EDUCACIÓN
SECRETARÍA DE EDUCACIÓN PÚBLICA

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

Research in Computing Science

Vol. 153 No. 6
June 2024

Research in Computing Science

Research in Computing Science, Año 23, Volumen 153, No. 6, junio de 2024, 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 junio de 2024.

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

Research in Computing Science, year 23, Volume 153, No. 6, June 2024, 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.

Volume 153(6)

Advances in Computing Science and Applications

Juan Carlos Chimal-Eguía
Carolina Palma Preciado (eds.)

Instituto Politécnico Nacional, Centro de Investigación en Computación
México 2024

ISSN: in process

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.

Indexed in LATINDEX, DBLP and Periodica

Electronic edition

ISSN 1870-4069

Table of Contents

ISSN 1870-4069

Prediagnóstico de enfermedades respiratorias a partir de imágenes de tomografías computarizadas torácicas mediante aprendizaje profundo

Francisco Javier Aragón-González[1], Abril Valeria Uriarte-Arcia[2],
Luz María Sánchez-García[1]

[1] Instituto Politécnico Nacional,
Escuela Superior de Cómputo,
México

[2] Instituto Politécnico Nacional,
Centro de Innovación y Desarrollo Tecnológico en Cómputo,
México

faragong1300@alumno.ipn.mx, {auriartea,lmsanchez}@ipn.mx

Resumen. En esta investigación, se evaluaron tres arquitecturas de redes neuronales convolucionales: VGG16, ResNet-50 e InceptionV3, en la tarea de clasificación sobre un dataset de imágenes de tomografías computarizadas torácicas (TC). El dataset proviene de un hospital en México, cuyo nombre se omite debido a razones de privacidad y confidencialidad. Estas imágenes representaban dos clases de pacientes: pacientes que egresaron por mejora y pacientes que fallecieron debido a neumonía por COVID-19. Esta clasificación es particularmente desafiante, ya que las imágenes entre estas dos clases son similares, dado que todos los individuos están enfermos, a diferencia de estudios donde las categorías pueden ser visiblemente diferentes, por ejemplo, pacientes sanos vs. pacientes enfermos. Se realizó una búsqueda por cuadrícula (*grid search*) para determinar los mejores valores de hiperparámetros para cada modelo. También se aplicó preprocesamiento a las imágenes como segmentación del pulmón y mejora del brillo. Los resultados obtenidos indicaron que la mejora de brillo tuvo el impacto más positivo en el rendimiento general. Los mejores resultados obtenidos fueron presentados por la VGG16, alcanzando F1-score de 90% y *accuracy* de 89%. Este trabajo demuestra que las redes neuronales convolucionales tienen el potencial de beneficiar significativamente al campo de la salud. A futuro, se propone explorar imágenes con mayor profundidad tonal, pasando de 8 bits a 16 bits por canal, para capturar variaciones más sutiles, lo que podría ser esencial en contextos médicos.

Palabras clave: Prognosis de pacientes, COVID-19, Redes neuronales convolucionales.

Prediagnosis of Respiratory Diseases from Computed Thoracic Tomogram Images Using Deep Learning

Abstract. In this research, three convolutional neural network architectures were evaluated: VGG16, ResNet-50 and InceptionV3, in the classification task on a dataset of thoracic computed tomography (CT) images. The dataset comes from a hospital in Mexico, whose name is omitted due to privacy and confidentiality reasons. These images represented two classes of patients: patients who were discharged due to improvement and patients who died due to COVID-19 pneumonia. This classification is particularly challenging, since the images between these two classes are similar, given that all individuals are sick, unlike studies where the categories may be visibly different, for example, healthy patients vs. sick patients. A grid search was performed to determine the best hyperparameter values for each model. Preprocessing such as lung segmentation and brightness enhancement was also applied to the images. The results obtained indicated that the brightness improvement had the most positive impact on overall performance. The best results obtained were presented by the VGG16, reaching F1-score of 90% and accuracy of 89%. This work demonstrates that convolutional neural networks have the potential to significantly benefit the healthcare field. In the future, it is proposed to explore images with greater tonal depth, going from 8 bits to 16 bits per channel, to capture more subtle variations, which could be essential in medical contexts.

Keywords: Patient prognosis, COVID-19, convolutional neural.

1. Introducción

La neumonía es una infección respiratoria aguda que afecta a los pulmones. En México es la novena causa de mortalidad en todos los grupos etarios, ocasionada generalmente por bacterias y en menor proporción por virus y hongos; daña en particular a niños y a adultos mayores. Cuando se detecta a tiempo, es controlable; ante las primeras manifestaciones, que llegan a confundirse con una gripe [1].

Para complementar un diagnóstico, los doctores se pueden apoyar de diferentes recursos para analizar el estado de los pulmones, por ejemplo, radiografía de tórax y tomografía computarizada. La tomografía computarizada torácica (TC) tiene una mayor sensibilidad que la radiografía de tórax y permite valorar tanto la afectación pulmonar como posibles complicaciones, además de proporcionar diagnósticos alternativos [2].

Dado que la gravedad de la enfermedad implica una amenaza directa para la vida del paciente, es necesario contar con herramientas que auxilien el diagnóstico médico del paciente con alguna enfermedad respiratoria.

La utilidad de esta investigación es aportar una herramienta para ayudar al personal de la salud a complementar su diagnóstico médico, basándose en métodos y técnicas de Deep learning para la clasificación de imágenes de tomografías torácicas. Uno de los principales problemas que enfrenta el sector salud es la falta de personal para atender la gran cantidad de pacientes, ocasionando que los diagnósticos se entreguen de forma tardía y en algunos casos erróneos. Se decidió utilizar imágenes de TC en lugar de imágenes de radiografías torácicas, debido a que en las imágenes de radiografías torácicas las estructuras se sobreponen, por ejemplo, las costillas se sobreponen a los pulmones [3].

De esta forma, y resaltando que el tiempo que dedica un doctor a un paciente en ocasiones es reducido debido a la carga de pacientes, es necesario contar con una herramienta de apoyo para complementar su diagnóstico.

Para esta investigación se utilizará un dataset de imágenes TC de pacientes que ingresaron a un hospital de México con un cuadro de neumonía con sospecha de ser ocasionada por COVID-19. El dataset se empleará para la tarea de clasificación, diferenciando entre pacientes que fueron dados de alta por mejoría y pacientes que recibieron alta por fallecimiento, de esta forma se abordará un problema de predicción utilizando algoritmos de clasificación. Así, se busca detectar a los pacientes con mayor riesgo de fallecimiento para intensificar el cuidado hacia ellos. El nombre del hospital no se puede revelar por cuestiones de confidencialidad al momento de ser redactado este documento.

2. Trabajo relacionado

En [4], los autores discuten el uso exitoso de algoritmos de machine learning para el diagnóstico asistido por computadora usando imágenes médicas. Resaltan las redes neuronales convolucionales por su eficacia en clasificar lesiones y anomalías cerebrales. También introducen un método que emplea transfer learning para distinguir radiografías torácicas de personas con y sin neumonía, utilizando la red preentrenada Xception con ImageNet.

