diferentes espacios de color algunas bandas son relevantes para la identificación de los núcleos celulares. Para seleccionar la información de color útil en la imagen, se evalúan cuatro espacios de color: YIQ, YCbCr, HSI, Lab. Cada uno de estos espacios de color tienen una banda en la que resaltan ciertos rasgos del color en la imagen. Las bandas seleccionadas en este trabajo son: Q de YIQ, Cb de YCbCr, H de HSI, b de Lab. Posteriormente, se calcula el promedio de las bandas seleccionadas para obtener cuatro descriptores de color.

Las siete características de textura y las cuatro de color descritas previamente se extraen de cada imagen para la clasificación utilizando como clasificadores una red MLP con una arquitectura de cuatro capas ocultas cuya estructura es [40 20 20 40], y 55 épocas. Estos parámetros fueron obtenidos por experimentación. También se probó con un SVM, RF y Bayes.



(a) (b) (c) (d)

Fig. 7. Etapas de la selección de regiones de interés de una imagen de la clase M3.

RF se inicializó con 500 árboles y la clasificación se realiza por mayoría de votos. La validación del aprendizaje es a través de una validación cruzada en 5 particiones.

4. Experimentos y resultados

En esta sección se presentan los experimentos y resultados del esquema de segmentación propuesto, así como los resultados de clasificación de los tipos de leucemia M0, M2, M3, M4, M5. Primero se describe el conjunto de imágenes, así como el hardware y software utilizados.

Posteriormente, se presentan los experimentos realizados en la fase segmentación de núcleos celulares, así como una evaluación comparativa del método de segmentación propuesto. Finalmente, se describe el proceso para la clasificación de imágenes, y los resultados obtenidos.

4.1. Base de datos, Harware y Software

El conjunto de datos utilizado se compone de 1085 imágenes de frotis de médula ósea de cinco subtipos de leucemia: M0, M2, M3, M4 y M5, 217 de cada clase. Las imágenes fueron tomadas de [14], son a color y están en formato PNG con resolución de 1280 × 1024 píxeles.

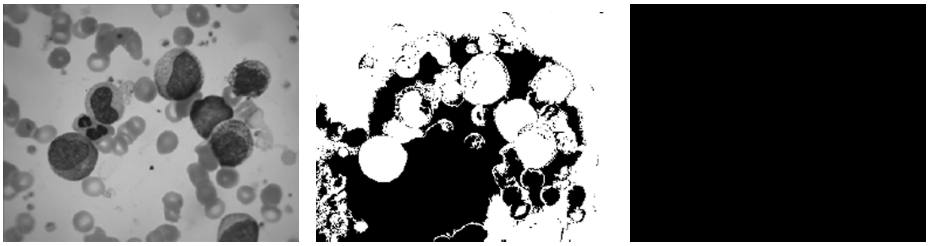
Para los experimentos se construyeron tres conjuntos de imágenes con diferente resolución cada uno: 128×160, 256×320, y 512×640. El redimensionamiento se hizo con una iterpolación bicúbica.

Las Figuras 1 y 4 muestran algunos ejemplos de estas imágenes, donde se puede observar que la iluminación es variable y en algunos casos puede ocasionar que regiones del fondo sean consideradas como células de interés, o viceversa.

Además en algunas clases de leucemia como M3 existen aglomeraciones de células y elementos del fondo con características similares, representando un reto para la segmentación. Los experimentos se ejecutaron en un equipo AMD Ryzen 5, 12.0 GB RAM, Windows 11 y MATLAB.

4.2. Segmentación en color

En la primera etapa de la segmentación se determina si la imagen de entrada requiere una mejora de contraste, para lo cual se siguen los pasos mostrados en el diagrama de la Figura 3.



(a) Imagen en escala de gris. (b) Segmentación Li et al. [13]. (c) Segmentación Ghane et al. [6].

Fig. 8. Resultados de la segementación de una imagen de la clase M3.

Tabla 2. Tiempo (minutos) de segmentación.

Dataset	Cant.Imgs	Propuesta	Li	Ghane
ALLIDB2	260	0.1390	01:36.4	01:05
Miktec	674	1.3238	07:26	03:43
Proyecto	1085	2.2887	06:46	04:56.6

En las Figuras 5 y 6 se presentan los resultados de la primera fase de segmentación aplicados a imágenes que si requieren mejora de contraste, así como a imágenes que no lo requieren. Esto se observa en los histogramas en los cuales el umbral indicado con una línea roja determina si la mejora es requerida.

Finalmente, en la tercera fase de la segmentación se lleva a cabo la selección de regiones de interés en donde se procesa la imagen obtenida en la segunda fase de la segmentación. En la Figura 7 es posible observar cada una de las etapas del post-procesamiento, la primera columna representa la imagen generada por Otsu, la segunda columna representa la binarización y selección de regiones, en la tercera columna se observa la imagen en RGB y finalmente en la cuarta columna se observa el resultado final de la segmentación.

El método se probó con las imágenes de cada conjunto propuesto en la sección 4.1 cuya resolución es distinta, mostrando que la segmentación es adecuada en todos los casos siendo invariante a los cambios de escala.

Para evaluar el método de segmentación propuesto se compararó con dos métodos de segmentación, Li et.al. [13] y Ghane et.al. [6]. En los experimentos se probó cada método con los datasets de los otros trabajos, respectivamente. En todos los casos la segmentación propuesta obtuvo resultados favorables, a diferencia de los métodos usados para la comparación, lo cual se muestra en la Figura 8 donde se presenta el resultado de la segmentación de estos métodos.

Considerando las características de los datasets, es posible notar que el método de Ghane et.al. [6] no logró aislar a las regiones de interés, lo cual es una desventaja con respecto a este trabajo. También se realizó una medición del tiempo para cada método de segmentación para comparar el desempeño del método propuesto. Los resultados se muestran en la Tabla 2, donde se puede observar que la segmentación propuesta tiene un mejor desempeño con relación a los métodos comparados.

Tabla 3. Resultados de la precisión de clasificación de prueba con la normalización $Z - score$.

Num. corrida	Máquina de vectores de soporte	Perceptron multicapa	Bosques Aleatorios	Bayes
1	93.53	**96.00**	93.53	76.30
2	93.23	94.46	92.30	77.23
3	93.23	95.69	92.30	77.53
4	93.53	95.69	92.30	77.23
5	93.84	94.15	93.53	76.92
6	93.84	95.69	91.69	76.61
7	92.92	94.15	92.61	76.92
8	92.61	92.92	94.15	76.61
9	92.61	95.69	94.15	76.30
10	92.92	93.54	93.84	76.61
11	93.53	94.46	92.30	77.23
12	92.92	95.08	93.84	76.30
13	93.84	94.46	92.92	76.61
14	93.53	95.08	92.61	77.23
15	94.15	94.77	92.92	76.00
16	92.92	94.15	93.84	76.61
17	93.84	94.15	93.23	76.61
18	92.30	94.46	92.30	76.61
19	93.23	92.92	93.53	76.00
20	94.15	95.08	93.84	76.61
21	93.23	92.92	92.30	76.92
22	93.84	94.46	92.61	77.23
23	93.53	92.92	94.76	76.30
24	93.53	94.77	92.61	76.92
25	93.23	95.08	92.61	77.23
26	93.53	94.77	93.23	77.23
27	93.23	95.69	92.92	77.23
28	93.53	95.15	92.92	76.61
29	93.23	93.54	93.53	76.30
30	93.23	94.15	93.53	77.23
Media	93.36	94.54	93.10	76.78
σ	$\pm$0,45	$\pm$0,85	$\pm$0,73	$\pm$0,42

4.3. Clasificación

Para la clasificación se utilizaron las 11 características descritas en la sección previa. Las características texturales fueron obtenidas a partir de la GLCM utilizando 8 niveles de gris, $\delta = 1$, and $\theta = \{0, 45, 90, 135\}$. De esta manera se calculó la uniformidad, entropía, disimilitud, contraste, diferencia inversa, momento de diferencia inversa, y la correlación.

Estas características se obtuvieron para cada valor de θ. Después, se promediaron las cuatro orientaciones para cada característica, obteniendo así 7 características para cada imagen. Las 4 características de color corresponden al promedio de las bandas de color Q de YIQ, Cb de YCbCr, H de HSI, y b de Lab.

	1	2	3	4	5	
1	60	4	0	1	0	92.30
2	1	62	0	1	1	95.38
3	0	1	63	0	1	96.92
4	0	2	0	63	0	96,92
5	0	1	0	0	64	98.46
						96

Fig. 9. Matriz de confusión de la corrida con mayor precisión de clasificación.

Previo al entrenamiento de los clasificadores MLP, SVM, RF y Bayes, la matriz de características se normalizó con la función Z-score. La arquitectura de la red MLP esta descrita en la sección 3.2. En todos los casos se realizaron 30 ejecuciones, donde el conjunto de datos se dividió en 70 % para entrenamiento y 30 % para prueba. Los resultados se muestran en la Tabla 3.

La matriz de confusión para el mejor de los casos utilizando una MLP se muestra en la Figura 9, donde se puede observar que la clase mejor reconocida es la M5, mientras que M0 alcanzó una precisión de clasificación más baja, con un 92,30 %. En los clasificadores SVM y RF el desempeño en general es ligeramente menor, mientras que con Bayes los resultados son más bajos debido probablemente a la existencia de algún tipo de dependencia de las características.

5. Conclusiones

En este trabajo se presentó un método de segmentación de imágenes para el reconocimiento de cinco tipos de leucemia aguda. Cabe señalar que el problema de reconocimiento de subtipos de leucemia mieloide aguda ha sido poco estudiado. Además, el método de segmentación es robusto con respecto a la resolución de las imágenes, logrando buenos resultados para distintos tamaños de imágenes.

En la comparación con dos métodos de segmentación existentes en la literatura, ha demostrado su utilidad al segmentar adecuadamente otros conjuntos de imágenes. Por otro lado, la precisión en la clasificación de las 5 clases de leucemia tuvo un desempeño adecuado, alcanzando un índice de precisión de clasificación superior al 90 % en la mayoría de las pruebas realizadas.

De esta manera, con el desarrollo de este trabajo se presenta un marco de referencia útil para como apoyo en el diagnóstico médico enfocado al reconocimiento de leucemia mieloide aguda a partir de imágenes de frotis de médula ósea.

El método propuesto es susceptible de mejora para lograr el reconocimiento de cualquier tipo de leucemia aguda a un bajo costo computacional y un buen desempeño. Otro uso es en el área educativa, como herramienta de autoevaluación en el reconocimiento visual de leucemias agudas para estudiantes que se forman en el área clínica.

Algunos aspectos por evaluar a futuro son la influencia de la calidad de las imágenes en el método de segmentación propuesto, lo cual implica ampliar los experimentos utilizando nuevos conjuntos de imágenes. De igual manera, el análisis de las características de las imágenes para determinar cuáles son los métodos de aprendizaje automático más adecuados para un mejor desempeño en el problema del reconocimiento de tipos de leucemia aguda, es una área de oportunidad en este trabajo.

Agradecimientos Los autores agradeceden a la Universidad Autónoma de Tlaxcala y a la Red de Inteligencia Computacional Aplicada (RedICA-CONACYT) por las facilidades para el desarrollo de este trabajo.

Referencias

1. Alagu, S., Bhoopathy, K.: Acute lymphoblastic leukemia diagnosis in microscopic blood smear images using texture features and SVM classifier. Alliance International Conference on Artificial Intelligence and Machine Learning, pp. 175–186 (2019)
2. American Cancer Society: Leucemia. (2023) https://www.cancer.org/es/cancer/leucemia.html
3. Bodzas, A., Kodytek, P., Zidek, J.: Automated detection of acute lymphoblastic leukemia from microscopic images based on human visual perception. Frontiers in Bioengineering and Biotechnology, vol. 8 (2020) doi: 10.3389/fbioe.2020.01005
4. Boldú, L., Merino, A., Alférez, S., Molina-Borrás, A., Acevedo, L. A., Rodellar, J.: Automatic recognition of different types of acute leukaemia in peripheral blood by image analysis. Journal of Clinical Pathology, vol. 72, no. 11 (2019) doi: 10.1136/jclinpath-2019-205949
5. Clausi, D. A.: An analysis of co-occurrence texture statistics as a function of grey level quantization. Canadian Journal of Remote Sensing, vol. 28, no. 1, pp. 45–62 (2002) doi: 10.5589/m02-004
6. Ghane, N., Vard, A., Talebi, A., Nematollahy, P.: Classification of chronic myeloid leukemia cell subtypes based on microscopic image analysis. EXCLI Journal, vol. 18, pp. 382–404 (2019) doi: 10.17179/excli2019-1292
7. Global Center Observatory, T.: International Agency for Research on Cancer. (2023) https://gco.iarc.fr/today/data/factsheets/cancers/33-Hodgkin-lymphoma-fact-sheet.pdf
8. Gonzalez, R. C., Woods, R. E., Eddins, S. L.: Digital image processing using MATLAB. Pearson (2010)
9. Haralick, R. M., Shamugam, K., Dinstein, I. H.: Textural features for image classification. IEEE Transactions on Systems, Man and Cybernetics, vol. 3, no. 6, pp. 610–621 (1973) doi: 10.1109/TSMC.1973.4309314
10. Howard, S., Zaidi, A., Cao, X., Weil, O., Bierre, B., Patte, C., Samudio, A., Haddad, L., Lam, C., Moreira, C., Pereira, A., Harif, M., Hessissen, L., Choudhury, S., Fu, L., Caniza, M. A., Lecciones, J., Traore, F., Ribeiro, R., Gagnepain-Lacheteau, A.: The my child matters programme: Effect of public–private partnerships on paediatric cancer care in low-income

and middle-income countries. The Lancet Oncology, vol. 19, no. 5, pp. e252–e266 (2018) doi: 10.1016/S1470-2045(18)30123-2
11. Khandekar, R., Shastry, P., Jaishankar, S., Faust, O., Sampathila, N.: Automated blast cell detection for acute lymphoblastic leukemia diagnosis. Biomedical Signal Processing and Control, vol. 68 (2021) doi: 10.1016/j.bspc.2021.102690
12. Kumar, J. K., Sekhar, D. H.: Nucleus and cytoplasm–based segmentation and actor-critic neural network for acute lymphocytic leukaemia detection in single cell blood smear images. Medical & Biological Engineering & Computing, vol. 58, no. 1, pp. 171–186 (2020) doi: 10.1007/s11517-019-02071-1
13. Li, Y., Zhu, R., Mi, L., Cao, Y., Yao, D.: Segmentation of white blood cell from acute lymphoblastic leukemia images using dual-threshold method. Computational and Mathematical Methods in Medicine, vol. 2016, pp. 1–12 (2016) doi: 10.1155/2016/9514707
14. Ochoa-Montiel, R., Sossa, H., Olague, G.: Improving leukemia image classification by extracting and transferring knowledge by evolutionary vision. Research in Computing Sicence, vol. 150, no. 11, pp. 167–176 (2021)
15. Otsu, N.: A threshold selection method from gray-level histograms. IEEE Transactions on Systems, Man, and Cybernetics, vol. 9, no. 1, pp. 62–66 (1979) doi: 10.1109/tsmc.1979.4310076
16. Rodak, B. F., Carr, J. H.: Clinical hematology atlas. Saunders (2016)
17. Ruiz-Argüelles, G. J., Ruiz-Delgado, G. J.: Fundamentos de hematología. Editorial Médica Panamericana (2014)
18. Vogado, L., Veras, R., Aires, K., Araújo, F., Silva, R., Ponti, M., Tavares, J. M.: Diagnosis of leukaemia in blood slides based on a fine-tuned and highly generalisable deep learning model. Sensors, vol. 21, no. 1 (2021) doi: 10.3390/s21092989

ISSN 1870-4069

Estudio del uso de GAN para el balanceo de datos en el conjunto BreakHis y los efectos en la clasificación de tumores de mama

Alfredo Gutiérrez-Alfaro[1], Angel E. Rosales-Morales[2], Andrés Espinal[1], Manuel Ornelas-Rodríguez[2], Marco Sotelo-Figueroa[1], Horacio Rostro-Gonzalez[3]

[1] Universidad de Guanajuato, Departamento de Estudios Organizacionales, División de Ciencias Económico Administrativas, México

[2] Instituo Tecnológico de León, Tecnológico Nacional de México, México

[3] Universidad de Guanajuato, Departamento de Electrónica, División de Ingenierías, México

{a.gutierrezalfaro, aespinal}@ugto.mx, m22240870@leon.tecnm.mx

Resumen. El conjunto BreakHis es una base de datos de imágenes histopatológicas de tejido de mamario, cuyo propósito es permitir el desarrollo de modelos para la clasificación automática de tumores (malignos y benignos). Las clases principales del conjunto están distribuidas en $5,429$ y $2,480$ imágenes para las clases maligno y benigno, respectivamente. El desbalance de datos en las clases, un problema común en el análisis de imágenes médicas, puede impactar negativamente en el desempeño de modelos de aprendizaje máquina. En este trabajo, se hace uso un modelo generativo (GAN) para el balanceo de imágenes. Posteriormente, se comprueba el comportamiento y desempeño de clasificadores basados en aprendizaje profundo usando los conjuntos no balanceado y balanceado. Los resultados obtenidos son comparados usando cuatro diferentes métricas.