En [5], los autores presentan un sistema para identificar casos de COVID-19 usando tomografías del tórax, aplicando modelos de Ensemble Learning. Usando transfer learning, transfieren conocimientos de una red previamente entrenada a una nueva tarea. Por otro lado, emplean Stacking y Weighted Average Ensemble (WAE) combinando tres modelos: VGG19, ResNet-50 y DenseNet201. Tras entrenar con imágenes de tomografías y evaluar con varias métricas, el método WAE demostró ser el más confiable, destacando en sensibilidad, vital para la detección precisa en contextos médicos.

En [6], los autores analizan imágenes médicas usando técnicas de machine learning y redes neuronales convolucionales. Discuten métodos de detección de enfermedades aplicados a imágenes de rayos X, tomografías y resonancias, resaltando sus pros y contras. Exploran arquitecturas de Deep learning detallando su funcionamiento. Entre los modelos de Deep learning utilizados en esta investigación destacan: LeNet, AlexNet, ZfNet, VGG16, GoogleNet, ResNet-50, ResNet-101 y DenseNet. Además, enfatizan la importancia de la segmentación de imágenes médicas en la detección de tumores y su relevancia en medicina.

En [7], los autores presentan un método de detección automatizada de infecciones pulmonares usando tomografías. Para mejorar la calidad de las imágenes, aplican una técnica de ecualización y eliminan áreas no relevantes segmentando el área de los pulmones.

Utilizan la arquitectura neuronal U-Net para segmentar las imágenes y una red convolucional de tres capas para la clasificación. Implementan validación cruzada dividiendo los datos en un 70% para entrenamiento y 30% para validación. Además, con el objetivo de obtener una estimación más precisa de la capacidad del modelo en datos no vistos, se ha empleado una cuádruple validación cruzada como procedimiento de remuestreo.

En este proceso, los datos se dividen en cuatro conjuntos, utilizándose tres de ellos para el entrenamiento y el restante para la validación en cada iteración. El sistema

Tabla 1. Distribución de imágenes del dataset por clase.

Clase (Egreso)	Cantidad
Defunción	660
Mejoría	544

exhibe una notable exactitud en diversas funciones críticas, alcanzando casi la perfección en la segmentación pulmonar y la clasificación general. Además, muestra una competencia muy alta en la detección de infecciones.

Con base en la investigación de trabajo relacionado se ha definido utilizar, en esta investigación, las redes neuronales convolucionales preentrenadas **ResNet50**, **VGG16** e **InceptionV3**, pues son las más usadas para resolver problemas similares al planteado en este trabajo. Las arquitecturas se utilizarán preentrenadas para reducir el tiempo de entrenamiento.

3. Materiales y método

3.1. Banco de datos

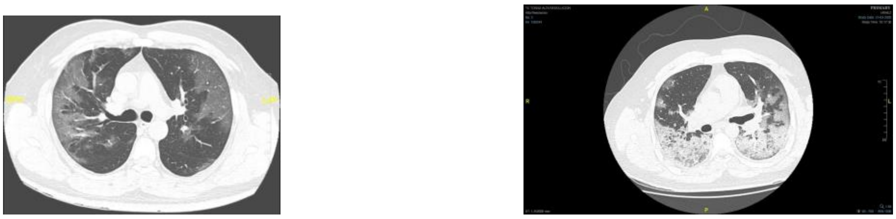
En esta investigación se utilizará un dataset de imágenes de tomografías torácicas (TC) anónimas de pacientes que ingresaron a un hospital del Estado de México, cuyo nombre no se puede revelar por cuestiones de confidencialidad. Las imágenes corresponden a tomografías torácicas de pacientes que fueron hospitalizados por un cuadro grave de neumonía bajo sospecha de COVID-19. El dataset inicial consiste en un total de **1,731 imágenes TC** nombradas con el ID del paciente, acompañadas de un archivo CSV que contiene 108 columnas con datos clínicos de cada paciente. Este archivo contiene una columna con el nombre **“Egreso”** que indica la forma en que el paciente salió del hospital. Dicha columna cuenta con 4 valores diferentes clases: “**Mejoría”, “Defunción”, “Voluntario” y “No se registró el tipo de egreso**”.

Se efectuó un preprocesamiento sobre el dataset para conservar información de interés, conservando solo los pacientes cuyo valor en esa columna fuera “**Mejoría” o “Defunción”**, siendo estas las clases que se usarán para plantear el problema de clasificación. Como resultado de este procesamiento el dataset cuenta finalmente con **1,204 imágenes**. La Tabla 1 muestra la distribución de las imágenes por cada clase.

El problema planteado en este trabajo es complejo, dado que al examinar las imágenes del dataset, se encontró que son notablemente similares entre sí, independientemente del paciente del cual provengan, puesto que todos padecen algún grado de afectación pulmonar debido a la neumonía.

Esta semejanza es tal que las variaciones entre las imágenes son mínimas. Esto contrasta con otros estudios o investigaciones, donde la tarea es clasificar entre pacientes sanos y enfermos.

En esos casos, las imágenes TC tienden a presentar diferencias más evidentes, ya que, por ejemplo, en un paciente enfermo podríamos observar anomalías como opacidades pulmonares, las cuales no estarían presentes en un individuo sano. Esta clara distinción facilita el proceso de clasificación.



(a) Imagen correspondiente a clase Defunción.

(b) Imagen correspondiente a clase Mejoría.

Fig. 1. Ejemplos de imágenes TC correspondientes a ambas clases.

Sin embargo, en nuestro caso, la homogeneidad visual de las imágenes eleva la complejidad de categorizar adecuadamente. Ver ejemplo en la Fig. 1 donde se observa que el área de los pulmones en ambas imágenes presenta opacidades que indican la presencia de afectación pulmonar.

3.2. Procesamiento digital de imágenes

Las redes convolucionales pueden verse beneficiadas del preprocesamiento de las imágenes [7], por lo que se plantea el realizar experimentos donde se apliquen técnicas de procesamiento digital de imágenes.

La primer técnica de procesamiento digital que se utilizó sobre las imágenes fue un algoritmo de segmentación automática con el objetivo de obtener la parte central de la imagen TC. Esta zona central contiene el área de interés, que es el pulmón. El algoritmo utilizado para esta tarea se encargaba de detectar el objeto más grande dentro de la imagen, calculando los contornos de los objetos y quedándose con el más grande.

La segunda técnica aplicada fue para mejorar el brillo de la imagen con el uso de la función **cv2.convertScaleAbs()**. Dicha función realiza una operación lineal en cada píxel de la imagen, definida por la relación:

$$output\ =\ \alpha \times\ input\ +\ \beta \tag{1}$$

donde:

- **input**: es el valor de intensidad del píxel original.
- **output**: es el valor de intensidad del píxel ajustado.
- α: es un factor de ganancia que controla el contraste.
- β: es un valor de desplazamiento que controla el brillo.

Factor de Ganancia α (Contraste):

- Si α> 1: aumenta el contraste de la imagen. Los valores de los píxeles se alejarán más de la intensidad media de la imagen.
- Si $\alpha\ =\ 1$: no se realiza ningún cambio en el contraste.
- Si $\alpha <\ 1$: reduce el contraste de la imagen. Los valores de los píxeles se acercarán más a la intensidad media de la imagen.
- El factor de ganancia α utilizado en este trabajo corresponde al valor de: **1.5**.