Palabras clave: Conjunto BreakHis, balanceo de datos, clasificación, aumento de datos, GAN.

Study of GAN Usage for Data Balancing in BreakHis Dataset and the Effects in Breast Tumor Classification

Abstract. The BreakHis set is a histopathologic images database of mammary tissue, which purpose is to enable the development of models for automatic

classification of tumors (malignant and benign). The main classes of the set are distributed in $5,429$ and $2,480$ images for the malignant and benign classes, respectively. Data imbalance in classes, a common problem in medical image analysis, can negatively impact the performance of machine learning models. In this work, a generative model (GAN) is used for image balancing. Subsequently, the behavior and performance of deep learning-based classifiers are tested using the unbalanced and the balanced datasets. The obtained results are compared using four different metrics.

Keywords: BreakHis dataset, data balancing, classification, data augmentation, GAN.

1. Introducción

En la actualidad, los casos de cáncer de mama han ido en aumento, no obstante, gracias a los avances en los métodos de detección temprana, la tasa de mortalidad ha decrementado considerablemente. De los diversos métdos para el diagnóstico de cáncer de mama, la examinación de biopsias histopatológicas suele ser infalible; este es llevado a cabo por patólogos al analizar láminas de tejido mamario con varios niveles de aumento microscópicos para resaltar las zonas de interés; sin embargo, la interpretación de los patológos es susceptible a errores humanos por factores como cansancio, distracción, etc [3].

En el área de medicina cada vez es más frecuente la adopción de sistemas para el diagnóstico asistido por computadora con el objetivo de disminuir los riesgos de un mal dianóstico. En particular, para el caso de detección de cáncer de mama a partir de imágenes histopatológicas existe una base de datos llamada BreakHis [24], la cuál recopila imágenes de biopsias con diferentes niveles de aumento microscópico de tumores de mama benignos y malignos.

Esta base de datos es de uso público con el objetivo de proveer un marco de referencia y para fines de investigación; sin embargo, a pesar de la riqueza de su contenido clínico, BreakHis aún posee anormalidades comunes en todos los conjuntos de datos médicos debido a la naturaleza de la enfermedad y limitaciones en las técnicas de adquisición de datos médicos [3].

El conjunto BreakHis tiene una proporción de imágenes del $68{,}6\,\%$ y $31{,}4\,\%$ para las clases maligno y benigno, respectivamente; este desbalance de datos entre las clases es una de las anomalías que se suelen presentar en los conjuntos de datos médicos.

El desbalanceo de datos puede tener efectos no deseados en un sistema de diagnóstico asistido por computadora, como sesgar su capacidad discriminativa hacia las clases más representativa en el conjunto de datos, es decir, la clase menos representativa puede ser considerada como datos atípicos o ruido [3, 22].

Para lidiar con el problema del desbalance de clases, existen diferentes técnicas o métodos. En la literatura se pueden encontrar técnicas como duplicar imágenes pertenecientes a la clase minoritaria, aplicar transformaciones a imágenes existentes (rotaciones, cambios de color o añadir ruido), combinar imágenes de la clase minoritaria para crear nuevas instancias (SMOTE por sus siglas en inglés -Synthetic Minority Oversampling Technique-) [22].

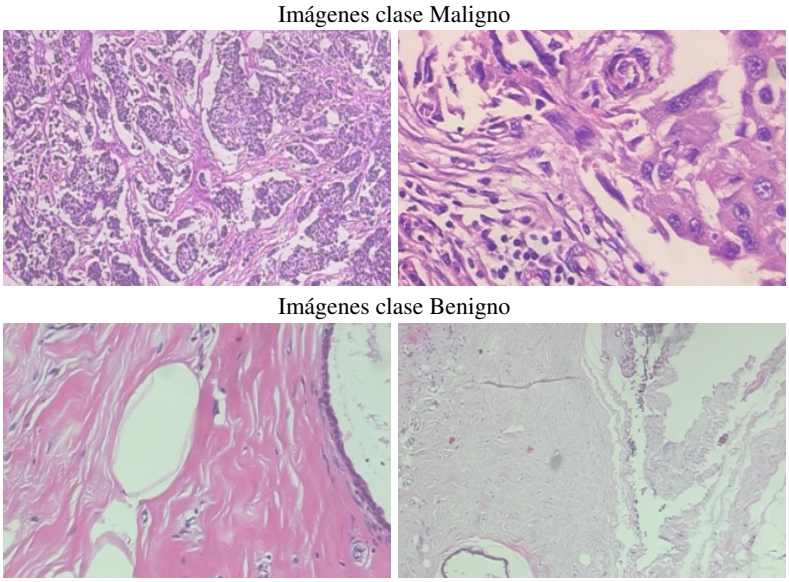


Fig. 1. Muestra de las imágenes del conjunto BreakHis.

Un método de los más recurridos últimamente es el aumento de datos artificiales mediante modelos generativos, en particular usando GAN [9] (por sus siglas en inglés -Generative Adversarial Networks-); las cuales son tratadas dentro de este artículo. Las GAN han sido utilizadas exitosamente en diversos ámbitos, como en el médico [26, 23]; por ejemplo, en [6] se logró mejorar los valores de métricas como la sensibilidad en la clasificación de lesiones en hígado.

No obstante, el uso de GAN en aplicaciones médicas para aumentar la cantidad de datos y mejorar los modelos de clasificación debe ser tratado con cuidado, ya que, los datos generados podrían añadir ruido a los modelos, ocasionando el efecto contrario al de mejorar su desempeño [4].

En el presente artículo se estudia el efecto de desarrollar modelos de clasificación usando el conjunto BreakHis original y una versión balanceada del mismo a través de una GAN; es decir, solo se agregarán imágenes artificiales de la clase benigna. Se realizará una experimentación exhaustiva del desempeño de modelos entrenados con ambos conjuntos de datos, midiendo métricas de exactitud, presición, recuperacón y especificidad, con las cuales se observará el comportamiento de los clasificadores basados en aprendizaje profundo.

La estructura del artículo es la siguiente. La sección 2 describe las características del conjunto BreakHis. El modelo de aprendizaje profundo que se utiliza para la tarea de clasificación, CNN (por sus siglas en inglés -Convolutional Neural Network-), se detalla en la sección 3.

El modelo generativo, GAN, para lograr el balanceo de clases es explicada en la sección 4. En la sección 5 se describe la experimentación realizada y los resultados obtenidos. Finalmente, las conclusiones y trabajo a futuro se mencionan en la sección 6.

Tabla 1. Distribución de las imágenes del conjunto BreakHis.

Aumento	Maligno	Benigno	Total
$40\times$	$1,370$	625	$1,995$
$100\times$	$1,437$	644	$2,081$
$200\times$	$1,390$	623	$2,013$
$400\times$	$1,232$	588	$1,820$
Total	**5,429**	**2,480**	**7,909**

2. Conjunto de datos BreakHis

El conjunto de datos BreakHis [24] es una base de datos de imágenes histopatológicas de tejidos mamarios, la cual fue recopilada con el propósido de ser un marco de referencia para el desarrollo de sistemas de clasificación para el diagnóstico asistido por computadora.

El conjunto de datos está dividido principalmente en dos clases de tumores, malignos y benignos; la Figura 1 muestra algunas de las imágenes contenidas en BreakHis para estas clases. Cada clase principal tiene cuatro subclases, siendo carcinoma ductal, carcinoma lobulillar, carcinoma mucinoso y carcinoma papilar para la clase maligno y adenosis, fibroadenoma, tumores filodes y adenoma tubular para la clase benigno; sin embargo, en este trabajo se limitará a trabajar únicamente con las clases principales.

El conjunto de datos está organizado en cuatro niveles de aumento. La Tabla 1 muestra la distribución de datos por nivel de aumento y clases. Enfocándose en las clases principales se puede observar el conjunto BreakHis sufre de desbalance de datos; siendo así que la clase predominante es la de tumores malignos con un total de $5,429$ imágenes y la clase con menor presencia la de tumores benignos con $2,480$ imágenes.

Dicho desbalance de datos puede repercutir en la generación de modelos de clasificación con comportamientos no deseados, como el hecho de que la clase menos representativa sea vista como casos raros por el modelo y llegue a ser clasificada incorrectamente [22].

Existen trabajos que han usado modelos generativos para lograr un balance de los datos en las clases y efectuar la tarea de clasificación de tumores reportando un incremento en la exactitud de los modelos [15, 7]; sin embargo, a diferencia de los trabajos previamente mencionados, en esta investigación se busca conocer el efecto que tiene el agregar datos artificiales mediante otras métricas como la precisión y recuperación que se enfocan en la clase positiva (maligno) y que no ha sido modificada y la especificidad que se enfoca en la clase negativa (benigno) la cual será balanceada mediante el uso de una GAN.

3. CNN: Convolutional Neural Network

Las CNN son un tipo de arquitectura de redes neuronales ampliamente usadas en tareas de visión por computadora, reconocimiento y clasificación de imágenes [25]; siendo la última el tipo de tarea que se resolverán en este trabajo. La Fig. 2 muestra un bosquejo general de la arquitectura y funcionamiento de una CNN.

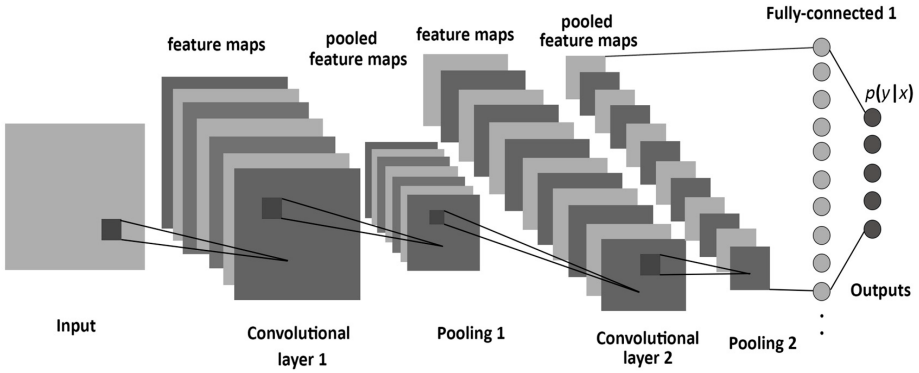


Fig. 2. Arquitectura de CNN para clasificación [1].

Estas redes trabajan mediante un proceso conocido como convolución, donde se aplican filtros para identificar y extraer patrones y características de las imágenes. Estos filtros son matrices de pesos que se aplican a lo largo de cada posición de la imagen, multiplicando los elementos de la imagen con los valores del filtro y sumando los resultados.

Este proceso se repite para cada posición dentro de la imagen, generando una nueva matriz de caracterśticas; conocido como mapa de características. A medida que se aplican capas convolucionales sucesivas, la red va aprendiendo representaciones cada vez más complejas de la imagen. Las capas posteriores suelen utilizar filtros más grandes para capturar características globales, como formas grandes o patrones complejos [8, 18].

Es común que una capa convolucional sea precedida por una capa conocida como pooling, la cuál sirve para extraer las características más relevantes del mapa generado por la capa convolucional; como resultado, el pooling proporciona un mapa de características importantes más pequeño. Finalmente, en la última capa se obtienen las probabilidades para cada clase y con base a ellas se realiza la clasificación [8].

4. GAN: Generative Adversarial Network

Las GAN [10] pertenecen a los modelos generativos, estas están formadas por dos redes que compiten entre sí; un generador y un discriminador. El generador tiene la tarea de tratar de engañar al discriminador, sintetizando información a partir de ruido aleatorio; es decir, el generador debe crear muestras que son lo suficientemente parecidas a las reales como para engañar al discriminador.

Mientras que el discriminador tiene la tarea de identificar si los datos son reales o falsos; es decir, creados artificialmente por el generador. Durante el proceso de entrenamiento de una GAN, se usa como función de pérdida la Ecuación 1, propuesta originalmente en [10]:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{data}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z)))], \quad (1)$$

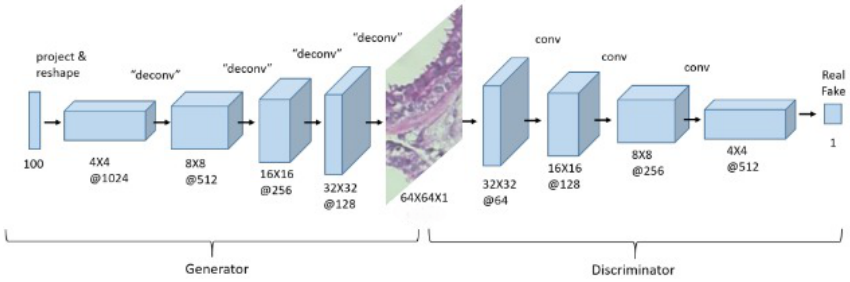


Fig. 3. Arquitectura de la DCGAN. (Imagen modificada de [6]).

donde $D(x)$ representa la estimación del discriminador para una muestra que proviene de una distribución de datos reales, mientras que $D(G(z))$ es la estimación del discriminador de que una muestra generada por $G(z)$ es real.

El generador busca minimizar la función, mientras que el discriminador busca maximizarla. La primera parte de la función $\mathbb{E}x \sim pdata(x)[\log D(x)]$ corresponde a la esperanza matemática del logaritmo de la probabilidad de que el valor x sea real. La segunda parte, $\mathbb{E}z \sim pz(z)[\log(1 - D(G(z)))]$, corresponde al valor esperado del logaritmo de la probabilidad asignada a la muestra generada $G(z)$ de que no sea real.

Las GAN se han utilizado en una amplia variedad de tareas y han dado lugar a múltiples variantes para tareas específicas [12, 17]. En el campo de la visión por computadora, se han utilizado principalmente para generar imágenes y tienen aplicaciones como la traducción de imagen a imagen [16], la generación de rostros [19], la generación de señales, entre otras que no están necesariamente relacionadas con imágenes [17].

En el ámbito médico, las GAN han tenido diversos usos [26], como la generación de imágenes, la reconstrucción, la segmentación, entre otras. La generación de imágenes es una de las categorías más populares [20], ya que permite generar nuevas imágenes en comparación con otras técnicas de aumento de datos.

Una de las múltiples aplicaciones de las GAN consiste en la generación de imágenes para balancear conjuntos de datos [22], lo cual puede ser de gran utilidad para compensar la falta de datos específicos de una clase en particular. Este enfoque puede ser relevante en conjuntos de datos médicos, donde la obtención de nueva información puede resultar complicada.

4.1. DCGAN: Deep Convolutional Generative Adversarial Network

Las DCGAN (por sus siglas en ingés -Deep Convolutional Generative Adversarial Networks-) son una modificación de la GAN original, la cual ha sido utilizada en arquitecturas actuales [21]. En lugar de emplear una capa completamente conectada, la DCGAN utiliza capas convolucionales, lo que permite la generación mejores imágenes.

En la Figura 3, se muestra en la parte izquierda la arquitectura del generador utilizado en la DCGAN, comenzando con una entrada de tamaño 100, que corresponde al tamaño del vector de ruido, se utilizan capas deconvolucionales para aumentar el tamaño de las salidas en cada capa hasta obtener una salida dé $64 \times 64 \times 3$ [11, 21].

Layer (type)	Output Shape	Param #
ConvTranspose2d-1	[-1, 512, 4, 4]	819,200
BatchNorm2d-2	[-1, 512, 4, 4]	1,024
LeakyReLU-3	[-1, 512, 4, 4]	0
ConvTranspose2d-4	[-1, 256, 8, 8]	2,097,152
BatchNorm2d-5	[-1, 256, 8, 8]	512
LeakyReLU-6	[-1, 256, 8, 8]	0
ConvTranspose2d-7	[-1, 128, 16, 16]	524,288
BatchNorm2d-8	[-1, 128, 16, 16]	256
LeakyReLU-9	[-1, 128, 16, 16]	0
ConvTranspose2d-10	[-1, 64, 32, 32]	131,072
BatchNorm2d-11	[-1, 64, 32, 32]	128
LeakyReLU-12	[-1, 64, 32, 32]	0
ConvTranspose2d-13	[-1, 3, 64, 64]	3,072
Tanh-14	[-1, 3, 64, 64]	0

Total params: 3,576,704
Trainable params: 3,576,704
Non-trainable params: 0

(a) Arquitectura del generador

Layer (type)	Output Shape	Param #
Conv2d-1	[-1, 64, 32, 32]	3,072
LeakyReLU-2	[-1, 64, 32, 32]	0
Conv2d-3	[-1, 128, 16, 16]	131,072
BatchNorm2d-4	[-1, 128, 16, 16]	256
LeakyReLU-5	[-1, 128, 16, 16]	0
Conv2d-6	[-1, 256, 8, 8]	524,288
BatchNorm2d-7	[-1, 256, 8, 8]	512
LeakyReLU-8	[-1, 256, 8, 8]	0
Conv2d-9	[-1, 512, 4, 4]	2,097,152
BatchNorm2d-10	[-1, 512, 4, 4]	1,024
LeakyReLU-11	[-1, 512, 4, 4]	0
Conv2d-12	[-1, 1, 1, 1]	8,192
Sigmoid-13	[-1, 1, 1, 1]	0

Total params: 2,765,568
Trainable params: 2,765,568
Non-trainable params: 0

(b) Arquitectura del discriminador

Fig. 4. Arquitecturas de los modelos de la GAN.