Tabla 2. *One-hot encoding* para la variable **Egreso.**

Valor Categórico	**Codificación**
Mejoría	[1 0]
Defunción	[0 1]

Valor de Desplazamiento β (Brillo):

- Si $\beta > 0$: aumenta el brillo de la imagen al añadir un valor constante a cada píxel.
- Si $\beta = 0$: no se realiza ningún cambio en el brillo.
- Si $\beta < 0$: reduce el brillo de la imagen al restar un valor constante a cada píxel.
- El valor de desplazamiento β utilizado en este trabajo corresponde al valor de: **50**.

El uso de la función **convertScaleAbs** garantiza que los valores resultantes estén en el rango válido [0, 255] para imágenes de 8 bits. Si el cálculo produce un valor por debajo de 0, se fija en 0, y si produce un valor por encima de 255, se fija en 255.

3.3. One-Hot Encoding

El método One-hot encoding permite convertir variables categóricas en valores numéricos usando vectores binarios. Esta técnica se emplea en machine learning cuando un algoritmo no puede usar valores categóricos o para mejorar resultados.

En este método, cada categoría única es representa por un vector cuyo tamaño corresponde al número de categorías, este vector tendrá el valor 1 en la posición que le corresponde a la categoría y el resto de posiciones estarán en cero.

Por ejemplo, para la variable Egreso de nuestro dataset, que tiene valores Mejoría y Defunción, se genera un vector de dos posiciones usando one-hot encoding, en donde la primera posición del vector corresponde a la categoría Mejoría y la segunda posición corresponde a la categoría Defunción. Si la categoría que se quiere codificar es Mejoría, la primera posición del vector tendrá un valor de 1 y la segunda posición un valor de cero y será lo contrario para la categoría Defunción. La Tabla 2 muestra esta codificación.

3.4. Método de validación

Como método de validación se usó **k-fold cross validation estratificado**. Este tipo de validación cruzada elimina el sesgo en la selección de datos que son usados para entrenamiento y prueba, así como garantizar que la proporcionalidad de los datos con respecto a la clase sea mantenida. Con este método los datos se dividen en k particiones, y el modelo se entrena *k* veces. En cada entrenamiento, se utilizan $k-1$ particiones como conjunto de entrenamiento y 1 partición como conjunto de prueba.

Esto se repite variando qué partes se emplean en el conjunto de entrenamiento y cuál se utiliza para el conjunto de prueba en cada ejecución. Esta técnica permite calcular métricas de rendimiento más precisas al minimizar el riesgo de resultados sesgados por

Tabla 3. Matriz de confusión: clase positiva = Defunción.

	Predicho	
Real	Mejoría	Defunción
Mejoría	VN	FP
Defunción	FN	VP

divisiones no representativas. Después de ejecutar el modelo utilizando las *k* particiones (folds), se calcula el promedio de los resultados.

3.5. Matriz de confusión y métricas

Una matriz de confusión es una herramienta utilizada en machine learning para evaluar el rendimiento de un algoritmo de clasificación. Muestra cómo se distribuyen las predicciones del modelo en comparación con los valores verdaderos.

La matriz se presenta en un formato tabular y, para una clasificación binaria, tiene cuatro componentes principales (ver Tabla 3):

- **Verdaderos Positivos (VP):** casos en los que el modelo predijo positivo y la verdadera clase también es positiva.
- **Verdaderos Negativos (VN):** casos en los que el modelo predijo negativo y la verdadera clase también es negativa.
- **Falsos Positivos (FP):** casos en los que el modelo predijo positivo pero la verdadera clase es negativa.
- **Falsos Negativos (FN):** casos en los que el modelo predijo negativo pero la verdadera clase es positiva.

Las métricas de evaluación sirven para cuantificar el rendimiento y la calidad de los modelos predictivos, especialmente en tareas de clasificación y regresión. Existen varias métricas de evaluación. A continuación, se presentan las métricas de evaluación para tareas de clasificación que serán usadas en este trabajo.

Accuracy (Exactitud): mide la proporción de predicciones correctas en el conjunto total de observaciones:

$$\text{Accuracy} = \frac{VP+VN}{VP+VN+FP+FN}.$$

Recall (Sensibilidad): indica la capacidad del modelo para identificar correctamente todas las muestras positivas reales:

$$\text{Recall} = \frac{VP}{VP+FN}.$$

Precision (Precisión): muestra la capacidad del modelo para no etiquetar como positiva una muestra que es negativa:

$$\text{Precision} = \frac{VP}{VP+FP}.$$

F1-Score: combina Precision y Recall en una única métrica, buscando un balance entre ambas:

$$\mathrm{F1-score} = 2 \times \frac{\mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}.$$

3.6. Modelos de redes neuronales convolucionales

VGG16: es una red neuronal convolucional profunda compuesta por 16 capas. Está estructurada en bloques: utiliza múltiples capas convolucionales con filtros 3x3 seguidas de una capa de max pooling. Estas capas convolucionales extraen características jerárquicas de las imágenes.

Después de procesar a través de todos los bloques convolucionales, la información se aplana y se pasa a través de tres capas totalmente conectadas para la clasificación final. La VGG16 es conocida por su simplicidad, utilizando consistentemente filtros 3x3 y su capacidad para capturar características complejas en imágenes [8].

ResNet-50: es una variante de las redes residuales o “ResNets”. La característica clave de las ResNets es el uso de conexiones residuales o “skip connections”, que permiten que las activaciones se salten una o más capas. Estas conexiones se añaden a la salida de una capa y se suman a la entrada de una capa posterior, facilitando el entrenamiento de redes muy profundas al abordar el problema de la desaparición del gradiente.

La ResNet-50 tiene 50 capas, incluyendo tanto capas convolucionales como totalmente conectadas. Se compone de bloques residuales donde cada bloque tiene tres capas convolucionales. Estas capas tienen filtros de diferentes tamaños (1x1, 3x3, 1x1) diseñados para reducir la dimensionalidad, capturar características espaciales y luego aumentar la dimensionalidad [9].

InceptionV3: es una variante de la familia de redes “Inception”, también conocidas como GoogleNet. El módulo de Inception contiene múltiples operaciones paralelas con diferentes tamaños de filtro convolucional (por ejemplo, 1x1, 3x3, 5x5) y también incluye pooling. Al combinar todas estas operaciones, la red puede capturar una variedad de características espaciales en diferentes escalas simultáneamente. InceptionV3, en particular, mejora las versiones anteriores con optimizaciones para mejorar la eficiencia y el rendimiento.

Una de las características distintivas de InceptionV3 es el uso de factorizaciones convolucionales, donde una convolución 2D se descompone en dos convoluciones 1D consecutivas, lo que reduce el número total de parámetros y acelera la red [10].

4. Diseño experimental

4.1. Flujo experimental

En este trabajo se realizaron varios experimentos con el objetivo de obtener los mejores resultados de los modelos. En la primera etapa de experimentos se ejecutaron los 3 modelos de redes convolucionales con el banco de datos original.

Con este banco de datos se ejecutó el proceso de grid search que se detalla en la sección 4.2. Una vez determinado los mejores hiperparámetros para cada modelo, se procedió a realizar la aumentación de datos, que se presenta en la sección 4.3, en cada modelo. Finalmente se selecciona el mejor modelo con base en la métrica F1-score. La

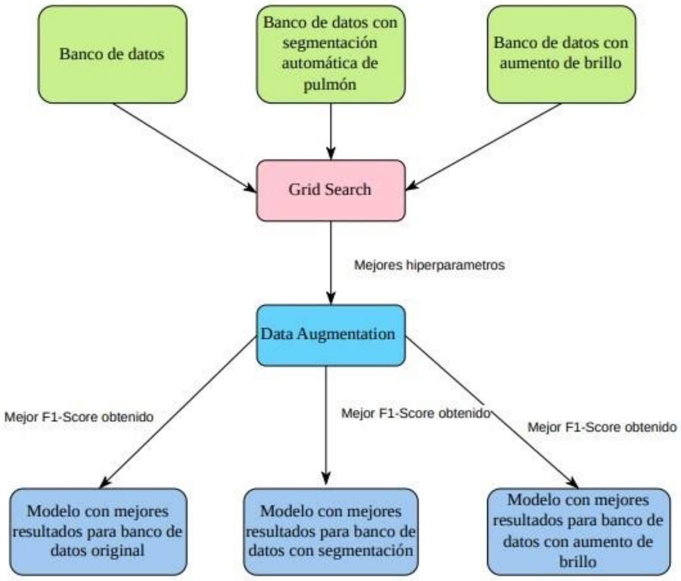

Banco de datos
Banco de datos con segmentación automática de pulmón
Banco de datos con aumento de brillo
Grid Search
Mejores hiperparametros
Data Augmentation
Mejor F1-Score obtenido
Mejor F1-Score obtenido
Mejor F1-Score obtenido
Modelo con mejores resultados para banco de datos original
Modelo con mejores resultados para banco de datos con segmentación
Modelo con mejores resultados para banco de datos con aumento de brillo

Fig. 2. Diagrama de flujo del diseño experimental.

segunda etapa de experimentación sigue el mismo flujo, pero se utiliza el banco de datos modificado mediante la técnica de segmentación que se presenta en la sección 3.2. El mismo flujo se sigue en la tercera etapa de experimentos, solo que en esta ocasión se usa el banco de datos modificado con la técnica de ajuste de brillo que también se detalla en la sección 4.3. En la Fig. 2 se muestra el flujo experimental.

4.2. Grid Search

El Grid Search o búsqueda en cuadrícula, es un método que realiza diferentes combinaciones de hiperparámetros y permite elegir la mejor combinación, es decir aquella con la que un determinado modelo produce un menor error. En este trabajo se realizó Grid Search para los modelos ResNet-50, VGG16 e InceptionV3 tomando en cuenta los hiperparámetros de learning rate y optimizador.

Se evaluaron **9 combinaciones** que se derivan de las combinaciones posibles de los valores de learning rate (0.00001, 0.0001 y 0.001) y los valores del optimizador (SGD, Adam y RMSprop). Por otro lado, la función de pérdida utilizada en todas las combinaciones y experimentos fue **nn.CrossEntropyLoss**, para evaluar la discrepancia entre las predicciones y las etiquetas reales.

Cada combinación fue probada para cada modelo mediante entrenamientos de **100 épocas**, y en cada iteración se guardaba el modelo cuando se alcanzaba el valor de pérdida más bajo en el conjunto de validación. Para concluir, se compararon los resultados de cada combinación para determinar cuál de todos tenía el valor más bajo de error para determinar la mejor combinación de hiperparámetros (ver Tabla 4).

4.3. Aumentación de datos

Una vez que identificada la mejor arquitectura con base en sus hiperparámetros mediante Grid Search, se llevaron a cabo experimentos adicionales para determinar si la incorporación de técnicas de aumentación de datos **(Data Augmentation)** podría mejorar aún más el rendimiento. Estos experimentos se realizaron usando diferentes combinaciones de los siguientes métodos de Data Augmentation: **"Random Horizontal Flip"**, **"Random Equalization"** y **"Normalization"**.

4.4. Selección del modelo

La selección del mejor modelo en cada experimento realizado se hizo con base en los resultados obtenidos por la métrica F1-Score, dado que dicha métrica toma en cuenta el promedio armónico entre la precisión (Precision) y la sensibilidad (Recall). Esta métrica es importante dado que, en aplicaciones médicas, tanto los falsos positivos como los falsos negativos pueden tener consecuencias significativas. Un falso positivo puede llevar a pruebas médicas adicionales innecesarias, por otro lado, un falso negativo puede resultar en una enfermedad o condición no diagnosticada a tiempo.

El F1-Score, al considerar tanto Precision como Recall, proporciona una métrica que penaliza ambos tipos de errores. Al buscar el mejor ajuste en esta métrica, estamos buscando un balance en el desempeño del clasificador entre Precision y Recall.

En este estudio, se designó como clase positiva a la categoría "Defunción". Esto se debe al interés particular en identificar a los pacientes con riesgo de fallecimiento.

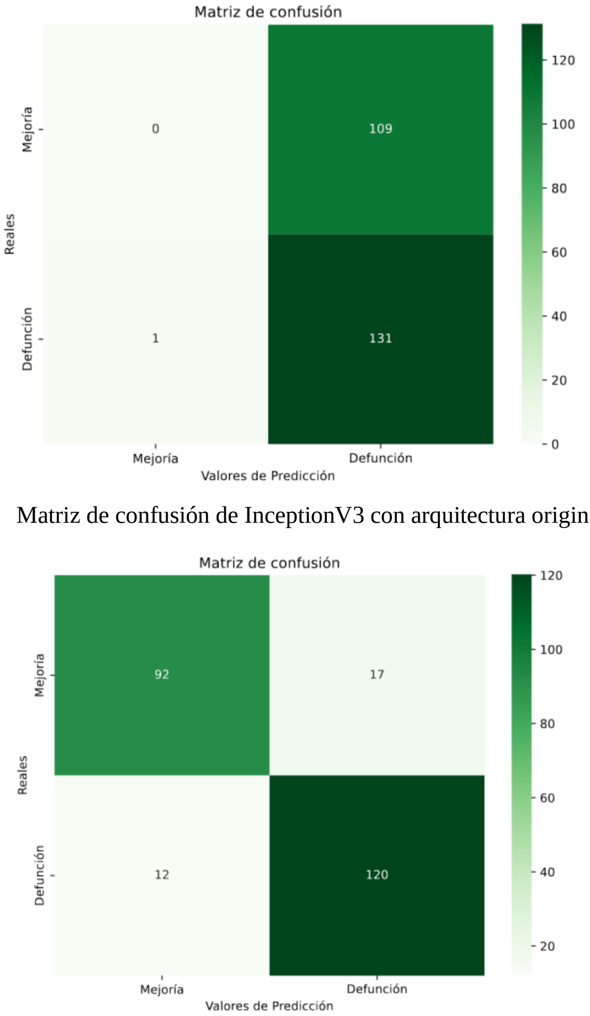
5. Resultados y discusión

5.1. Modificaciones a modelos

En este estudio, se emplearon tres modelos preentrenados de redes neuronales convolucionales: **InceptionV3, ResNet-50 y VGG16.** Durante los primeros experimentos realizados se notaron algunos problemas en el entrenamiento de 2 de los modelos, con el objetivo de mejorar los resultados se realizaron algunas modificaciones a las arquitecturas respectivas.

Para el modelo InceptionV3, los resultados de la clasificación eran muy bajos, como se puede observar en la Fig. 3a. Para mejorar estos resultados se modificó la arquitectura añadiendo tres capas completamente conectadas. Esta modificación fue necesaria debido a que la configuración original del modelo tenía dificultades para clasificar adecuadamente la categoría "Mejoría", donde las imágenes de esa categoría se estaban clasificando en la clase "Defunción" como se puede observar en los 109 Falsos Positivos que muestra la matriz de confusión. En la Fig. 3b se observan los resultados obtenidos después de realizar la modificación a la arquitectura.