El discriminador, mostrado en la parte derecha de la Figura 3, tiene una arquitectura similar, pero de forma inversa, ya que en este caso, el discriminador recibe una imagen de tamaño $64 \times 64 \times 3$ y de salida tiene la predicción de sí dicha entrada es real o falsa; esta red es básicamente una CNN como se explicón en la Sección 3.

5. Experimentos y resultados

En esta sección se describe el procedimiento experimental y los resultados obtenidos. El desarrollo de los modelos fue realizado usando el lenguaje de programación Python y la librería de PyTorch.

5.1. Generación de imágenes para balanceo de datos

El balanceo de datos se enfocará a igualar la cantidad de imágenes de la clase benigno a las que conforman la clase maligno. Dicho proceso consistirá en la creación de $2,949$ imágenes artificiales que asemejen lo más posible las imágenes reales contenidas en la clase benigno mediante el entrenamiento y uso de una DCGAN.

La DCGAN usada está basada principalmente en el modelo publicado en [21] con modificaciones propuestas en [9]. Las Figuras 4a y Figura 4b muestran la configuración de los modelos que conforman la DCGAN; generador y discriminador respectivamente.

Para el entrenamiento de la DCGAN, será necesario preprocesar las imágenes, esto incluye su normalización a través del método *Z-Score*. La Tabla 2 muestra los hiperparámetros de los modelos usados para la DCGAN.

En la Figura 5 se muestran algunas imágenes creadas a partir del generador de la DCGAN utilizada.

5.2. Clasificación de tumores

Para la clasificación de tumores, se proponen dos configuraciones experimentales. La primera configuración consiste en obtener modelos a partir del conjunto BreakHis desbalanceado, es decir, el conjunto original.

Tabla 2. Hiperparámetros de la DCGAN.

DCGAN	Tasa de aprendizaje	Adam, Beta	Épocas	Batch
Generador	0.0001	0.5, 0.999	700	512
Discriminador	0.0001	0.5, 0.999	700	512

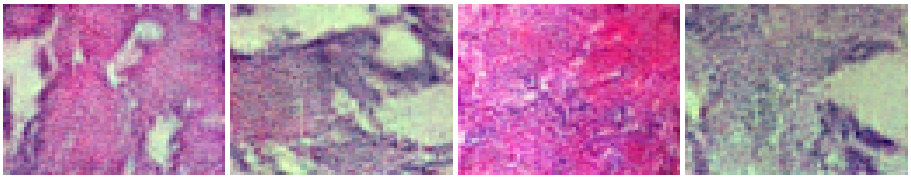


Fig. 5. Imágenes artificialmente generadas de la clase benigna.

Para la segunda configuración se utilizará el conjunto balanceado con las imágenes complementarias de la clase benigno creadas por la DCGAN.

Para ambas configuraciones experimentales se realizaron 30 entrenamientos independientes, en los que para cada uno se crearon particiones aleatorias de conjuntos de entrenamiento, validación y prueba; usando el $64\,\%$, $16\,\%$ y $20\,\%$ respectivamente para cada conjunto.

La cantidad de 10 épocas de entrenamiento para ambas configuraciones experimentales fue seleccionada empíricamente analizando el comportamiento de la función de pérdida para el conjunto de entrenamiento y validación.

Para poder medir el desempeño de los clasificadores basados en DCNN se usarán cuatro métricas, para las cuales es necesario definir que en este trabajo la clase positiva corresponde a los tumores malignos y la clase negativa a los tumores benignos. Para la cantidad de datos positivos cuyas predicciones son positivas se le denomina TP (por sus siglas en inglés -True Positive-), la cantidad de datos positivos con predicciones negativas entonces se conoce como FN (por sus siglas en inglés -False Negative-).

Por otra parte, la cantidad de datos negativos, con predicciones negativas, se le conoce como TN (por sus siglas en inglés -True Negative-), de lo contrario, la cantidad de datos negativos con predicciones positivas se le nombra como FP (por sus siglas en inglés -False Positive-) [2]. A continuación, se describen las métricas a usar y el objetivo de las mismas [5].

La exactitud (acc, Ecuación 2) indica la proporción de los datos correctamente clasificados. Esta métrica nos permite conocer de manera general el rendimiento del modelo sin poner énfasis en alguna clase en particular:

$$acc = \frac{TP + TN}{TP + FN + FP + TN}. \tag{2}$$

La precisión (prec, Ecuación 3) indica la proporción de datos positivos reales entre los datos que el modelo identifica como positivos:

$$prec = \frac{TP}{TP + FP}. \tag{3}$$

Tabla 3. Tabla de experimentos.

	No balanceado				***Balanceado***			
Exp	**acc**	**prec**	**rec**	**spec**	**acc**	**prec**	**rec**	**spec**
1	0.848	0.861	0.931	0.658	0.875	0.863	0.897	0.851
2	0.839	0.887	0.881	0.745	0.868	0.814	0.961	0.771
3	0.851	0.857	0.942	0.643	0.860	0.795	0.977	0.737
4	0.851	0.886	0.903	0.734	0.864	0.907	0.817	0.912
5	0.848	0.881	0.904	0.722	0.887	0.858	0.932	0.839
6	0.832	0.823	0.967	0.525	0.879	0.859	0.914	0.843
7	0.846	0.888	0.891	0.743	0.878	0.877	0.885	0.870
8	0.859	0.896	0.902	0.761	0.883	0.860	0.922	0.843
9	0.829	0.910	0.836	0.811	0.883	0.841	0.951	0.813
10	0.856	0.872	0.929	0.689	0.878	0.834	0.950	0.803
11	0.836	0.827	0.965	0.539	0.864	0.823	0.936	0.789
12	0.849	0.876	0.912	0.705	0.874	0.822	0.963	0.782
13	0.853	0.883	0.910	0.724	0.859	0.923	0.789	0.931
14	0.848	0.899	0.880	0.774	0.887	0.875	0.908	0.864
15	0.856	0.895	0.898	0.759	0.876	0.900	0.851	0.901
16	0.837	0.896	0.866	0.770	0.882	0.874	0.897	0.865
17	0.851	0.868	0.927	0.678	0.884	0.849	0.940	0.825
18	0.846	0.854	0.940	0.633	0.875	0.869	0.888	0.861
19	0.848	0.894	0.886	0.759	0.881	0.839	0.948	0.811
20	0.813	0.925	0.795	0.853	0.870	0.840	0.920	0.817
21	0.850	0.891	0.894	0.751	0.844	0.780	0.968	0.715
22	0.846	0.874	0.909	0.701	0.874	0.832	0.945	0.800
23	0.842	0.838	0.957	0.579	0.862	0.899	0.823	0.903
24	0.839	0.879	0.891	0.720	0.872	0.907	0.835	0.911
25	0.841	0.907	0.860	0.799	0.878	0.834	0.950	0.803
26	0.834	0.915	0.840	0.822	0.872	0.891	0.854	0.891
27	0.851	0.858	0.941	0.645	0.881	0.894	0.870	0.893
28	0.847	0.894	0.884	0.761	0.874	0.828	0.950	0.794
29	0.841	0.837	0.957	0.575	0.889	0.872	0.917	0.860
30	0.853	0.875	0.919	0.701	0.881	0.841	0.946	0.814
min	0.813	0.823	0.795	0.525	0.844	0.780	0.789	0.715
max	0.859	0.925	0.967	0.853	0.889	0.923	0.977	0.931
$\boldsymbol{\mu}$	0.844	0.878	0.902	0.708	0.874	0.856	0.909	0.836
$\boldsymbol{\sigma}$	0.011	0.027	0.044	0.088	0.011	0.037	0.053	0.057

La recuperación (rec, Ecuación 4) indica la proporción de datos positivos correctamente clasificados. Puede entenderse como un estimador de la probabilidad de clasificar correctamente un dato positivo arbitrario:

$$rec = \frac{TP}{TP + FN}. \tag{4}$$

La especificidad (spec, Ecuación 5) indica la proporción de datos negativos correctamente clasificados. Puede entenderse como un estimador de la probabilidad de clasificar correctamente un dato negativo arbitrario:

$$spec = \frac{TN}{FP + TN}. \tag{5}$$

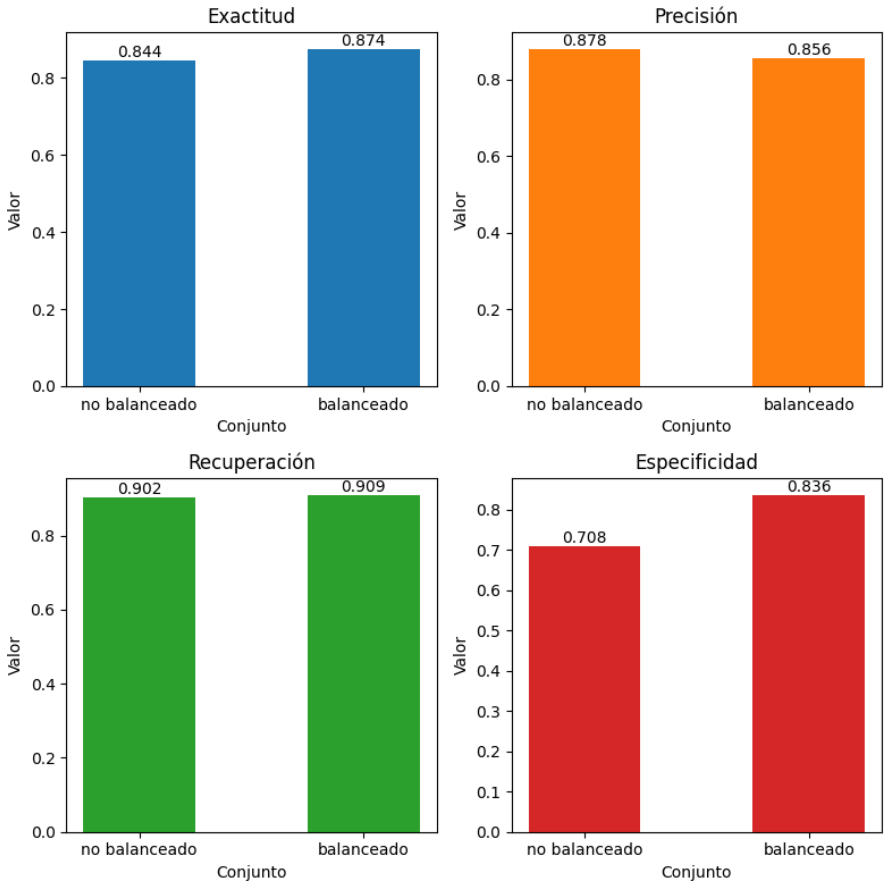


Fig. 6. Comparación de medias de las métricas de los conjuntos no balanceados y balanceados.

La Tabla 3 muestra los resultados de 30 ejecuciones utilizando el conjunto BreakHis original (parte izquierda) y el conjunto balanceado mediante GAN (parte derecha). Para cada ejecución se reportan los valores obtenidos por las métricas mencionadas usando conjuntos de pruebas, también para cada una de ellas se calculan los estadísticos mínimo, máximo, media y desviación estándar.

Con respecto al estado del arte, generalmente se reportan porcentajes de exactitud entre el $80\,\%$ y $90\,\%$ [24, 3], otras métricas como las mencionadas en este trabajo no son utilizadas en la literatura revisada. En la Tabla 3 se pueden observar exactitudes promedio $84{,}4\,\%$ y $87{,}4\,\%$ para modelos entrenados con los conjuntos no balanceado y balanceado, respectivamente.

La Figura 6 muestra gráficas de barra para comparar los valores medios de las métricas utilizadas para ambas configuraciones experimentales. La gráfica a destacar es la que involucra la especificidad, ya que en esta es donde se ve reflejado de cierta manera el impacto del balanceo de la clase negativa, teniendo un incremento aproximado del $13\,\%$ en su valor con respecto a los resultados del conjunto no balanceado.

Los resultados de exactitud y recuperación no se ven mayormente afectados y los de la precisión muestran un ligero decremento porcentual de la experimentación con datos balanceados con respecto al de los datos no balanceados.

6. Conclusiones y trabajo futuro

El uso del DCGAN puede producir imágenes que ayuden a balancear conjuntos de datos, aunque estas pueden introducir problemas de ruido [22], como se pudo observar en los resultados de la precisión de la Figura 6.

A pesar de que las imágenes sintéticas pueden no llegan a tener la misma calidad o característcas a las imágenes reales, estas pueden contribuir a mejorar en el rendimiento de clasificadores cuando existe una clase con pocos datos, caso común en el ámbito médico [17]; como se pudo observar en los resultados de la exactitud, recuperación y especificidad de la Figura 6.

Aunque los resultados obtenidos en el trabajo demuestran que el uso de una DCGAN para el balanceo de datos del conjunto BreakHis puede llegar a mejorar la exactitud, la especificidad y la recuperación, existen áreas en las que se puede trabajar para intentar conseguir mejores resultados a la hora de realizar clasificación.

Una de las posibilidades es explorar otras técnicas como SMOTE, así como arquitecturas emergentes de GAN específicas para el balanceo, como lo es BAGAN [14], que se enfoca en generar imágenes para el balanceo de conjuntos de datos, o arquitecturas GAN condicionales, en las que se puede generar imágenes de una clase en específico. Además de explorar métricas que permitan evaluar la calidad de las imágenes generadas como la Fréchet Inception Distance [13].

Otra área de trabajo a futuro podría ser la generación de imágenes sintéticas de subclases específicas en lugar de trabajar únicamente con un enfoque binario. Por ejemplo, en tareas de clasificación de tumores, sería más útil generar imágenes de diferentes tipos de tumores y en diferentes resoluciones, lo que podría mejorar la capacidad del modelo para generalizar y clasificar con mayor rendimiento los diferentes tipos de tumores presentes en el conjunto de datos BreakHis [3].

Referencias

1. Albelwi, S., Mahmood, A.: A framework for designing the architectures of deep convolutional neural networks. Entropy, vol. 19, no. 6 (2017) doi: 10.3390/e19060242
2. Alpaydin, E.: Introduction to machine learning, The MIT Press (2010)
3. Benhammou, Y., Achchab, B., Herrera, F., Tabik, S.: BreakHis based breast cancer automatic diagnosis using deep learning: Taxonomy, survey and insights. Neurocomputing, vol. 375, pp. 9–24 (2020) doi: 10.1016/j.neucom.2019.09.044
4. Bissoto, A., Valle, E., Avila, S.: GAN-based data augmentation and anonymization for skin-lesion analysis: A critical review. Computer Vision and Pattern Recognition, pp. 1847–1856 (2021) doi: 10.1109/cvprw53098.2021.00204
5. Flach, P.: Machine learning: The art and science of algorithms that make sense of data (2012)
6. Frid-Adar, M., Klang, E., Amitai, M., Goldberger, J., Greenspan, H.: Synthetic data augmentation using GAN for improved liver lesion classification. In: 2018 IEEE 15th International Symposium on Biomedical Imaging, pp. 289–293 (2018) doi: 10.1109/isbi.2018.8363576

7. Gheshlaghi, S. H., Enoch-Kan, C. N., Ye, D.: Breast cancer histopathological image classification with adversarial image synthesis. In: 2021 43rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 3387–3390 (2021) doi: 10.1109/EMBC46164.2021.9630678
8. Glassner, A.: Deep learning: A visual approach. No Starch Press, (2021) https://books.google.com.mx/books?id=NgTyDwAAQBAJ
9. Goodfellow, I.: NIPS 2016 tutorial: Generative adversarial networks. Cornell University, (2016) doi: 10.48550/arXiv.1701.00160
10. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in Neural Information Processing Systems, vol. 27 (2014)
11. Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G., Cai, J., Chen, T.: Recent advances in convolutional neural networks. Pattern Recognition, vol. 77, pp. 354–377 (2018) doi: 10.1016/j.patcog.2017.10.013
12. Gui, J., Sun, Z., Wen, Y., Tao, D., Ye, J.: A review on generative adversarial networks: Algorithms, theory, and applications. IEEE Transactions on Knowledge and Data Engineering, vol. 35, no. 4, pp. 3313–3332 (2023) doi: 10.1109/TKDE.2021.3130191
13. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs trained by a two time-scale update rule converge to a local nash equilibrium. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, vol. 30, pp. 6629–6640 (2017)
14. Huang, G., Jafari, A.: Enhanced balancing GAN: Minority-class image generation. Neural Computing and Applications, vol. 35, no. 7, pp. 5145–5154 (2020) doi: 10.1007/s00521-021-06163-8
15. Huynh-Thuy, M. B., Hoang, V. T.: Fusing of deep learning, transfer learning and GAN for breast cancer histopathological image classification. Advanced Computational Methods for Knowledge Engineering, vol. 1121, pp. 255–266 (2019) doi: 10.1007/978-3-030-38364-023
16. Isola, P., Zhu, J. Y., Zhou, T., Efros, A. A.: Image-to-image translation with conditional adversarial networks. Computer Vision and Pattern Recognition, pp. 5967–5976 (2017) doi: 10.1109/cvpr.2017.632
17. Jabbar, A., Li, X., Omar, B.: A survey on generative adversarial networks: Variants, applications, and training. ACM Computing Surveys, vol. 54, no. 157, pp. 1–49 (2021) doi: 10.1145/3463475
18. Jernelv, I. L., Hjelme, D. R., Matsuura, Y., Aksnes, A.: Convolutional neural networks for classification and regression analysis of one-dimensional spectral data (2020) doi: 10.48550/arxiv.2005.07530
19. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. Computer Vision and Pattern Recognition, pp. 4396–4405 (2019) doi: 10.1109/cvpr.2019.00453
20. Li, X., Jiang, Y., Rodriguez-Andina, J. J., Luo, H., Yin, S., Kaynak, O.: When medical images meet generative adversarial network: Recent development and research opportunities. Discover Artificial Intelligence, vol. 1, no. 5 (2021) doi: 10.1007/s44163-021-00006-0
21. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks (2015) doi: 10.48550/arXiv.1511.06434
22. Sampath, V., Maurtua, I., Aguilar-Martín, J. J., Gutierrez, A.: A survey on generative adversarial networks for imbalance problems in computer vision tasks. Journal of Big Data, vol. 8, no. 27 (2021) doi: 10.1186/s40537-021-00414-0
23. Shorten, C., Khoshgoftaar, T. M.: A survey on image data augmentation for deep learning. Journal of Big Data, vol. 6, no. 60 (2019) doi: 10.1186/s40537-019-0197-0