Por otro lado, el modelo VGG16 mostraba problemas en el ajuste del error durante el entrenamiento, como se puede observar en la Fig. 4a, lo que hacía que el entrenamiento de la red, además de consumir más tiempo, no alcanzará un mejor desempeño. Para mejorar este comportamiento se incorporó la técnica de Batch Normalization en la arquitectura del modelo. La decisión se basó en la capacidad de



(a) Matriz de confusión de InceptionV3 con arquitectura original.

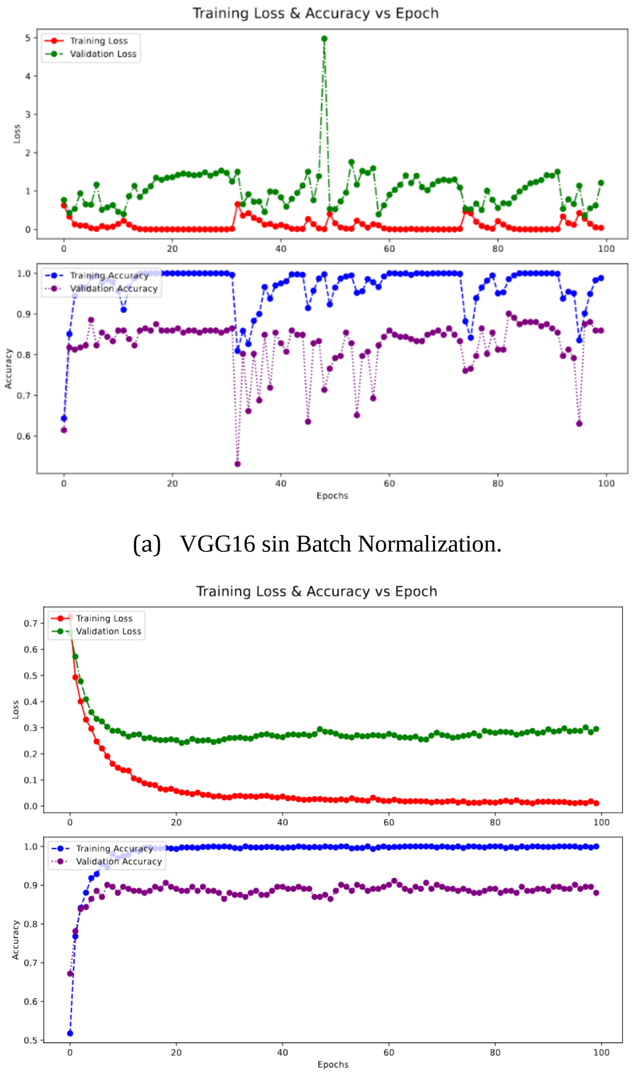
(b) Matriz de confusión de InceptionV3 con arquitectura modificada.

Fig. 3. Matrices de confusión de InceptionV3.

esta técnica para normalizar las activaciones de las capas, lo que puede promover un entrenamiento más estable y acelerar la convergencia. El comportamiento de la VGG16 se mejoró como podemos ver en la Fig. 4b.

2. Experimentos realizados con el Dataset original

Al realizar el Grid Search en los 3 modelos de redes neuronales convolucionales usando el banco de datos sin modificar, se determinaron los mejores hiperparámetros para cada arquitectura.



(a) VGG16 sin Batch Normalization.

(b) VGG16 con Batch Norma
(c) lization.

Fig. 4. Gráficas de entrenamiento VGG16. – InceptionV3: los mejores hiperparámetros para esta arquitectura dentro del espacio de búsqueda descrito en la sección 4.2 son learning rate = 0.0001 y optimizador = Adam.

Hiperparámetros obtenidos con Grid Search

- **VGG16:** los mejores hiperparámetros para esta arquitectura dentro del espacio de búsqueda descrito en la sección 4.2 son **learning rate = 0.00001** y **optimizador = Adam**.

Tabla 5. Resultados VGG16 (dataset original).

	VGG16			
	Accuracy	Precision	Recall	F1-Score
Sin Data Augmentation	0.8672	0.8672	0.8863	0.8796
H [1] y Normalización	0.8838	0.8838	0.8863	0.8931
E [1] y Normalización	0.8713	0.8713	0.8836	0.8803
H, E [1] y Normalización	0.8672	0.8672	0.8484	0.8750
H, E[1] sin Normalización	0.8796	0.8796	0.9015	0.8913

[1] H: Inversión Horizontal, E: Ecualizado.

Tabla 6. Resultados ResNet-50 (dataset original).

	ResNet-50			
	Accuracy	Precision	Recall	F1-Score
Sin Data Augmentation	0.8423	0.8423	0.8257	0.8515
H 1 y Normalización	0.8755	0.8755	0.8863	0.8863
E 1 y Normalización	0.8174	0.8174	0.8787	0.8405
H, E 1 y Normalización	0.8340	0.8340	0.7954	0.8400
H, E1 sin Normalización	0.8340	0.8340	0.8636	0.8507

[1] H: Inversión Horizontal, E: Ecualizado.

Tabla 7. Resultados InceptionV3 (dataset original).

	InceptionV3			
	Accuracy	Precision	Recall	F1-Score
Sin Data Augmentation	0.8796	0.8796	0.9090	0.8921
H 1 y Normalización	0.8838	0.8838	0.9015	0.8947
E 1 y Normalización	0.8298	0.8298	0.9166	0.8551
H, E 1 y Normalización	0.8506	0.8506	0.9015	0.8686
H, E1 sin Normalización	0.9004	0.9004	0.9242	0.9104

[1] H: Inversión Horizontal, E: Ecualizado.

- **ResNet-50:** los **mejores** hiperparámetros para esta arquitectura dentro del espacio de búsqueda descrito en la sección 4.2 son **learning rate = 0.0001** y **optimizador = Adam**.

Resultados obtenidos con Data Augmentation

Tras determinar los hiperparámetros mediante Grid Search, se sometieron los modelos a diversas estrategias de Data Augmentation previamente descritas en la sección 4.3, con el objetivo de discernir si alguna de estas técnicas podía potenciar los resultados obtenidos. Las tablas 5, 6 y 7 muestran los desempeños al aplicar Data Augmentation en los tres modelos empleados para abordar este problema.

En la Tabla 5 se muestra que la VGG16 alcanzó sus mejores resultados cuando se implementó Data Augmentation de Inversión Horizontal. En cuanto al modelo ResNet-50, mostrado en la Tabla 6, también este alcanzó su mejor desempeño al aplicar Data Augmentation mediante Inversión Horizontal.

Para el modelo InceptionV3, cuyos resultados se muestran en la Tabla 7, se observa que el modelo alcanzó su mejor rendimiento cuando se aplicaron técnicas de Inversión Horizontal, Ecualización y omitiendo la Normalización de la imagen.

5.3. Experimentos realizados con el dataset con segmentación automática de pulmón

Al realizar el Grid Search en los 3 modelos de redes neuronales convolucionales usando el banco de datos con imágenes segmentadas, se determinaron los mejores hiperparámetros para cada arquitectura.

Hiperparámetros obtenidos con Grid Search

- **VGG16:** los mejores hiperparámetros para esta arquitectura dentro del espacio de búsqueda descrito en la sección 4.2 son **learning rate = 0.00001** y **optimizador = Adam**.

- **ResNet-50:** los **mejores** hiperparámetros para esta arquitectura dentro del espacio de búsqueda descrito en la sección 4.2 son **learning rate = 0.0001** y **optimizador = Adam**.

- **InceptionV3:** los mejores hiperparámetros para esta arquitectura dentro del espacio de búsqueda descrito en la sección 4.2 son **learning rate = 0.0001** y **optimizador Adam**.

Resultados obtenidos con Data Augmentation sobre el dataset con segmentación

Tras determinar los hiperparámetros mediante Grid Search, se sometieron los modelos a diversas estrategias de Data Augmentation previamente descritas en la sección 4.3, con el objetivo de discernir si alguna de estas técnicas podía potenciar los resultados obtenidos.

Las tablas 8, 9 y 10 muestran los desempeños al aplicar Data Augmentation en los tres modelos utilizando el dataset de imágenes con segmentación.

En la Tabla 8 se observa que VGG16 obtuvo los mejores resultados sin utilizar ninguna técnica de data augmentation. El modelo ResNet-50 mostrado en la Tabla 9 alcanzó su mejor desempeño al aplicar Data Augmentation mediante Ecualización y Normalización.

Por otro lado, en la Tabla 10, se observa que el modelo InceptionV3 alcanzó mejores resultados utilizando Data Augmentation con Inversión Horizontal y Normalización.

Tabla 8. Resultados VGG16 (dataset con segmentación).

	VGG16 con Segmentación			
	Accuracy	Precision	Recall	F1-Score
Sin Data Augmentation	0.8962	0.8962	0.8863	0.9034
H [1] y Normalización	0.8755	0.8755	0.8484	0.8818
E [1] y Normalización	0.8423	0.8423	0.8560	0.8560
H, E [1] y Normalización	0.8464	0.8464	0.8181	0.8537
H, E[1] sin Normalización	0.8464	0.8464	0.8257	0.8549

[1] H: Inversión Horizontal, E: Ecualizado.

Tabla 9. Resultados ResNet-50 (dataset con segmentación).

	ResNet-50 con Segmentación			
	Accuracy	Precision	Recall	F1-Score
Sin Data Augmentation	0.7634	0.7634	0.7196	0.7692
H [1] y Normalización	0.8174	0.8174	0.7500	0.8181
E [1] y Normalización	0.8464	0.8464	0.8787	0.8624
H, E [1] y Normalización	0.6473	0.6473	0.4090	0.5595
H, E[1] sin Normalización	0.8132	0.8132	0.9015	0.8409

[1] H: Inversión Horizontal, E: Ecualizado.

Tabla 10. Resultados InceptionV3 (dataset con segmentación).

	InceptionV3 con Segmentación			
	Accuracy	Precision	Recall	F1-Score
Sin Data Augmentation	0.8091	0.8091	0.7651	0.8145
H [1] y Normalización	0.8796	0.8796	0.8863	0.8897
E [1] y Normalización	0.8257	0.8257	0.8787	0.8467
H, E [1] y Normalización	0.8091	0.8091	0.8787	0.8345
H, E[1] sin Normalización	0.8381	0.8381	0.8560	0.8528

[1] H: Inversión Horizontal, E: Ecualizado.

5.4. Experimentos con dataset con aumento de brillo

Al realizar el Grid Search en los 3 modelos de redes neuronales convolucionales usando el banco de datos con ajuste de brillo en las imágenes, se determinaron los hiperparámetros para cada arquitectura.

Hiperparámetros obtenidos con Grid Search

- **VGG16:** los mejores hiperparámetros para esta arquitectura dentro del espacio de búsqueda descrito en la sección 4.2 son **learning rate = 0.00001** y **optimizador = Adam**.

Tabla 11. Resultados VGG16 (dataset con ajuste de brillo).

	VGG16 con ajuste de brillo			
	Accuracy	Precision	Recall	F1-Score
Sin Data Augmentation	0.8796	0.8796	0.8257	0.8825
H [1] y Normalización	0.8879	0.8879	0.8636	0.8941
H [1] sin Normalización	0.8630	0.8630	0.8484	0.8715

[1] H: Inversión Horizontal.

Tabla 12. Resultados ResNet-50 (dataset con ajuste de brillo).

	ResNet-50 con ajuste de brillo			
	Accuracy	Precision	Recall	F1-Score
Sin Data Augmentation	0.7842	0.7842	0.7803	0.7984
H [1] y Normalización	0.7551	0.7551	0.9318	0.8065
H [1] sin Normalización	0.8796	0.8796	0.9015	0.8913

[1] H: Inversión Horizontal

Tabla 13. Resultados InceptionV3 (dataset con ajuste de brillo).

	InceptionV3 con ajuste de brillo			
	Accuracy	Precision	Recall	F1-Score
Sin Data Augmentation	0.7676	0.7676	0.8636	0.8028
H [1] y Normalización	0.8174	0.8174	0.8030	0.8281
H [1] sin Normalización	0.8796	0.8796	0.9015	0.8913

[1] H: Inversión Horizontal

- **ResNet-50:** los mejores hiperparámetros para esta arquitectura dentro del espacio de búsqueda descrito en la sección 4.2 son **learning rate = 0.0001** y **optimizador = Adam**.
- **InceptionV3:** los mejores hiperparámetros para esta arquitectura dentro del espacio de búsqueda descrito en la sección 4.2 son **learning rate = 0.001** y **optimizador = SGD**.

Resultados obtenidos con data augmentation sobre el conjunto con aumento de brillo

Tras determinar los hiperparámetros mediante Grid Search, se sometieron los modelos a diversas estrategias de Data Augmentation previamente descritas en la sección 4.3, con el objetivo de discernir si alguna de estas técnicas podía potenciar los resultados obtenidos.

Las tablas 11, 12 y 13 muestran los desempeños al aplicar Data Augmentation de los tres modelos utilizando el dataset de imágenes con ajuste de brillo. En la Tabla 11 se observa que VGG16 obtuvo los mejores resultados utilizando Data Augmentation con inversión horizontal. El modelo ResNet-50 mostrado en la Tabla 12 alcanzó su mejor desempeño al aplicar Data Augmentation mediante la Inversión Horizontal, Ecualización y omitiendo la Normalización de la imagen.

Tabla 14. Comparación de resultados para VGG16 y sus variantes.

Modelo	Métrica	Resultado promedio
VGG16	Accuracy	0.8970
	Precision	0.9141
	Recall	0.8969
	F1-Score	0.9046
VGG16 (segmentación)	Accuracy	0.8753
	Precision	0.8976
	Recall	0.8727
	F1-Score	0.8844
VGG16 (ajuste de brillo)	Accuracy	0.8920
	Precision	0.9085
	Recall	0.8939
	F1-Score	0.9011

En la Tabla 13, se observa que el modelo InceptionV3 alcanzó mejores resultados utilizando Data Augmentation mediante la Inversión Horizontal, Ecualización y omitiendo la Normalización de la imagen.

5.5. Resultados finales de los modelos

Las tablas 14, 15 y 16 nos presentan los mejores desempeños de cada modelo para cada dataset (original, con segmentación y ajuste de brillo) con la mejor técnica de data augmentation. Estos resultados se obtuvieron al promediar las métricas de los 5 ejecuciones sobre diferentes particiones de cada banco de datos usando el método de 5-folds cross-validation.