24. Spanhol, F. A., Oliveira, L. S., Petitjean, C., Heutte, L.: A dataset for breast cancer histopathological image classification. IEEE Transactions on Biomedical Engineering, vol. 63, no. 7, pp. 1455–1462 (2016) doi: 10.1109/TBME.2015.2496264
25. Yamashita, R., Nishio, M., Gian-Do, R. K., Togashi, K.: Convolutional neural networks: An overview and application in radiology. Insights Into Imaging, vol. 9, no. 4, pp. 611–629 (2018) doi: 10.1007/s13244-018-0639-9
26. Yi, X., Walia, E., Babyn, P.: Generative adversarial network in medical imaging: A review. Medical Image Analysis, vol. 58 (2019) doi: 10.1016/j.media.2019.101552

ISSN 1870-4069

Etiquetado, clasificación y análisis de calidad de imagen para detección de retinopatía diabética usando modelos convolucionales profundos

Pedro de J. Bermejo-Guerrero[1], Abraham Sánchez[2], E. Ulises Moya-Sánchez[1,2], Ulises Cortés[3]

[1] Universidad Autónoma de Guadalajara, Posgrado en Ciencias Computacionales, México

[2] Gobierno de Jalisco, Coordinación General de Innovación Gubernamental, México

[3] Barcelona Supercomputing Center, España

eduardo.moya@jalisco.gob.mx

Resumen. La retinopatía diabética (RD) es la principal causa de ceguera en el mundo en edad laboral. La detección-tratamiento temprano ha demostrado ser efectiva para disminuir las complicaciones visuales (y la ceguera) de los pacientes diabéticos. Los modelos profundos han mostrado recientemente que pueden ayudar con la detección temprana analizando masivamente un gran número de imágenes. Uno de los problemas con esos modelos en su aplicación en el mundo real es que la calidad de la imagen puede afectar significativamente el desempeño de los modelos haciendo inviable la evaluación automática. En este trabajo presentamos tres elementos principales i) Etiquetado de calidad: criterios, herramientas, etiquetas en imágenes públicas; ii) comparación de modelos convolucionales y iii) análisis sobre el desempeño de los modelos con degradaciones sintéticas. Los resultados muestran que se puede alcanzar una exactitud del 97 % y los análisis ayudan a entender los tipos de degradaciones que afectan más a las redes. Estos modelos y resultados serán especialmente útiles para quienes tengan interés de desplegar modelos de RD en ambientes reales.

Palabras clave: Retinopatía diabética, calidad de imagen, modelos convolucionales profundos.

Labeling, Classification and Image Quality Analysis for Detection of Diabetic Retinopathy Using Deep ConvNets Models

Abstract. Diabetic retinopathy (DR) is the leading cause of blindness in the working-age population worldwide. Early detection and treatment have been

shown to be effective in reducing visual complications (and blindness) in diabetic patients. Deep learning models have recently demonstrated that they can aid in early detection by analyzing a massive number of images. One problem with these models in real-world applications is that image quality can significantly impact model performance, making automatic evaluation unfeasible. In this work, we present three main elements: i) Quality labeling - criteria, tools, and labels in public images; ii) Comparison of convolutional models; and iii) Analysis of model performance with synthetic degradations. The results show that an accuracy of 97% can be achieved, and the analyses help understand the types of degradation that most affect the networks. These models and results will be especially useful for those interested in deploying DR models in real-world environments.

Keywords: Diabetic rethinopathy, deep Conv-Nets, image quality.

1. Introducción

En los últimos años, el aprendizaje profundo (AP) o Deep Learning se ha convertido en una poderosa herramienta para clasificar imágenes médicas [14]. En Retinopatía Diabética (RD) se han usado exitosamente las redes neuronales convolucionales (ConvNets) para la clasificación del nivel de RD en las imágenes de fondo de ojo [7, 15, 8]. Sin embargo, el desempeño de estos modelos puede disminuir significativamente cuando se usan datos del mundo real [19, 1, 8, 12, 11].

Aunque los sistemas de AP generalmente se entrenan y prueban en conjuntos de datos de imágenes de alta calidad, no se puede asumir que las imágenes del mundo real sean de calidad aceptable. Además, se ha demostrado que los modelos profundos convolucionales son muy sensibles a cambios de calidad de imagen, por ejemplo al cambio de: contraste, borrosidad y ruido [6]. Por lo tanto, es necesario tener herramientas para una evaluación de la calidad de la imagen [15].

En este trabajo presentamos tres elementos principales para poder evaluar y clasificar la calidad de imágenes en el contexto especifico de la RD: i) etiquetado de imágenes públicas: herramientas, criterios específicos para la RD; ii) comparación de modelos convolucionales y iii) análisis sobre el desempeño de los modelos con degradaciones sintéticas.

Una de las contribuciones de este trabajo es ir más allá de los criterios clásicos de calidad de imagen: ruido, contraste, definición, agregando la presencia de elementos anatómicos: nervio óptico, y macula los cuales son fundamentales para detectar el nivel de RD de manera más confiable (incluso para el médico especialista).

Además, consideramos que es una contribución presentar una comparación del desempeño de varios modelos convolucionales. Por último, nuestro análisis con imágenes degradadas sintéticamente nos ayuda a evaluar el desempeño de los clasificadores en ambientes controlados, algo que no hemos visto en otros trabajos que abordan este tema.

Algunos trabajos similares son por ejemplo Zago ([20]) y Chalakkal ([3]), quienes obtienen un desempeño inferior del 98 % y 91.4 % en exactitud respectivamente. Sin embargo no toman en cuenta análisis de desempeño después del entrenamiento.

Además por desgracia no podemos comparar directamente estos trabajos ya que se usan datos diferentes y los códigos no están disponibles. Actualmente desarrollamos un conjunto de modelos para datos de México y las pruebas preliminares (sólo con datos públicos) muestran buenos resultados. Esperamos que estos resultados ayuden a evitar datos fuera de la distribución (por ejemplo datos de mala calidad) y así tener modelos más confiables para su aplicación en el mundo real.

2. Conceptos y fundamentos

Para etiquetar la calidad de imagen en RD no sólo usamos conceptos básicos de calidad de imagen, como ruido, contraste, nitidez, sino que además agregamos aspectos particulares RD como son artefactos o lesiones visibles y aspectos anatómicos particulares únicos de estas imágenes. A continuación enlistamos los conceptos de calidad de imagen más relevantes para el etiquetado de imagen de fondo de ojo con RD:

1. El *ruido* (cociente señal a ruido) es una variación aleatoria de los valores de intensidad de una imagen. La explicación del fenómeno tiene su origen frecuentemente en la poca señal detectada por el sensor (por ejemplo poca luz) [2].
2. El *rango dinámico*, es el intervalo de niveles de luz que el sensor (cámara) puede capturar [9]. En general los rango dinámicos de las imágenes médicas son mayores (> 8 bits) que las de las imágenes convencionales. Además, el rango dinámico está ligado con *el contraste*, que está vinculada a la diferencia relativa entre los valores intensidad de imagen en dos puntos de la imagen.
3. La *nitidez* determina el nivel de detalle (objeto más pequeño que se puede separar) en una imagen.
4. Los *artefactos* en RD son objetos, texturas o regiones de saturación que no corresponden al objeto real. Por ejemplo, elementos inherentes de la cámara, o destellos de luz en las imágenes de fondo de ojo.
5. Aspectos del *campo de visión* (anatómicos o lesiones en RD) tales como el nervio óptico, macula y lesiones como microaneurismas. Según la experiencia de médicos especialistas estos dos elementos son fundamentales y si no se encuentran debería de repetirse la captura o evaluarse para otras enfermedades.

3. Datos

3.1. Etiquetar imágenes

Se etiquetaron imágenes del conjunto de datos EyePACS-Kaggle [4], IDRID [13] y MESSIDOR [5] por tres personas. Usamos varios conjuntos de datos independientes y varias personas para disminuir la probabilidad de sesgos en el etiquetado y en los datos.

3.2. Clasificación

En la Tabla 1 mostramos como se dividieron el número de imágenes en los conjuntos de entrenamiento, validación y pruebas.

Tabla 1. División de conjunto de datos.

Conjunto	Buena calidad	Mala calidad
Entrenamiento	715	685
Validación	146	125
Pruebas	136	167
Total	997	977

3.3. Análisis de desempeño

Para analizar que tan sensibles eran estos modelos, se usaron 1,813 imágenes de buena calidad que no estaban en el conjunto de entrenamiento, validación o pruebas. Se eligieron tres tipo de degradaciones de imagen: blur (desenfoque), gauss noise (ruido) y random fog (artefactos y cambios de contraste) mediante el uso de Albumentations[4].

4. Metodología y diseño experimental

4.1. Etiquetado de imágenes

Se tomaron en cuenta cinco factores para hacer el etiquetado de calidad de las imágenes de fondo de ojo de retinopatía (ver Figura 1):

1. Artefactos visibles. Se encuentran o no artefactos en el campo visual.
2. Nitidez. Distinción (binario) de los elementos anatómicos (separables). Por ejemplo, se pueden distinguir microaneurismas, neovasos y capilares hemorragias, exudados.
3. Campo de visión. Detección de disco óptico y mácula.
4. Lesiones graves. Existencia de problemas graves como hemorragias, dilatación venosa, desprendimientos, cirugías. En nuestra experiencia era importante etiquetar imágenes con estas lesiones ya que algunas de estas lesiones se puede confundir con artefactos.
5. Evaluación general. Clasificación evaluable (imagen de calidad), no evaluable (imagen de mala calidad).

El proceso de etiquetado fue realizado por tres personas y la etiqueta final fue dada por la mayoría. En la figura 1 se presenta un ejemplo de la herramienta interna utilizada para etiquetar las imágenes de fondo de ojo. Es importante notar que los factores antes mencionados se presentan como botones binarios que ayudan al usuario a etiquetar más rápidamente los factores. También queremos destacar que el tamaño de la imagen (número de pixeles) fue el tamaño original, ya que nos permitía ver con más detalle las lesiones o artefactos pequeños dentro de la imagen.

4.2. Clasificación de calidad

Se eligieron cuatro redes para comparar su rendimiento: InceptionV3 [18], MobilenetV2 [16], Resnet50 [10] y VGG19 [17]. Se eligieron estas redes ya que han tenido buenos resultados en la clasificación de imágenes para tareas similares.

[4] https://albumentations.ai/

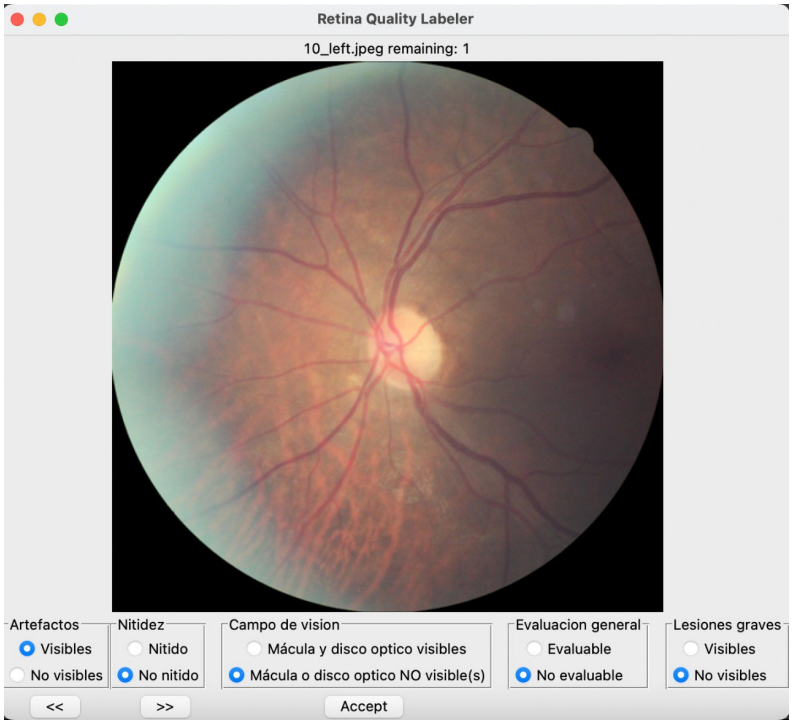


Fig. 1. Herramienta que se usó para etiquetar imágenes de fondo de ojo.

Sin embargo su tamaño (número de parámetros) y arquitecturas diferentes nos motivó a comparar su desempeño. A continuación se describe una lista de las arquitecturas y sus características más importantes.

- InceptionV3: 159 capas y 23.8 millones de parámetros.
- MobilenetV2: 53 capas y 3.4 millones de parámetros.
- Resnet50: 50 capas y 25.6 millones de parámetros.
- Vgg19: 19 capas y 143 millones de parámetros.

Podemos ver ciertas diferencias, InceptionV3 tiene un gran numero de capas pero un número medio de parámetros, MobilenetV2 tiene pocas capas y pocos parámetros siendo la red más ligera, Resnet50 tiene pocas capas y un número medio de parámetros, por último Vgg19 tiene muy pocas capas pero una gran cantidad de parámetros, siendo la red más pesada.

Los entrenamiento fueron usando los modelos pre-entrenados (imagenet) por lo que fue necesario sólo entrenar durante 20 épocas. Las imágenes de entrada tienen un resize de width=300, height=300 y una normalización con valores mean=[0, 0, 0], std=[1, 1, 1]. No se realizó ningún tipo de aumento de datos, ya que en estas imágenes las modificaciones más simples pueden alterar su clasificación; una imagen de buena calidad podría cambiar a ser una de mala calidad.

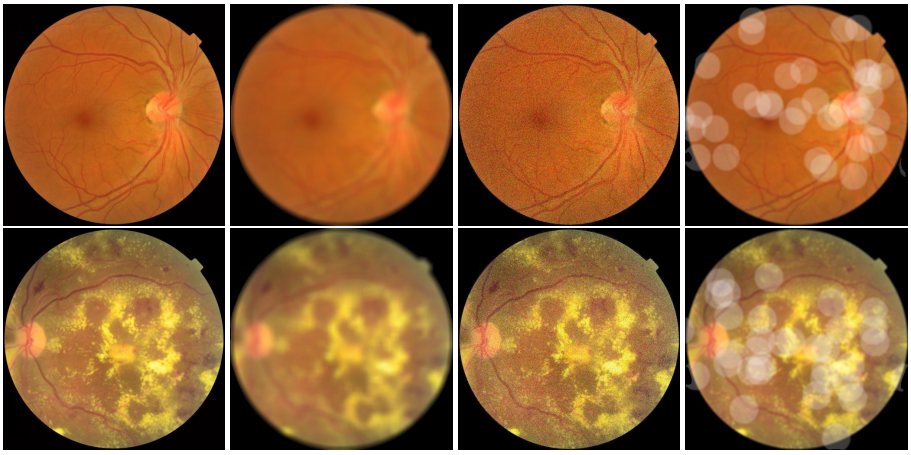


Fig. 2. De izquierda a derecha se muestra: la imagen original, con borrosidad, ruido y niebla aleatoria.

Además se modificó la última capa de acuerdo al número de clases (2). Para entrenar se usó Pytorch como framework de AP. Para evaluar el desempeño de las redes convolucionales de clasificación de calidad de imagen, se usaron las siguientes métricas:

$$Exactitud = \frac{TP + TN}{TP + TN + FP + FN}, \tag{1}$$

$$Sensibilidad = \frac{TP}{TP + FN}, \tag{2}$$

$$Especificidad = \frac{TP}{TN + FP}, \tag{3}$$

En que TP = True Positive, TN = True Negative, FP = False Negative, y FP = False Positive.

4.3. Análisis de desempeño

El análisis de desempeño nos ayuda a evaluar la robustez/sensibilidad de los modelos ante tres ataques o degradaciones de imágenes. Se compara el intervalo de confianza con las imágenes originales (good) de buena calidad y se compara la distribución cuando se hace la degradación sintética de las imágenes. En la Figura 2 se muestran dos imágenes y sus degradaciones sintéticas, borrosidad (blur), ruido (noise), y niebla aleatoria (random fog).

5. Resultados y análisis

5.1. Etiquetado de imágenes

El principal resultado de esta subsección es haber creado una herramienta simple para etiquetar offline las imágenes de fondo de ojo con los criterios específicos de la RD.