5.6. Discusión de resultados

La Tabla 14 presenta una comparativa de los resultados obtenidos utilizando el modelo VGG16 y sus variantes: segmentación automática del pulmón y aumento de brillo. La VGG16 con el dataset original sirve como modelo de referencia y obtuvo un F1-score de 0.9046, mientras que el Accuracy alcanzado fue de 0.8970.

La segmentación del pulmón tiene como objetivo centrarse únicamente en las regiones de interés, eliminando posibles ruidos que podrían afectar el rendimiento del modelo. A pesar de esto, el F1-score de la VGG16 con segmentación fue 0.8844, ligeramente inferior al modelo de referencia.

El Accuracy también disminuyó a 0.8753. Esto podría sugerir que la segmentación automática, aunque útil en teoría, en este caso particular no aportó una mejora significativa. Es posible que el modelo VGG16 ya era lo suficientemente robusto para manejar el ruido en las imágenes sin segmentar. El aumento de brillo puede ayudar a resaltar características sutiles en las imágenes, lo que podría ser útil especialmente en imágenes médicas donde los detalles son cruciales.

Tabla 15. Comparación de resultados para ResNet-50 y sus variantes.

Modelo	Métrica	Resultado promedio
ResNet-50	Accuracy Precision Recall F1-Score	0.8729 0.8844 0.8848 0.8844
ResNet-50 (segmentación)	Accuracy Precision Recall F1-Score	0.8089 0.8278 0.8394 0.8244
ResNet-50 (ajuste de brillo)	Accuracy Precision Recall F1-Score	0.8629 0.8878 0.8606 0.8721

Esta variante, VGG16 con ajuste de brillo, mostró una mejora en comparación con la segmentación del pulmón, obteniendo un F1-Score de 0.9011 y un Accuracy de 0.8920, por lo que el aumento de brillo podría ser una técnica de preprocesamiento más efectiva para este dataset en particular que la segmentación del pulmón.

La aplicación de diferentes técnicas de preprocesamiento y la variación de la arquitectura VGG16 con batch normalization mostró diferencias en el rendimiento. Mientras que el aumento de brillo mejoró el rendimiento en comparación con la segmentación automática del pulmón, ninguna de las técnicas superó significativamente al modelo VGG16 de referencia en términos de F1-Score.

Esto destaca la robustez del modelo VGG16 y sugiere que futuras investigaciones podrían centrarse en otras técnicas de preprocesamiento o ajustes de arquitectura para mejorar aún más el rendimiento. La Tabla 15 presenta una comparativa de los resultados obtenidos utilizando el modelo ResNet-50 y sus variantes: segmentación automática del pulmón y aumento de brillo.

La ResNet-50 con el banco de datos original obtuvo un F1-Score de 0.8844 y Accuracy de 0.8729. La variante de ResNet-50 con segmentación experimentó una disminución en todas las métricas comparadas con el modelo base, el F1-Score se redujo a 0.8244 y el Accuracy a 0.8089.

Estos resultados podrían indicar que, para el modelo ResNet-50, la segmentación automática no necesariamente aporta mejoras. Podría ser que la segmentación estuviera eliminando información útil o que ResNet-50 ya tiene la capacidad de manejar el ruido o información adicional presente en las imágenes sin segmentar.

ResNet-50 con aumento de brillo alcanzó F1-Score de 0.8721 y Accuracy de 0.8629. Aunque el aumento de brillo mostró una mejora con respecto a la segmentación del pulmón, todavía son inferiores al modelo ResNet-50 base. Estos resultados sugieren

Tabla 16. Comparación de resultados para InceptionV3 y sus variantes.

Modelo	Métrica	Resultado promedio
InceptionV3	Accuracy Precision Recall F1-Score	0.8488 0.8616 0.8666 0.8377
InceptionV3 (segmentación)	Accuracy Precision Recall F1-Score	0.8555 0.8701 0.8651 0.8673
InceptionV3 (ajuste de brillo)	Accuracy Precision Recall F1-Score	0.8787 0.9010 0.8757 0.8877

que, aunque el aumento de brillo puede aportar ciertas ventajas, no es suficiente para superar el rendimiento del modelo base.

La comparativa entre las diferentes variantes del modelo ResNet-50 revela que el modelo estándar, sin modificaciones adicionales, ofrece el mejor rendimiento en este dataset específico. Tanto la segmentación automática del pulmón como el aumento de brillo no lograron superar las métricas del modelo con el banco de datos original.

La Tabla 16 presenta una comparativa de los resultados obtenidos utilizando el modelo InceptionV3 y sus variantes: segmentación automática del pulmón y aumento de brillo. InceptionV3 con el banco de datos original presentó un F1Score de 0.8377 y Accuracy de 0.8488. Esta misma arquitectura con segmentación obtuvo un ligero aumento en el F1-Score, alcanzando 0.8673.

Además, se observa una mejora en todas las métricas en comparación con el modelo base, lo que sugiere que la segmentación automática beneficia al modelo InceptionV3 al trabajar con este dataset. La variante con ajuste de brillo mostró mejores resultados, con F1-Score de 0.8877 y un Accuracy de 0.8787. La Precision alcanzó 0.9010, lo que indica una alta proporción de verdaderos positivos en relación con los falsos positivos.

Mientras que en los modelos previamente discutidos (VGG16 y ResNet-50) no siempre se observó una mejora con las técnicas de preprocesamiento, InceptionV3 mostró un incremento en el rendimiento con ambas variantes. De hecho, el aumento de brillo demostró ser la técnica más beneficiosa para este modelo, llevando al F1-Score más alto.

Estos hallazgos subrayan cómo diferentes arquitecturas pueden responder de manera distinta a técnicas de preprocesamiento similares, y cómo es crucial experimentar y evaluar múltiples combinaciones para encontrar la mejor solución para un dataset específico.

5.7. Discusión integrada de resultados globales

Las tablas y las discusiones individuales presentadas anteriormente ofrecen una perspectiva detallada sobre el rendimiento de tres arquitecturas prominentes de redes neuronales convolucionales: VGG16, ResNet-50 e InceptionV3. A lo largo de este análisis, nos enfocaremos en el F1-Score, ya que es una métrica que combina Precision y Recall, ofreciendo una visión más completa del rendimiento del modelo.

VGG16:

- 1-Score dataset original: 0.9046
- F1-Score dataset con segmentación: 0.8844
- F1-Score dataset con aumento de brillo: 0.9011

VGG16 en el dataset original mostró el F1-Score más alto entre sus variantes. Curiosamente, la segmentación, que suele ser beneficiosa para resaltar áreas de interés, redujo ligeramente el rendimiento. Sin embargo, el aumento de brillo mostró resultados casi idénticos al modelo estándar.

ResNet-50:

- F1-Score dataset original: 0.8844
- F1-Score dataset con segmentación: 0.8244
- F1-Score dataset con aumento de brillo: 0.8721

ResNet-50 en el dataset original no solo superó a sus variantes, sino que la segmentación automática redujo significativamente el F1-Score. Esto podría sugerir que ResNet-50, podría ya estar extrayendo características esenciales sin la necesidad de un preprocesamiento tan enfocado.