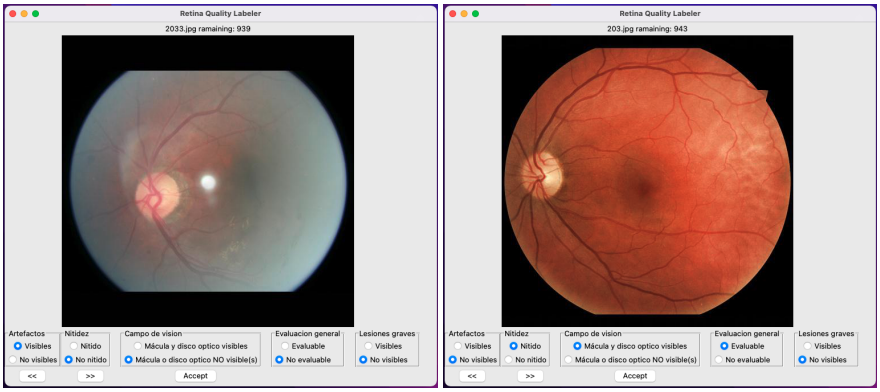


Fig. 3. Ejemplo de la herramienta que se usó para etiquetar imágenes.

En la Figura 3 se muestran dos ejemplos del etiquetado de imágenes. Los ejemplos son imágenes de buena calidad y con artefactos. El segundo resultado más importante de esta subsección es que se etiquetaron más de tres mil imágenes por tres personas tomando en cuenta aspectos cualitativos de la imagen.

Esperamos que esto reduzca los sesgos en las etiquetas de calidad y ayuden a un mejor aprendizaje del modelo. El aprendizaje que tuvimos al etiquetar estás imágenes nos ayudo a definir parámetros reales (acertados) para degradar las imágenes sintéticas (ver subsección Análisis de desempeño).

Es importante mencionar que las etiquetas estarán disponible públicamente en nuestro repositorio[5]. Esperamos que estas herramientas y etiquetas sirvan para que otras personas o grupos de investigación y además puedan comparar nuestro trabajo e incluso puedan pre-etiquetar sus imágenes de fondo de ojo.

5.2. Clasificación de calidad

En la Tabla 2 se muestran los resultados del entrenamiento de cuatro redes. de acuerdo a los resultados de la tabla, la red que tiene mejor exactitud, especificidad (Espe) y sensibilidad (Sen) es la Vgg19. Sin embargo los tiempos de entrenamiento y predicción son los más altos de todos los modelos.

Estos modelos como mencionamos anteriormente se han usado para tareas similares con gran éxito. Para poder entender porqué este modelo Vgg19 funciona mejor proponemos el análisis que hacemos en la siguiente sección. Además, para la explicabilidad, debemos realizar pruebas donde se muestra el gradiente de las últimas capas con imágenes de otro conjunto de datos.

5.3. Análisis de desempeño

En la Tabla 3 se presenta el porcentaje de imágenes clasificadas como mala calidad después de una degradación. El mayor porcentaje se asocia con una red más sensible a una degradación.

[5] https://github.com/PedroBermejo/retinopathy

Tabla 2. Comparación de desempeño de clasificación de calidad de imagen de fondo de ojo. En que T. Entr es el tiempo de entrenamiento, T. Pred es el tiempo de predicción, Sen es la $Sensibilidad$ y Espe es la $Especificidad$.

Modelo	Exactitud	Sen	Espe	T. Entr. (s)	T. Pred (s)
InceptionV3 [18]	0.91	0.91	0.91	605	52
MobileNetV2[16]	0.89	0.90	0.89	382	32
ResNet50 [10]	0.95	0.95	0.95	556	68
Vgg19[17]	0.97	0.98	0.95	983	145

Tabla 3. Porcentaje de imágenes que se clasificaron como de mala calidad después de las degradaciones. Un porcentaje mayor representa un mayor cambio o sensibilidad ante esas degradaciones.

Modelo	Borrosidad	Ruido	Niebla aleatoria
InceptionV3 [18]	62 %	4 %	11 %
MobilenetV2 [16]	47 %	4 %	11 %
Resnet50 [10]	9 %	1 %	6 %
Vgg19 [17]	95 %	8 %	76 %

Se puede observar que la borrosidad (blur) y niebla aleatoria (random fog) son las degradaciones que más afectan a las redes. Por otra parte el ruido es la degradación que menos afecta a las redes. Se puede observar que la red Vgg19 es la red más sensible a cada uno de los tipos de degradación. De acuerdo a este análisis y a los experimentos de clasificación consideramos que Vgg19 es red más sensible para esta tarea a pesar de ser la red más pesada y con mayor tiempo de entrenamiento y predicción.

En la Figura 4 se presenta la comparación de la distribución del intervalo de confianza (IC) mediante cajas y bigotes. El IC = 0 representa imágenes clasificadas con buena calidad mientras que el valor IC = 1 es el correspondiente a las clasificadas de mala calidad. Se puede observar como las imágenes originales (etiquetada como *referencia*) su distribución es cercana al cero.

Después de degradar esas imágenes podemos ver que el intervalo de confianza se distribuye con valores más cercanos a uno en el caso de Vgg10 e InceptionV3. Esta figura se puede considerar como representación gráfica de la tabla 3 y nos indica que tan sensibles son las redes a los tipos de degradación.

6. Conclusiones

En este trabajo presentamos tres elementos principales: etiquetado, clasificación y análisis. El etiquetado de calidad de imagen presenta los elementos y herramientas en el contexto específico de RD, ya que consideramos en el proceso de etiquetado los elementos de calidad de imagen y elementos anatómicos asociados al fondo de ojo y lesiones vinculadas a la enfermedad de RD.

Además, presentamos la comparación del desempeño de clasificación cuatro modelos convolucionales.

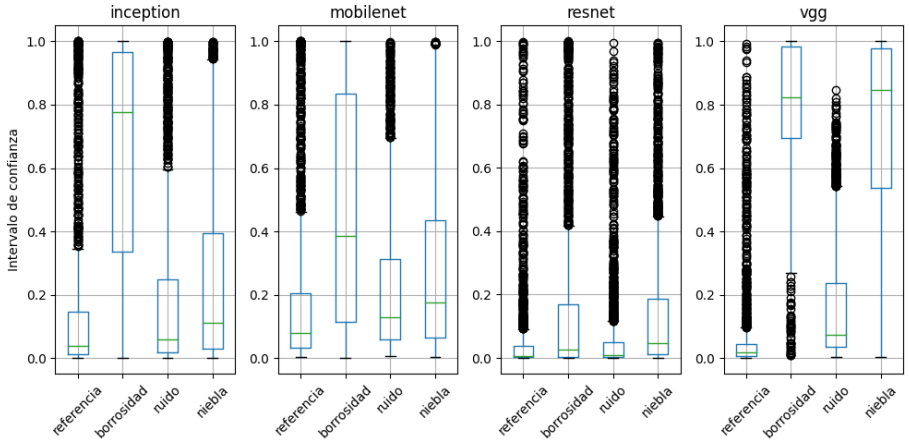


Fig. 4. Análisis del desempeño de clasificación con imágenes sintéticas. Se presenta en intervalo de confianza de las imágenes de referencia y de las mismas imágenes degradadas sintéticamente. De izquierda a derecha referencia (good quality), borrosidad (blur), ruido (gauss noise), niebla (random fog).

El modelo Vgg19 presenta mejores métricas de exactitud, especificidad y sensibilidad, pero con el mayor tiempo de entrenamiento e inferencia. Por último presentamos un análisis de los modelos cuando se hace la inferencia con imágenes degradadas sintéticamente.

El análisis del desempeño nos confirma que el modelo Vgg19 es el más sensible ante cambios de borrosidad o niebla aleatoria. Un modelo con más sensibilidad nos ayuda a una mejor clasificación de imágenes de buena o mala calidad.

En contraste un modelo con muy poca sensibilidad como lo es Resnet50 podría ser muy buena opción para la etapa de clasificación de niveles de retinopatía.

Estos modelos y resultados serán especialmente útiles para quienes tengan interés de desplegar modelos en ambientes reales ya que consideran aspectos de robustez y rapidez (menor tiempo de inferencia).

Consideramos que en un futuro este análisis nos servirá para implementar un sistema de modelos profundos con imágenes regionales que ya han sido colectadas por el Gobierno del Estado de Jalisco.

Referencias

1. Beede, E., Baylor, E., Hersch, F., Iurchenko, A., Wilcox, L., Ruamviboonsuk, P., Vardoulakis, L. M.: A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pp. 1–12 (2020) doi: 10.1145/3313831.3376718
2. Bushberg, J. T., Seibert, J. A., Leidholdt, E. M., Boone, J. M.: The essential physics of medical imaging, Lippincott Williams and Wilkins (2011)

3. Chalakkal, R. J., Abdulla, W. H., Thulaseedharan, S. S.: Quality and content analysis of fundus images using deep learning. Computers in Biology and Medicine, vol. 108, pp. 317–331 (2019) doi: 10.1016/j.compbiomed.2019.03.019
4. Cuadros, J., Bresnick, G.: EyePACS: An adaptable telemedicine system for diabetic retinopathy screening. Journal of diabetes science and technology, vol. 3, no. 3, pp. 509–516 (2009) doi: 10.1177/193229680900300315
5. Decencière, E., Zhang, X., Cazuguel, G., Lay, B., Cochener, B., Trone, C., Gain, P., Ordonez, R., Massin, P., Erginay, A., Charton, B., Klein, J. C.: Feedback on a publicly distributed database: The Messidor database. Image Analysis & Stereology, vol. 33, no. 3, pp. 231–234 (2014) doi: 10.5566/ias.1155
6. Dodge, S., Karam, L.: Understanding how image quality affects deep neural networks. In: 2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX), pp. 1–6 (2016) doi: 10.1109/QoMEX.2016.7498955
7. Fan, R., Liu, Y., Zhang, R.: Multi-scale feature fusion with adaptive weighting for diabetic retinopathy severity classification. Electronics, vol. 10, no. 12 (2021) doi: 10.3390/electronics10121369
8. Gaspar-González, B., Sánchez, A., Ortega-Cisneros, S., Pinedo-Diaz, G., García-Contreras, M. S., Alvarado-Castillo, B., Moya-Sánchez, E. U.: Automatic cropping of retinal fundus photographs using convolutional neural networks. Research in Computing Science, vol. 149, no. 5, pp. 161–167 (2020)
9. Gonzalez, R. C., Woods, R. E.: Digital image processing. Pearson Prentice Hall (2001)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016) doi: 10.1109/CVPR.2016.90
11. Moya-Sánchez, E. U., Xambó-Descamps, S., Sánchez-Pérez, A., Salazar-Colores, S., Cortés, U.: A trainable monogenic ConvNet layer robust in front of large contrast changes in image classification. IEEE access, vol. 9, pp. 163735–163746 (2021) doi: 10.1109/ACCESS.2021.3128552
12. Pinedo-Diaz, G., Ortega-Cisneros, S., Moya-Sanchez, E. U., Rivera, J., Mejia-Alvarez, P., Rodriguez-Navarrete, F. J., Sanchez, A.: Suitability classification of retinal fundus images for diabetic retinopathy using deep learning. Electronics, vol. 11, no. 16 (2022) doi: 10.3390/electronics11162564
13. Porwal, P., Pachade, S., Kokare, M., Deshmukh, G., Son, J., Bae, W., Liu, L., Wang, J., Liu, X., Gao, L., Wu, T., Xiao, J., Wang, F., Yin, B., Wang, Y., Danala, G., He, L., Choi, Y. H., Lee, Y. C., Jung, S. H., et al.: IDRiD: Diabetic retinopathy – segmentation and grading challenge. Medical Image Analysis, vol. 59 (2020) doi: 10.1016/j.media.2019.101561
14. Rajpurkar, P., Chen, E., Banerjee, O., Topol, E. J.: AI in health and medicine. Nature medicine, vol. 28, no. 1, pp. 31–38 (2022) doi: 10.1038/s41591-021-01614-0
15. Ruamviboonsuk, P., Tiwari, R., Sayres, R., Nganthavee, V., Hemarat, K., Kongprayoon, A., Raman, R., Levinstein, B., Liu, Y., Schaekermann, M., Lee, R., Virmani, S., Widner, K., Chambers, J., Hersch, F., Peng, L., Webster, D. R.: Real-time diabetic retinopathy screening by deep learning in a multisite national screening programme: A prospective interventional cohort study. The Lancet Digital Health, vol. 4, no. 4, pp. e235–e244 (2022) doi: 10.1016/s2589-7500(22)00017-6
16. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L. C.: MobileNetV2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4510–4520 (2018) doi: 10.1109/CVPR.2018.00474
17. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: 3rd International Conference on Learning Representations (ICLR 2015) (2015) doi: 10.48550/arXiv.1409.1556

18. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2818–2826 (2016) doi: 10.1109/CVPR.2016.308
19. Wei-Ting, D. S., Cheung, C., Lim, G., Wei-Tan, G. S., Quang, N. D., Gan, A., Hamzah, H., Garcia-Franco, R., San-Yeo, I. Y., Lee, S. Y., Wong, E., Sabanayagam, C., Baskaran, M., Ibrahim, F., Tan, N. C., Finkelstein, E. A., Lamoureux, E. L., Wong, I. Y., Bressler, N. M., Sivaprasad, S., et al.: Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. JAMA, vol. 318, no. 22, pp. 2211–2223 (2017) doi: 10.1001/jama.2017.18152
20. Zago, G. T., Andreão, R. V., Dorizzi, B., Teatini-Salles, E. O.: Retinal image quality assessment using deep learning. Computers in Biology and Medicine, vol. 103, pp. 64–70 (2018) doi: 10.1016/j.compbiomed.2018.10.004

ISSN 1870-4069

Anime Success Prediction Based on Synopsis Using Traditional Classifiers

Jesús Armenta-Segura, Grigori Sidorov

Instituto Politécnico Nacional,
Centro de Investigación en Computación,
Mexico

{jarmentas2022, sidorov}@cic.ipn.mx

Abstract. For predicting the success of an anime in its early stages of development, a baseline is proposed in this paper, based on the synopsis of its plot. AniSyn7 is presented, which is a corpus consisting of $6,928$ anime synopsis with binary labels of successful/unsuccesful. The corpus was explored by vectorizing the synopsis using n-grams and dependence trees, so three traditional machine learning classifiers (Support Vector Machine, Gaussian Naive Bayes, and Logistic Regression) can be employed in order to study correlation between synopsis and success.

Keywords: Anime success prediction, entertainment success prediction, machine learning, natural languaje processing.

1 Introduction

In recent years, the anime industry has become highly profitable. For example, Jujutsu Kaisen, the most lucrative anime of 2022,[1] generated $76 million in Japan alone. However, failure is equally punished, as demonstrated by the anime movie Final Fantasy: The Spirits Within, which lost more than $50 million.[2] Given that high risk on the investment, being able to predicting the success of an anime in early stages of development is crucial for the industry.

The success of an anime depends on several factors, including plot quality, animation, voice acting, soundtrack, marketing, among others. However, in the initial stages of development, plot features are the most accessible and cost-effective to assess as it happens with movie productions [7]. Hence, inspecting the correlation between plot quality and success is a prudent starting point for research.

Plots can be summarized through its synopsis, which can also be marketed as a standalone product, as it happens in Hollywood [5]. Hence, reducing the plot to its synopsis is a very effective and low-cost way to start this inspection. About studying correlation between variables, in recent times, machine learning classification techniques have emerged as powerful tools for tackle that issue, because their ability to

[1] According to Oricon's success rating: www.oricon.co.jp/special/61353/5/ An english explanation can be find in erzat.blog/oricons-yearly-manga-sales-ranking-2022/

[2] www.boxofficemojo.com/release/rl3008595457/

		Predicted by the method	
		Successful	**Unsuccessful**
Real	**Successful**	True Positives (TP)	False Negatives (FN)
	Unsuccessful	False Positives (FP)	True Negatives (TN)

Fig. 1. True is for correctly labeled and False for wrong labeled. In this case, positive means successful and negative means unsuccessful.

learn and identify patterns and relationships within corpora. Support Vector Machine (SVM), Naive Bayes (NB), and Logistic Regression (LogReg) are three of the most prominent classifiers, and had proven to be useful by establishing performance baselines in other text related tasks [8, 13, 14, 19, 12, 6, 17, 18].

Hence, inspecting the correlation between plot and success through them will conform a suitable baseline for the anime success prediction task. In this paper we present AniSyn7: a balanced corpus with $6,928$ anime synopsis scrapped from MyAnimeList (MAL)[3]. They are associated with a binary golden label based in MAL's weighted score.

We explore this corpus with the three machine learning classifiers mentioned above, and we gain valuable insights about the correlation between synopsis and success, through their confusion matrices. Our most successful approach was Bag of Words + Bigrams + Trigrams + Character Trigrams classified with Support Vector Machine and Logistic Regresion (BBTC+SVM and BBTC+LogReg), which both achieved an F1 score of 0.55.

2 Related Work

Several work has been done in the most general task of entertainment success prediction through natural language processing (NLP) and machine/deep learning techniques. For instance, in [7] the authors used deep learning for predicting movie success. In [8, 9] the authors tackled the book success prediction task:

In [9], they used a multitask setting to predict both success and genre, which led to a significant improvement in performance with respect to their state of the art; whereas in [8] they used neural networks and transformers with an original vectorization based on the emotional lexicon of the text, which yield an improvement of the results from [9].

Focusing on anime success, there are various works on the closely related task of recommender systems design [4, 11, 16]. The most closely related work to the success prediction task is [1], where the authors studied the correlation between reviews and success using sentiment analysis in the reviews. However, they utilized a very unbalanced dataset with less than 5% of negative reviews, so they used data augmentation techniques.

[3] myanimelist.net/

[4] Image source: https://scikit-learn.org/stable/modules/cross_validation.html.

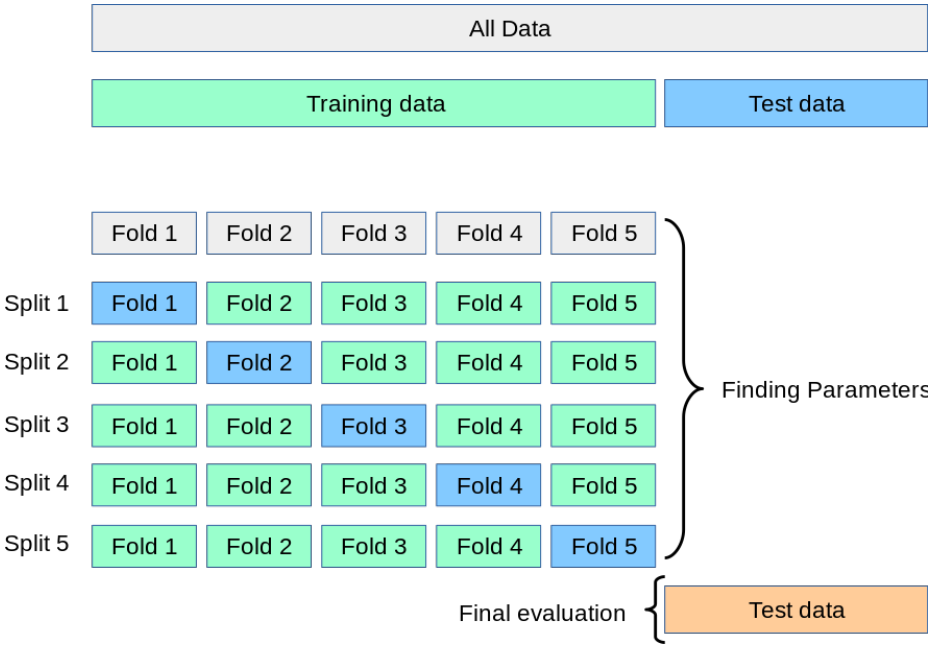


Fig. 2. In K-fold cross validation, the MLC must to be trained K times, each time taking a different fold as test set (in dark blue). At the end, the model is checked with the test data[4].

3 Methods

3.1 Vectorization

Machine learning classifiers (MLC) must to be feed with a vectorial representation of the data. In this paper we focus in three vectorizations: two based in N-grams and one based in dependence trees.

***N*-grams:** A token is defined as a fundamental unit of text, such as a word or a single character. An N-gram is defined as a contiguous sequence of N tokens. By counting the frequencies of appearance of N-grams, it is possible to vectorize a text, as follows:

1. We collect all the N-grams from all synopsis in the entire dataset, and create a vocabulary consisting of all non-repeated N-grams.
2. Finally, we represent each text in the dataset as a vector, where the i-th entry corresponds to the frequency of the i-th N-gram in the vocabulary, who can be zero.

This approach captures some lexical information and have prove to be useful in tackling NLP tasks as sentiment analysis, authorship attribution and fake news spreaders detection [2, 3, 10, 15, 12]. Usually, this vectorizations also comes along with lemmatization and stopword removal, so we do the same.

In this paper we used Bag of Words (BoW), where $N = 1$ as well as trigrams, where $N = 3$, with tokens defined as words. Besides BoW alone, we also applied various N-grams vectorizations together, at the same time:

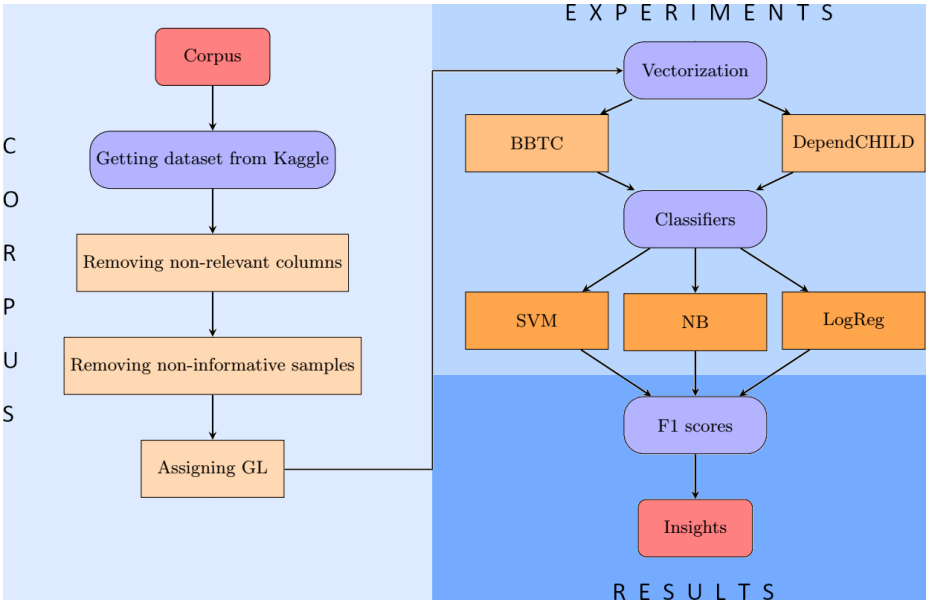


Fig. 3. Workflow for the anime success prediction task.

BBTC is an abbreviation for "Bag of Words, bigrams, trigrams, and character trigrams", which are $1, 2, 3$-grams respectively. In the first three cases, the tokens are defined as words, whereas in the fourth case the tokens are defined as individual characters. We used the Stanza python library for lemmatizing and removing stopwords, as well of the re package to remove special characters through regex.

Dependency trees: They are a way to represent the syntactical structure of a sentence by breaking it down into a hierarchical structure of grammatical dependencies between words. In a dependency tree, each word in a sentence is represented as a node, and the syntactical relationships between them corresponds to directed edges.

A naive baseline to take advantage of this syntactical representation is by counting the children of every node in a dependency tree (DependCHILD). This allows to vectorize a text, as follows:

1. We get the dependency tree of a synopsis, using Spacy python library.
2. We create a vector whose i-th entry corresponds to the amount of children of the i-th word.
3. We ensure uniform vector length by adding zeroes as needed until we reach the size of the largest vector.

As long as we know, this naive approach has never been used before in other NLP tasks. However, we admit that is extremely rudimentary and should be used only for stating baselines purposes.

Table 1. Examples of animes whose synopsis are not informative.

Anime Title	Synopsis
Poo Poo kids	A comedy series about farting children
Kabukichou Sherlock OVA	A fun collection of six short side stories depicting the past and present of Sherlock, Watson, Moriarty, and the rest of the Row House as they go about their everyday lives in their home of Kabukicho
Virtual-san wa Miteiru	Virtual-san wa Miteiru is an anime that brings Virtual YouTubers to life. The term Virtual YouTuber, or VTuber, refers to (...) Virtual-san wa Miteiru is a one-of-a-kind comedy certain to bring delight to any fan of Virtual YouTubers!

3.2 Classifiers

SVM: It is a supervised MLC that approximates the best separation between data. In this case, we ask SVM to finds the best hyperplane (line, plane, volume, etc.) whose defined semi-spaces divided successful animes from unsuccesful. The performance of SVM depends on how well the data can be separated by such hyperplane.

If several successful animes result to be nearby to several unsuccessful animes, SVM may not be able to make a proper separation. Since the vectors are defined in terms of selected features, this failure of SVM means that the phenomenom is not properly characterized by those features.

NB: Is a probabilistic MLC that assigns a label through a random variable who is calibrated using a dataset with binary golden labels. In this case, Naive Bayes computes the probability of an anime being successful, given the vectorial representation of its synopsis, and viceversa. For that computation, it uses the generalized Bayes theorem applied to every feature f_i, assuming that their appearance in the synopsis are independent of each other[5]:

$$P(\text{Success}|f_1,\ldots,f_n) = \frac{P(f_1,\ldots,f_n|\text{Success}) \times P(\text{Success})}{P(f_1,\ldots,f_n)}. \tag{1}$$

The performance of NB strongly depends on the presence of a feature in every labeled data. If a synopsis contains a majority of features that NB assigned a higher probability of being related with success, then that synopsis will be labeled as successful.

Wrong labeling can happens if either several features of the synopsis has probabilities very near to $1/2$, which means that every feature has similar presence in both labelings, and hence there is no statistical correlation. In this work we use Gaussian Naive Bayes, who assumes a normal distribution for the random variable.

Logistic Regression (LogReg): It is a supervised MLC similar to linear regression, but for binary predictions (successful/unsuccessful). The algorithm uses a sigmoid function to map the vectorized synopsis to a probability value.

[5] Since that rarely happens in practice, this classifier is considered as naive.

Table 2. Discarding of underrepresented genres.

Genre	Description	Merged with	Justification
Unknown	No genre	eliminated	For noise reasons.
Gourmet	About gastronomy	eliminated	It is actually a theme.
Award Winning	Won a prestigious award	eliminated	Feature beyond of plot.
Avant Garde	Experimental animes	eliminated	Although narrative is distinct, contents are not.
Mystery/Suspense	About riddling mysteries	eliminated	The most painful loss since they are well-stablished genres beyond anime.
Boys/Girls love	Softporn romance	Romance	Softporn is not considered Hentai (proof: Fairy Tail is not considered Hentai).
Erotica	Explicit erotism	Hentai	Although Hentai (変態) means pervert, we decided to take it as a synonym of porn.
Sports	About sports	Action	Sports are full of action.
Horror	Frightening content	Supernatural	Usually horror is based in paranormal events (not always).

If the probability is greater than a certain threshold, the anime is predicted as successful, otherwise it is predicted as unsuccessful. The algorithm learns the coefficients of the input variables by minimizing the difference between the predicted probability and the actual output value in the training data.

3.3 Score Measures

$F1$-score is a popular measure for MLC correctness. It is calculated in base of how many successful/unsuccessful animes were correctly classified or not.

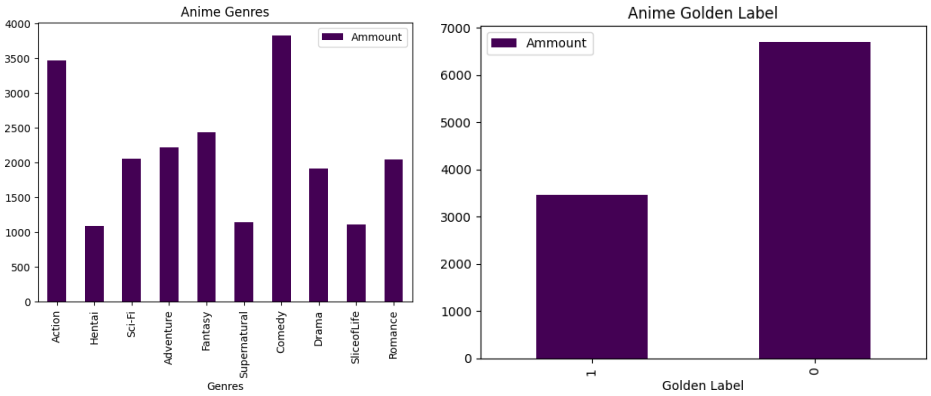
– **Confusion matrix:** It is a nice way to visualize the raw performance of a MLC. It is a 2×2 matrix, defined in Figure 1.
– **Accuracy:** Is the mean of all correct labeled samples by the MLC, within the corpus:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \tag{2}$$

where TP are the true positives, TN the true negatives, FP the false positives and FN the false negatives.

– **Presicion:** Measures how effective was the successful-labeling of the MLC. It consider all animes labeled as successful and compares them with the only the correct labeled:

$$\text{Presicion} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{3}$$



(a) (b)

Fig. 4. The dataset after preprocessing but before genre and golden label balance. Hentai is the less represented with $1,092$ samples.

– **Recall:** Measures how effective the MLC labeled successful animes. It consider all successful animes of the dataset and compares them with the correctly labeled ones:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{4}$$

– **F1 score:** It is the harmonic mean of precision and recall. This is the most widespreaded score for binary classification:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Presicion} + \text{Recall}}. \tag{5}$$

3.4 K-fold Cross Validation

For training a MLC, the dataset must be splitted into a training data, used for calibrates the classifier, and the test data, used for checking the performance. K-fold cross validation consist in splitting the training data into K parts, and then to train the MLC taking one of that parts as a test set (Figure 2). Finally, we check the model by doing predictions in the test data.

4 Experiments

We created a suitable corpus, named AniSyn7, by processing an anime synopsis dataset, focusing it on the success prediction task. We vectorized them with BBTC and DependCHILD, and finally we applied SVM, NB, and LogReg for classify them. We gathered the confusion matrices, calculated accuracy and $F1$ scores, in order to study correlation between success and synopsis. See Figure 3 for the workflow.

Algorithm 1 Balancing the corpus.

procedure BALANCER(corpus in xlsc format)
  $\ell \leftarrow$ genre with less animes in the corpus
  $\#\ell \leftarrow$ ammount of animes with genre ℓ
  $dik \leftarrow \{$genre name : $\#$animes with that genre $- \#\ell\}$

  $g, G \leftarrow$ Golden label with minimum/maximum ammount of animes
  $\#g, \#G \leftarrow$ Ammount of animes with gl g/G

  while $\#G - \#g > 0$ **do**
    Victim $\leftarrow$ An anime with golden label G
    if for all genre μ of Victim, $\mathrm{dik}[\mu] > 0$ **then**
      Remove Victim from the corpus

4.1 AniSyn7 Corpus

Dataset The initial dataset was gathered from Kaggle[6], and consists in 21,460 anime series, movies and musical videos scrapped from MAL by the user Harits Fadlilah. This dataset includes genre, theme, demographics, year, format, and also includes the MAL's **weighted score** W, calculated with the follow formula[7]:

$$W = S\left(\frac{v}{v+m}\right) + C\left(\frac{m}{v+m}\right), \tag{6}$$

where S is the average score for the anime, v is the number of users who scored, m is the minimum number of scored users to get a score and C is the mean score across the entire Anime/Manga database.

Weighted score justification: A naive baseline for measuring anime success is the mean of all user scores, given by:

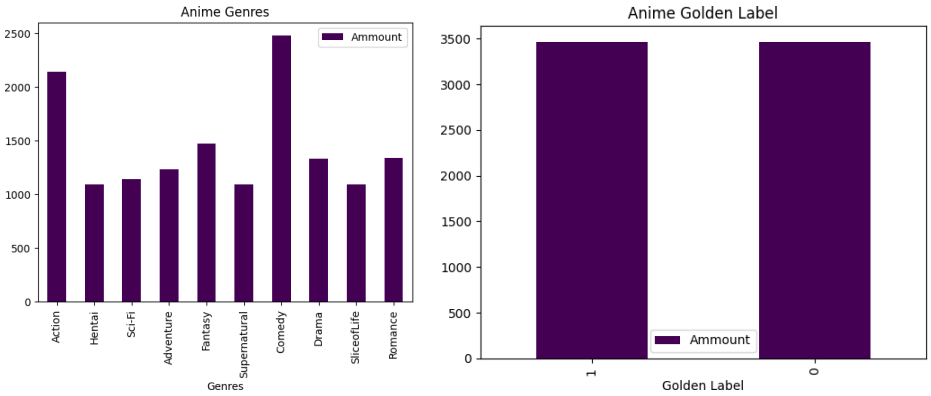
$$S = \frac{\text{Sum of all users scores of the anime}}{\text{Total ammount of users who scored the anime}}. \tag{7}$$

However, this can be biased by fake reviews, so MAL filters out votes from users who have not watched at least a fifth part of the anime. Also, they determined a minimum of $m = 50$ ratings for S to be statistically significant,[8] which also associates S with a weight, in terms of m, given by:

$$\text{Weight of } S \text{ in the score} = \left(\frac{v}{v+m}\right). \tag{8}$$

If v grows, $v/(v+m)$ tends to 1. When $v = 1$, is reached the minimum nonzero value $1/(m+1)$, who also is the statistical importance of a single review. For an anime with no human scores, MAL assigns a default score of:

[6] www.kaggle.com/datasets/harits/anime-database-2022
[7] myanimelist.net/info.php?go=topanime
[8] Value of m at least until 2023.



(a) (b)

Fig. 5. The corpus after genre and golden label balance. No Hentai was eliminated, and Comedy stayed as the most represented genre, with $2,478$ samples.

$$C = \frac{\text{Sum of all valid scores in the database}}{\text{Total ammount of valid scores in the database}}. \tag{9}$$

Who is actualized twice per day,[9] but is treated as a constant in every calculation. Analogous to S, the weight of C in the score is defined by:

$$\text{Weight of } C \text{ in the score} = \left(\frac{m}{v+m}\right). \tag{10}$$

If v grows, $m/(v+m)$ tends to 0. If $v = 0$, then C's weight is 1, which reflects the fact that MAL assigns C as a default score for an anime that has never been reviewed by any user.

Making the corpus: In order to focus on the anime success prediction task, we extracted only the follow relevant columns from the dataset: title, synopsis, score, genre, demographics, and theme. We were inspired by the multitask setting proposed in [9] who yielded better results for the book success prediction task, but we will not explicitly use it in this work. Instead, we implicitly incorporate genre information through balancing (Algorithm 1), leaving open the possibility of explicitly explore this approach in further work.

Preprocesing: We cleaned the dataset by eliminating noisy entries, according to the follow criteria:

1. **Non-informative synopsis:** There are some animes whose synopsis describes the product instead the plot, or is very ambiguous. In Table 1, are depicted three examples. We filter out such entries with a simple but effective criterion: we only included synopses with a minimum length of $1,000$ characters. While a few entries, such as Virtual-san wa Miteiru, may have slipped into the final corpus, they are not significant enough to introduce statistical noise.

[9] myanimelist.net/topanime.php

Table 3. F1 and accuracy scores for the synopsis-based anime success classification task. BBTC +LogReg/ SVM had the best performance.

Methods	TP	FP	TN	FN	Accuracy	F1
BBTC+LogReg	336	191	502	357	0.60	0.55
BBTC+SVM	342	208	485	351	0.59	0.55
DependCHILD+LogReg	279	169	524	414	0.57	0.49
DependCHILD+SVM	210	116	577	483	0.57	0.41
DependCHILD+NB	34	8	685	659	0.51	0.10
BBTC+NB	171	121	572	522	0.53	0.35
BoW+LogReg	1	0	693	692	0.50	0
BoW+SVM	1	0	693	692	0.50	0
BoW+NB	693	692	1	0	0.50	0.66

2. **Elimination of underrepresented/noisy genres:** MAL recognize several niche-focused genres, such as "Gourmet" or "Boys Love", who are very low representated in the dataset. We tackle that issue in an unwanted way, by merging them with highest represented genres. Is an unwanted measure because they are considered independent genres for a very good reason. In table 2 we explain those merges. If an anime lost all of its genres, we remove it from the dataset.

After that preprocessing, the amount of samples got reduced from $21,460$ to $10,168$. The genre proportion is depicted in Figure 4.

Golden label: We use a binary golden label as a baseline: animes with weighted score of 7 or higher are considered successful, while animes with score lower than 7 are labeled as unsuccessful. We arbitrarily decided that, guided by the fact that the dataset was splitted in an approximated proportion of 2:1 unsuccessful/successful by it (Figure 4). Moreover, the current value of C in march of 2023 is around 6.7, which is very near to 7.

Balancing the corpus: Given the preprocessed dataset with the associated binary golden labels, we implemented Algorithm 1 in order to balance it. Note that Algorithm 1 do not balance perfectly the genres, because sacrifies it in order to properly balance the golden labels, according to the value of μ. The new genres and golden label proportion after executing our implementation are depicted in Figure 5.

4.2 Experimental Setup

We splitted AniSyn7 in 80:20, having $5,542$ samples for the training data and $1,386$ samples for the test data. We used SVM, Gaussian NB and LogReg with the SciKitLearn's default hyperparameters. We made the training with 5-fold cross validation.

BoW+SVM

	P	N
P	1	693
N	0	692

BoW+LogReg

	P	N
P	1	693
N	0	692

BoW+NB

	P	N
P	693	1
N	692	0

BBTC+SVM

	P	N
P	342	485
N	208	351

BBTC+LogReg

	P	N
P	336	502
N	191	357

BBTC+NB

	P	N
P	171	572
N	121	522

DCHILD+SVM

	P	N
P	210	577
N	116	483

DCHILD+LogReg

	P	N
P	279	524
N	169	414

DCHILD+NB

	P	N
P	34	659
N	8	685

Fig. 6. Confusion matrices.

5 Results

Results are depicted in Table 3. Confusion matrices can be found in Table 6. The vectorization with the worst performance was BoW, whose MLC's only predicted a single class. It improved when we added bigrams, trigrams and character trigrams in the BBTC vectorization, who suggests that the correlation between anime success and lexical features must be highly detailed in order to be significant. DependCHILD, on the other hand, had an average performance.

However, in NB, the model tended to predict only unsuccess, who suggests that the correlation between anime failure and syntactical features could be weak. This hypothesis must to be reinforced/invalidate through better syntactical vectorizations, in a similar way that we enhanced the BoW vectorization.

6 Conclusions and Further Work

In this paper, we presented the corpus AniSyn7 for the synopsis based anime success prediction task. From a MyAnimeList dataset gathered from Kaggle, we cleaned noisy data, defined a binary Golden Label and made a genre balancing with a view to a multitask setting, for further work.

We explored this corpus with three vectorizations, the first and the second (bag of words and bag of words + bigrams + trigrams + character trigrams) based in lexical features, and the third (Counting the childs on dependency trees) based in syntactical features. Through three MLC, we obtained several insights about the statistical correlation between syntactical and lexical features with success/unsuccess.

Further work will be to consider the multitask setting, and also to turn it into a multimodal project: early stages of development also includes character and element visual designs, so we plan to introduce image vectorizations to the task. We will also enhance our text-based baseline by adding more MLC's such as K-nearest neighbour and perceptron, as well as deep learning techniques such as recurrent neural networks and convolutional neural networks.

We also plan to upgrade the experimental setup by using different values of μ. Vectorizations will be also enhanced by adding word embeddings, attention mechanisms and transformers, in order to obtain a deeper understanding about the correlation between plot and success.

References

1. AlSulaim, S. M., Qamar, A. M.: Prediction of anime series' success using sentiment analysis and deep learning. In: International Conference of Women in Data Science at Taif University, pp. 1–6 (2021) doi: 10.1109/WiDSTaif52235.2021.9430244
2. Balouchzahi, F., Shashirekha, H. L., Sidorov, G.: Cosad-code-mixed sentiments analysis for dravidian languages. In: Central Europe Workshop Proceedings, vol. 3159, pp. 887–898 (2021)
3. Buda, J., Bolonyai, F.: An ensemble model using n-grams and statistical features to identify fake news spreaders on twitter notebook for pan at clef 2020. In: Conference and Labs of the Evaluation Forum (2020)
4. Cho, H., Schmalz, M. L., Keating, S. A., Lee, J. H.: Information needs for anime recommendation: Analyzing anime users' online forum queries. In: ACM/IEEE Joint Conference on Digital Libraries (2017) doi: 10.1109/jcdl.2017.7991602
5. Field, S.: Screenplay: The foundations of screenwriting. Delta Trade Paperbacks (2005)
6. Gemeda-Yigezu, M., Tonja, A. L., Kolesnikova, O., Shahiki Tash, M., Sidorov, G., Gelbukh, A.: Word level language identification in code-mixed Kannada-English texts using deep learning approach. In: Proceedings of the 19th International Conference on Natural Language Processing (ICON): Shared Task on Word Level Language Identification in Code-mixed Kannada-English Texts, Association for Computational Linguistics, pp. 29–33 (2022)
7. Kim, Y. J., Cheong, Y. G., Lee, J. H.: Prediction of a movie's success from plot summaries using deep learning models. In: Proceedings of the Second Workshop on Storytelling, Association for Computational Linguistics, pp. 127–135 (2019) doi: 10.18653/v1/W19-3414
8. Maharjan, S., Kar, S., Montes-y-Gómez, M., González, F. A., Solorio, T.: Letting emotions flow: Success prediction by modeling the flow of emotions in books. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, vol. 2, pp. 259–265 (2018) doi: 10.18653/v1/n18-2042
9. Maharjan, S., Ovalle, J. E. A., Montes-y-Gómez, M., González, F. A., Solorio, T.: A multi-task approach to predict likability of books. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, vol. 1, pp. 1217–1227 (2017) doi: 10.18653/v1/e17-1114
10. Martín-del Campo-Rodríguez, C., Pérez-Alvarez, D. A., Maldonado-Sifuentes, C. E., Sidorov, G., Batyrshin, I., Gelbukh, A.: Authorship attribution through punctuation n-grams and averaged combination of SVM. Proceedings of the Conference and Labs of the Evaluation Forum, pp. 9–12 (2019)

11. Nuurshadieq, Wibowo, A. T.: Leveraging side information to anime recommender system using deep learning. In: 3rd International Seminar on Research of Information Technology and Intelligent Systems, pp. 62–67 (2020) doi: 10.1109/ISRITI51436.2020.9315363
12. Ojo, O. E., Gelbukh, A., Calvo, H., Feldman, A., Adebanji, O. O., Armenta-Segura, J.: Language identification at the word level in code-mixed texts using character sequence and word embedding. In: Proceedings of the 19th International Conference on Natural Language Processing (ICON): Shared Task on Word Level Language Identification in Code-mixed Kannada-English Texts, Association for Computational Linguistics, pp. 1–6 (2022)
13. Olumide Ebenezer, O., Thang-Ta, H., Gelbukh, A., Calvo, H., Sidorov, G., Oluwayemisi-Adebanji, O., Armenta-Segura, J.: Automatic hate speech detection using deep neural networks and word embedding. Computacion y Sistemas, vol. 26, no. 2, pp. 1007–1013 (2022) doi: 10.13053/CyS-26-2-4107
14. Ortiz, G., Enguix, G. B., Gómez-Adorno, H., Ameer, I., Sidorov, G.: Job offers classifier using neural networks and oversampling methods (2022) doi: 10.48550/ARXIV.2207.06223
15. Pizarro, J.: Profiling bots and fake news spreaders at PAN'19 and PAN'20 : Bots and gender profiling 2019, profiling fake news spreaders on twitter 2020. In: 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA), pp. 626–630 (2020) doi: 10.1109/DSAA49011.2020.00088
16. Reynaldi, Istiono, W.: Content-based filtering and web scraping in website for recommended anime. Asian Journal of Research in Computer Science, vol. 15, no. 2, pp. 32–42 (2023) doi: 10.9734/ajrcos/2023/v15i2318
17. Shahiki-Tash, M., Ahani, Z., Tonja, A. L., Gemeda, M., Hussain, N., Kolesnikova, O.: Word level language identification in code-mixed Kannada-English texts using traditional machine learning algorithms. In: Proceedings of the 19th International Conference on Natural Language Processing: Shared Task on Word Level Language Identification in Code-mixed Kannada-English Texts, Association for Computational Linguistics, pp. 25–28 (2022)
18. Tonja, A. L., Yigezu, M. G., Kolesnikova, O., Tash, M. S., Sidorov, G., Gelbuk, A.: Transformer-based model for word level language identification in code-mixed kannada-english texts (2022)
19. Wang, P., Yan, Y., Si, Y., Zhu, G., Zhan, X., Wang, J., Pan, R.: Classification of proactive personality: Text mining based on weibo text and short-answer questions text. IEEE Access, vol. 8, pp. 97370–97382 (2020) doi: 10.1109/ACCESS.2020.2995905

ISSN 1870-4069

Traducción automática entre lenguas indígenas de México y el español

Abdul Gafar Manuel Meque, Jason Angel,
Grigori Sidorov, Alexander Gelbukh

Instituto Politécnico Nacional,
Centro de Investigación en Computación,
México

{gafar_meque, sidorov}@cic.ipn.mx,
ajason08@gmail.com, gelbukh@gelbukh.com

Resumen. En los últimos años, hemos presenciado mejoras significativas en la precisión y velocidad de los sistemas de procesamiento de lenguaje natural. Particularmente, los métodos de traducción automática han abierto la posibilidad de conseguir mejores traducciones para lenguas de escasos recursos, tales como las lenguas indígenas de México. En este estudio usamos el modelo encoder-decoder Fairseq para evaluar la traducción automática desde el español de lenguas indígenas de México, incluyendo: el Huichol, el Mixteco, el Mazateco, el Mazahua, el Náhuatl de Guerrero y el Náhuatl de Puebla. Utilizando ROUGE y BLEU score como métricas de desempeño, nuestros resultados superan a trabajos anteriores para estas lenguas. Nuestras contribuciones incluyen la propuesta de un fuerte baseline para la evaluación de la traducción automática y la publicación de libre acceso del código y el dataset empleado.

Palabras clave: Traducción automática, lenguas indígenas, Náhuatl, Mazateco, Mixteco, Huichol, Mazahua.

Automatic Translation between Indigenous Languages of Mexico and Spanish

Abstract. In recent years, we have witnessed significant improvements in the accuracy and speed of natural language processing systems. In particular, automatic translation methods have opened the possibility of achieving better translations for low-resource languages, such as the indigenous languages of Mexico. In this study, we use the Fairseq encoder-decoder model to evaluate automatic translation from Spanish into Mexican indigenous languages, including: Huichol, Mixteco, Mazateco, Mazahua, Guerrero's Náhuatl, and Puebla's Náhuatl. Using ROUGE and BLEU score as performance metrics, our results outperform previous work for these languages. Our contributions include proposing a strong baseline for automatic translation evaluation and the open-source publication of the code and dataset used.

Keywords: Automatic translation, indigenous languages, Náhuatl, Mazateco, Mixteco, Huichol, Mazahua.

1. Introducción

La traducción automática de lenguas ha sido un tema de investigación y desarrollo en el área de Procesamiento del Lenguaje Natural durante varias décadas. Sin embargo, a pesar de las mejoras significativas en la calidad y velocidad de los sistemas de traducción automática, todavía es un desafío lograr traducciones precisas y fiables para lenguas de escasos recursos, es decir aquellas con cantidades muy limitadas de recursos digitales, como corpora, léxicos, pues al carecer de suficientes datos de entrenamiento los sistemas de traducción automática no pueden aprender con precisión los patrones linguisticos que deben utilizarse para leer o escribir en dichas lenguas.

Más aún, las lenguas de escasos recursos presentan retos significativos desde el punto de vista lingüístico para los modelos de procesamiento de lenguaje natural más efectivos de la actualidad, entre estos se evidencian la carencia de reglas ortográficas consistentes, amplias variaciones dialectales, mezcla de dialectos, neologismos en español y falta de consenso con respecto a los estándares ortográficos.

En este paper, revisamos los últimos avances en la traducción automática para lenguas de escasos recursos, y lo ejemplificamos usando 6 lenguas indígenas de México como lenguas objetivo para traducir textos de La Biblia desde el español. Nuestras contribuciones incluyen la propuesta de un fuerte baseline para la evaluación de la traducción automática y la publicación de libre acceso del código y el dataset empleado[1].

2. Antecedentes

La competencia de AmericasNLP 2021 [5] sobre traducción automática se centró en la traducción de lenguas indígenas habladas en el continente américano. El reto tuvo como objetivo promover la investigación en el área de traducción automática para idiomas de bajos recursos, particularmente aquellos con desafíos únicos como variaciones ortográficas, diferencias dialectales y falta de recursos escritos.

Durante la competencia se incluyeron diez lenguas indígenas alineadas con el español (ellas fueron: wixarika, Náhuatl, guaraní, bribri, rarámuri, aymara, shipibo-konibo, quechua, asháninka y Otomí) y se solicitó que participantes presentaran sistemas de traducción en ambas direcciones (esto es, español a lengua indígena y lengua indígena a español).

En AmericasNLP 2021 el modelo de referencia fue un modelo de secuencia a secuencia (sequence-to-sequence) implementado con Fairseq [6]. Los equipos participantes utilizaron varios enfoques, incluido el entrenamiento previo con datos monolingües, la incorporación de información fonética y el uso de modelos a nivel de caracteres.

Y aunque para la mayoría de los idiomas, muchos modelos pudieron mejorar considerablemente la línea de base, se hizo evidente la gran brecha que existe para traducir estas lenguas pues los sistemas con mejor desempeño lograron puntajes relativamente bajos en las métricas BLEU [7] y ChrF [8] respecto a los puntajes que obtienen las lenguas de con mayores recursos.

[1] huggingface.co/mekjr1

Tabla 1. Lenguas indígenas empleadas en esta investigación.

Lengua indígena	ISO 639-3	Hablantes	Estados de México
Mazahua	maz	120000	México y Michoacán
Huichol	hch	45000	Nayarit, Jalisco, Durango, y Zacatecas
Mazateco	maq	145000	Oaxaca
Mixteco	mim	490000	Guerrero
Náhuatl (de Puebla)	azz	170000	Puebla
Náhuatl (de Guerrero)	ngu	1500000	Guerrero, Puebla y Veracruz

De esta competencia se destaca que el ganador [11] logró los mejores resultados al combinar datos procedentes de La Biblia, Wikipedia y fuentes menores como constituciones politicas. Por otro lado, en cuanto a creación de corpus paralelos para lenguas indígenas, los autores en [3] describen el proyecto de creación de un corpus paralelo de español y Náhuatl junto con su interfaz de búsqueda.

El corpus se compiló a partir de libros no digitales, que presentaron varios desafíos durante el proceso de digitalización y alineación. El corpus paralelo comprende textos de diferentes fuentes que incluyen variaciones en dialecto, ortografía y cronología.

Su artículo enfatiza la escasez de recursos digitales para idiomas de bajos recursos como el Náhuatl (uno de los idiomas presentados en el trabajo actual) y cómo este corpus paralelo puede ser útil para los estudios lingüísticos y el desarrollo de tecnologías lingüísticas.

El documento proporciona ejemplos de cómo los corpus paralelos son valiosos para la traducción automática, la recuperación de textos multilingües y los estudios contrastivos y de traducción. Los autores discuten las diferencias entre las lenguas española y Náhuatl en términos de morfología, sintaxis y ortografía.

El documento está organizado en secciones que describen el proceso de compilación de los documentos paralelos, la interfaz de búsqueda, sus aplicaciones. Un trabajo similar al nuestro es [4] en el cual los autores crean un corpus paralelo entre el inglés y 5 lenguas africanas, utilizando distintos modelos neurales y el BLEU score como métrica de evaluación.

3. Fuentes de datos

Lo recursos lingüísticos utilizados en esta investigación son seis lenguas indígenas de México recuperadas de La Biblia [1], la cual es una fuente de datos bastante conocida al trabajar con lenguas de escasos recursos debido a que su contenido a sido traducido a una gran variedad de lenguas por motivos religiosos; ademas, resulta especialmente conveniente para la creación de modelos de traducción automática, pues la Biblia al estar estructurada en capítulos y versiculos permite una apropiada alineación entre los textos escritos en distintas lenguas.

A continuación se listan las lenguas indígenas empleadas en esta investigación, incluyendo el código ISO 639-3, el número de hablantes actuales y los estados de México donde principalmente se encuentran esos hablantes (Tabla 1).

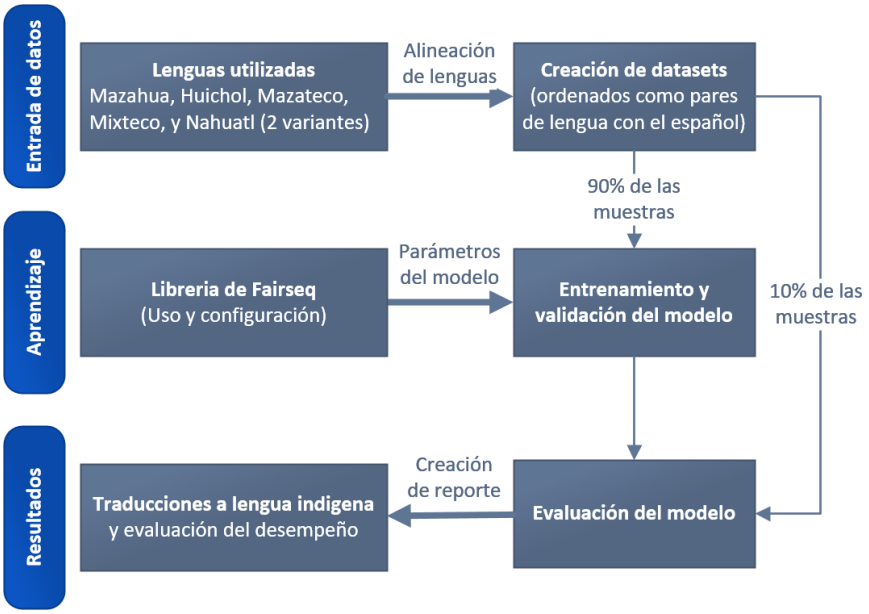


Fig. 1. Arquitectura del sistema de propuesto para implementar el baseline.

Con ello, el total de oraciones alineadas para cada lengua fue de 7930 aproximadamente, siendo el Mazateco (de Huautla de Jiménez) la lengua con menos oraciones alineadas al contar con un total de 7850 oraciones alineadas correctamente con el español.

Vale anotar además, que para esta investigación para la Biblia en español empleamos un "español simplificado" el cual utiliza un vocabulario mas sencillo que el que podría encontrarse en las biblias convenciales escritas en español, esto con el fin de facilitar las tareas de traducción desde el español hacia las lenguas indígenas (y viceversa si fuese necesario), pues como es sabido, las lenguas de escasos recursos cuentan con un vocabulario muy límitado que podria afectar el desempeño del modelo [2].

4. Método

El siguiente diagrama presenta la arquirectura del sistema implementado, el cual se divide en tres etapas: 'Entrada de datos' donde se preparan los datos para su uso en el modelo, 'Aprendizaje' donde se entrena el modelo, y 'Resultados' donde se evaluan las traducciones del modelo y se organizan los resultados generados para su posterior análisis y evaluación.

En cada una de estas etapas utilizamos Fairseq [6], una libreria escrita en PyTorch para entrenar modelos del tipo encoder-decoder que pueden ser personalizados para resolver una variedad de tareas de generación de texto, tales como traducir, resumir, parafrasear, entre otras. A continuación describimos en mayor detalle cada una de las etapas del sistema:

Tabla 2. Resultados del preprocesamiento de conjuntos de datos mediante fairseq-preprocess.

Lengua	Vocabulario	Valid UNK	Eval UNK
maz	25,480	2.41 %	2.43 %
hch	43,168	21.20 %	21.70 %
maq	22,000	1.97 %	1.81 %
mim	19,224	0.64 %	0.61 %
azz	37,632	7.31 %	7.20 %
ngu	33,296	8.80 %	8.11 %

1. **Entrada de datos:** El primer paso para crear nuestro conjunto de datos fue tomar cada una de las lenguas de interés y llevar a cabo un preprocesamiento riguroso mediante el cual se alinean los textos de La Biblia escrita en lengua indígena con La Biblia escrita en español. Este proceso se realiza utilizando los capítulos y versículos correspondientes de cada texto, lo que asegura una alineación precisa y confiable entre las dos versiones, de esta forma conseguimos representar los textos como pares de oraciones (i.e., ”texto-español, texto-lengua-indígena”). Una vez completada la alineación usamos la funcionalidad de Fairseq para preprocesar el texto antes de enviarlo al modelo de traducción automática. Este proceso implica tokenizar las palabras, segmentarlas y generar el vocabulario. El resultado final es un conjunto de datos apropiado para la tarea de traducción automática de cada lengua indígena.
2. **Aprendizaje:** Durante la etapa de aprendizaje se utilizó el 80 % de las muestras disponibles para entrenamiento y un 10 % para validación del modelo. Fairseq también permite una gran flexibilidad para personalizar los modelos generados usando distintos hiperparámetros para optimizar el desempeño del modelo. A continuación se presentan los parámetros empleados principalmente para configurar el modelo, pero puede ver la lista completa de ellos en el codigo fuente del sistema. encoder/decoder-embed-dim = 256, encoder/decoder layers = 2, dropout/attention-dropout = 0.2, learning-rate = 0.0005, optimizer adam. Finalmente, Fairseq también provee un conjunto de métricas para evaluar los resultados conseguidos, entre ellos el BLEU score que empleamos en esta investigación
3. **Resultados:** se lleva a cabo la evaluación del modelo utilizando el 10 % de las muestras disponibles. Para realizar la evaluación se emplea la métrica BLEU score, que es ampliamente utilizada en la evaluación de modelos de traducción automática. Además, se reporta el desempeño del modelo por n-gramas, desde 1-gram hasta 4-grams, lo que permite una evaluación detallada del rendimiento del modelo considerando diferentes niveles de precisión.

Cabe destacar que Fairseq también incluye una opción para manejar instancias de palabras (tokens) desconocidas, es decir, palabras que no fueron parte del vocabulario en la etapa de aprendizaje. Esta opción se llama –replace-unk y permite al usuario reemplazar tokens desconocidos con un token especial “unk” o con un token específico.

Al reemplazar tokens desconocidos con un token específico, el modelo aún puede aprender del contexto de las palabras circundantes y mejorar su capacidad para generar resultados precisos.

Tabla 3. Desempeño del modelo en los datos de validación usando las métricas ROUGE y BLEU.

Lengua	R-1	R-2	R-L	BLEU	1-gram	2-gram	3-gram	4-gram
maz	0.429	0.172	0.165	1.53	15.5	2.81	0.67	0.19
hch	0.465	0.203	0.189	2.42	17.18	3.75	1.19	0.45
maq	0.444	0.193	0.174	2.02	15.27	3.22	1.03	0.33
mim	0.538	0.251	0.196	2.43	17.61	3.97	1.27	0.4
azz	0.367	0.148	0.159	1.77	15.84	3.27	0.86	0.22
ngu	0.441	0.184	0.169	1.8	14.66	3.01	0.87	0.27

5. Análisis de resultados

En este artículo, presentamos un conjunto de datos para la traducción automática del español a seis lenguas indígenas. El conjunto de datos se preprocesó con fairseq-preprocess, siguiendo [10, 9] e informamos los resultados de referencia y las métricas de evaluación para cada par de idiomas.

5.1. Resultados del preprocesamiento del conjunto de datos

La tabla 2 muestra los resultados del preprocesamiento del conjunto de datos mediante el preprocesamiento de fairseq. La tabla informa el vocabulario de cada lengua, donde notamos que como varían entre cada una, siendo la mazahua la que tiene el vocabulario mas amplio (43,168) y el Mazateco el de menor vocabulario (22,000). También se informan los porcentajes de palabras desconocidas (UNK) para los conjuntos de de validación y evaluación con datos que van desde el 1,81 % hasta el 21,1 %.

5.2. Baseline results

La tabla 3 y la tabla 5.2 muestran los resultados de nuestro modelo de línea base en los conjuntos de validación y prueba, respectivamente, utilizando las métricas ROUGE y BLEU para medir el desempeño. Estas métricas son comúnmente utilizadas en la evaluación de modelos generativos de lenguahje, tales como resumen y traducción automática.

Especificamente, ROUGE mide la similitud entre el texto generado y el texto de referencia y para ello se consideran tres variantes: R1 que cuantifica la precisión de las palabras individuales que se superponen entre el resumen generado y el resumen de referencia, R2 mide la precisión de las secuencias de dos palabras superpuestas entre el resumen generado y el resumen de referencia, y RL mide la precisión de las secuencias de palabras superpuestas, teniendo en cuenta la longitud de la secuencia.

BLEU por otro lado mide la calidad de la traducción comparando la salida del sistema al contar el número de n-gramas en la traducción candidata que coinciden con los n-gramas en las traducciones de referencia. De manera complementaria ambas métricas proporcionan una medida cuantitativa de la calidad de la salida de los sistemas de resumen y traducción automática.

Tabla 4. Desempeño del modelo en los datos de evaluación usando las métricas ROUGE y BLEU.

Lengua	R-1	R-2	R-L	BLEU	1-gram	2-gram	3-gram	4-gram
maz	0.448	0.180	0.172	1.41	13.79	2.53	0.64	0.18
hch	0.466	0.206	0.192	2.33	17.12	3.74	1.15	0.4
maq	0.450	0.192	0.179	1.95	14.95	3.08	0.99	0.32
mim	0.525	0.247	0.192	2.62	18.74	4.25	1.32	0.45
azz	0.370	0.150	0.158	1.38	12.61	2.54	0.68	0.17
ngu	0.456	0.197	0.175	1.95	15.04	3.16	0.97	0.32

Cuanto mayor sea el valor de ROUGE o BLEU, mayor será la similitud entre el texto generado y el texto de referencia, o entre la traducción generada y la traducción de referencia, respectivamente. Como puede notarse los puntajes de las tablas 3 y 5.2 son relativamente bajos, siendo el Náhuatl de Puebla (azz) el que menor desempeño obtuvo en estos experimentos, mientras que el Mixteco y el Huichol obtuvieron los resultados mas altos.

Una posible razón podría ser el tamaño limitado del conjunto de datos. Como el número de oraciones y tokens en el conjunto de entrenamiento para cada par de idiomas es relativamente pequeño, es posible que el modelo no haya tenido suficientes datos para aprender los matices de los idiomas objetivo. Además, la complejidad y diversidad de los idiomas indígenas pueden representar un desafío significativo para los modelos de traducción automática.

Otra razón podría ser el preprocesamiento del conjunto de datos. Aunque utilizamos fairseq-preprocess, una herramienta ampliamente utilizada para el preprocesamiento de conjuntos de datos, es posible que una optimización adicional de los pasos de preprocesamiento pueda mejorar los resultados de la traducción.

Además, la calidad de las traducciones también puede verse afectada por la elección de la arquitectura del modelo, los hiperparámetros y el algoritmo de optimización. Por lo tanto, es necesario realizar más experimentos con diferentes modelos y técnicas de optimización para mejorar el rendimiento de la traducción.

Vale la pena señalar que, aunque los resultados de traducción obtenidos en nuestro estudio fueron relativamente bajos, son comparables a los informados en otros estudios con pares de lenguas de escasos recursos, como el benchmark de [5]. Por lo tanto, nuestros resultados brindan información valiosa sobre los desafíos para la traducción automática de estas lenguas y pueden informar futuras investigaciones en esta área.

6. Conclusión

La diversidad lingüística es un componente vital de la riqueza cultural de cualquier sociedad. Sin embargo, muchas lenguas indígenas están en peligro de extinción debido a factores como la globalización, la urbanización y la asimilación cultural. Con el progreso de las tecnologías para el procesamiento del lenguaje natural y su aplicación a las lenguas de escasos recursos creemos que es posible conseguir que estas lenguas no se pierdan, y con ello intentar preservar la historia y la identidad cultural de estas comunidades, lo que a su vez puede contribuir a la construcción de una sociedad más justa y equitativa.

En este artículo, creamos un conjunto de datos para la traducción automática del español a seis lenguas indígenas y evaluamos el rendimiento de la traducción utilizando las métricas de ROUGE y BLEU. Los resultados mostraron que las traducciones producidas por el modelo no fueron muy precisas, lo que indica la necesidad de una mayor mejora.

Para mejorar el rendimiento de la traducción, el trabajo futuro podría incluir el aumento del tamaño del conjunto de datos, la optimización de los pasos de preprocesamiento, la experimentación con diferentes arquitecturas de modelo, hiperparámetros y algoritmos de optimización, y la incorporación de conocimientos lingüísticos y culturales adicionales en el proceso de traducción.

En general, el desarrollo de sistemas de traducción automática efectivos para lenguas indígenas es crucial para preservar y promover la diversidad lingüística y el patrimonio cultural. Adicionalmente, como trabajo futuro planeamos agregar más lenguas indígenas considerando ademas sus respectivas variantes.

Específicamente, para el caso de México, existen 68 lenguas indígenas y según datos del Instituto Nacional de Lenguas Indígenas (INALI) se estima que existen alrededor de 364 variantes lingüísticas de estas 68 lenguas. Estas variantes reflejan la riqueza y la diversidad cultural de los pueblos indígenas de México y su patrimonio lingüístico.

Sin embargo, dada la escasez de recursos para estas lenguas, y mas aún, para sus variantes, es necesario explorar nuevas fuentes de datos que podamos incorporar para el entrenamiento de los modelos, algunas opciones incluyen: la constitución politica de México, y el uso de informacion multimodal como audio, el cual está publicamente disponible en algunos formatos de La Biblia.

Agradecimientos. Los autores agradecen al CONACYT los recursos de cómputo brindados a través de la plataforma de aprendizaje profundo para tecnologías del lenguaje del Laboratorio de Supercómputo del INAOE. Así como el uso de los recursos lingüísticos empleados en esta investigación.

Referencias

1. Bible (2023) Bible.com
2. Gu, J., Hassan, H., Devlin, J., Li, V.: Universal neural machine translation for extremely low resource languages. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, vol. 1 (2018) doi: 10.18653/v1/n18-1032
3. Gutierrez-Vasques, X., Sierra, G., Hernandez-Pompa, I.: Axolotl: A web accessible parallel corpus for Spanish-Nahuatl. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation, pp. 4210–4214 (2016)
4. Lakew, S. M., Negri, M., Turchi, M.: Low resource neural machine translation: A benchmark for five african languages. In: Proceedings of the 13th Conference on Language Resources and Evaluation, pp. 6654–6661 (2022)
5. Mager, M., Oncevay, A., Ebrahimi, A., Ortega, J., Rios, A., Fan, A., Gutierrez-Vasques, X., Chiruzzo, L., Giménez-Lugo, G., Ramos, R., Meza Ruiz, I. V., Coto-Solano, R., Palmer, A., Mager-Hois, E., Chaudhary, V., Neubig, G., Vu, N. T., Kann, K.: Findings of the AmericasNLP 2021 shared task on open machine translation for indigenous languages of

the americas. In: Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas, Association for Computational Linguistics, pp. 202–217 (2021) doi: 10.18653/v1/2021.americasnlp-1.23
6. Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., Auli, M.: fairseq: A fast, extensible toolkit for sequence modeling. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations), Association for Computational Linguistics, pp. 48–53 (2019) doi: 10.18653/v1/N19-4009
7. Papineni, K., Roukos, S., Ward, T., Zhu, W. J.: BLEU: A method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, pp. 311–318 (2001) doi: 10.3115/1073083.1073135
8. Popović, M.: chrF: Character n-gram F-score for automatic MT evaluation. In: Proceedings of the Tenth Workshop on Statistical Machine Translation, Association for Computational Linguistics (2015) doi: 10.18653/v1/w15-3049
9. Ramesh-Harsha, R., Prasad-Sankaranarayanan, K.: Neural machine translation for low resource languages using bilingual lexicon induced from comparable corpora. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, vol. 1, pp. 1748–1759 (2018)
10. Sennrich, R., Zhang, B.: Revisiting low-resource neural machine translation: A case study. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, pp. 211–221 (2019) doi: 10.18653/v1/p19-1021
11. Vázquez, R., Scherrer, Y., Virpioja, S., Tiedemann, J.: The Helsinki submission to the AmericasNLP shared task. In: Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas, Association for Computational Linguistics (2021) doi: 10.18653/v1/2021.americasnlp-1.29

Electronic edition
Available online: http://www.rcs.cic.ipn.mx

http://rcs.cic.ipn.mx

Centro de Investigación
en Computación