InceptionV3:

- F1-Score dataset original: 0.8377
- F1-Score dataset con segmentación: 0.8673
- F1-Score dataset con aumento de brillo: 0.8877

InceptionV3 mostró una tendencia interesante. Aunque su rendimiento en el dataset original fue el más bajo de los tres, las técnicas de procesamiento mostraron ser beneficiosas, con el aumento de brillo se obtuvo el mayor incremento en el F1-Score.

Al comparar los tres modelos, VGG16 demostró tener el F1-Score más alto en el dataset original. Sin embargo, es crucial notar cómo diferentes modelos responden a técnicas de procesamiento sobre el dataset. Mientras que VGG16 y ResNet-50 mostraron un rendimiento reducido con la segmentación, InceptionV3 se benefició de ella. Esto subraya la importancia de considerar el dataset, la arquitectura y las técnicas de preprocesamiento como un sistema integrado.

6. Conclusiones

Las arquitecturas de redes neuronales convolucionales, específicamente VGG16, ResNet-50 e InceptionV3, han demostrado ser herramientas potentes para la clasificación de imágenes de tomografía computarizada. Cada modelo tiene sus propias fortalezas y debilidades, y su eficacia puede variar dependiendo del tipo de procesamiento aplicado sobre el dataset.

VGG16 presentó el F1-Score más alto en el dataset original, lo que sugiere que esta arquitectura es robusta y capaz de capturar características esenciales sin la necesidad de procesamiento especializado de las imágenes.

ResNet-50 aunque tuvo un buen rendimiento, mostró una mayor sensibilidad a la segmentación en comparación con las otras arquitecturas, lo que indica que las conexiones residuales podrían estar optimizadas para trabajar con la información original sin modificaciones sustanciales.

InceptionV3 a pesar de tener el rendimiento más bajo en el dataset original, mostró una notable mejora con técnicas de procesamiento, subrayando su flexibilidad y adaptabilidad.

Tras un análisis exhaustivo de los resultados, se puede inferir que las diferentes arquitecturas de redes neuronales convolucionales tienen peculiaridades que pueden influir en su rendimiento en función de la naturaleza del dataset. La VGG16, por ejemplo, es una arquitectura más compacta comparada con InceptionV3. Esta característica podría hacerla más apta cuando se requiere un costo computacional más bajo.

Por otro lado, arquitecturas como ResNet-50 e InceptionV3, al ser más profundas y complejas, podrían ser más adecuadas para conjuntos de datos con variaciones más amplias y características más intrincadas. Sin embargo, en nuestro dataset, que consta de imágenes TC de pacientes que egresaron y aquellos que fallecieron por neumonía de COVID-19, la similitud entre las imágenes podría haber desafiado la capacidad discriminativa de estas redes más complejas.

Estas observaciones indican que no existe una solución única para la clasificación de imágenes TC. La elección del modelo y las técnicas de procesamiento deben adaptarse a los datos y el contexto clínico específico. A partir de los hallazgos de nuestra investigación, queda en evidencia el potencial de las redes neuronales convolucionales como herramientas de apoyo a los sistemas de salud.

Los resultados conseguidos no alcanzan una cifra superior al 90% en términos de Accuracy y F1-Score, estos valores subrayan la necesidad de continuar con experimentaciones variadas con el objetivo de reducir el margen de error. Es esencial ampliar los métodos y técnicas implementados para lograr una mayor precisión en los diagnósticos y potenciar aún más la contribución de la inteligencia artificial en el ámbito médico.

7. Trabajo a futuro

Mencionamos las siguientes direcciones de trabajo futuro.

- **Técnicas de procesamiento digital:** aunque se exploraron la segmentación y el ajuste de brillo, existen muchas otras técnicas de procesamiento, como la ampliación, la rotación y morfología matemática, que podrían mejorar el rendimiento de la clasificación.

– **Modelos más recientes:** explorar arquitecturas otras arquitectura, como EfficientNet o Vision Transformers, que podrían ofrecer mejoras en términos de eficacia y rendimiento.
– **Redes neuronales generativas (GANs):** se puede considerar un método para **ampliar** el conjunto de datos utilizando redes neuronales generativas, como las GANs (Generative Adversarial Networks). Estas redes son capaces de crear imágenes sintéticas que, aunque no son reales, comparten características similares con las imágenes originales del conjunto de datos. Implementando GANs, podríamos generar un mayor volumen de imágenes, enriqueciendo así el dataset y posiblemente mejorando el desempeño de los modelos en entrenamientos futuros.

Además, en consideración de aspectos futuros, también se está considerando el uso de imágenes de 16 bits por las razones que se mencionan a continuación:

– **Mejor resolución de detalles:** dada la alta precisión de las imágenes a 16 bits, sería posible identificar y clasificar estructuras y patologías en la imagen con un nivel de detalle anteriormente inaccesible.
– **Preprocesamiento avanzado:** las imágenes a 16 bits permiten técnicas de preprocesamiento más sofisticadas, aprovechando el mayor rango dinámico para, por ejemplo, realzar características sutiles o correcciones de imagen.
– **Redes neuronales profundas especializadas:** diseñar y entrenar redes neuronales específicamente optimizadas para trabajar con la profundidad de 16 bits, teniendo en cuenta la naturaleza de los datos.

Considerando que las imágenes convencionales utilizan una profundidad de 8 bits por canal, lo que ofrece de 0 a 255 niveles de tonalidad, se propone explorar el uso de imágenes a 16 bits por canal, dado que una imagen a 16 bits tiene 65,536 niveles de tonalidad (desde 0 hasta 65,535). La utilización de imágenes con mayor profundidad podría potenciar la capacidad de identificación y clasificación de modelos de Deep learning en contextos donde la precisión es esencial.

Referencias

1. Romero, L.: Neumonía, novena causa de mortalidad en México (2019) https://www.gaceta.unam.mx/neumonia-novenacausa-de-mortalidad-en-mexico/
2. Chamorro, E. M., Tascón, A. D., Sanz, L. I., Vélez, S. O., Nacenta, S. B. Diagnóstico radiológico del paciente con COVID-19. Radiología, vol. 63, no. 1, pp. 56–73 (2021). doi: 10.1016/j.rx.2020.11.001
3. ¿Cuáles son los beneficios de las exploraciones por TC? https://www.radiologyinfo.org/es/info/safety-hiw_04 (2022)
4. Luján-García, J. E., Yáñez-Márquez, C., Villuendas-Rey, Y., Camacho-Nieto, O.: A transfer learning method for pneumonia classification and visualization. Applied Sciences, vol. 10, no. 8, pp. 2908 (2020) doi: 10.3390/app10082908

5. Mouhafid, M., Salah, M., Yue, C., Xia, K.: Deep ensemble learning-based models for diagnosis of Covid-19 from chest CT images. Healthcare, vol. 10, no. 1, p. 166 (2022) doi: 10.3390/healthcare10010166
6. Iqbal, S., Qureshi, A. N., Li, J., Mahmood, T.: On the analyses of medical images using traditional machine learning techniques and convolutional neural networks. Archives of Computational Methods in Engineering, vol. 30, no. 5, pp. 3173–3233 (2023) doi: 10.1007/s11831-023-09899-9
7. Mahmoudi, R., Benameur, N., Mabrouk, R., Mohammed, M. A., Garcia-Zapirain, B., Bedoui, M. H.: A deep learning-based diagnosis system for COVID-19 detection and pneumonia screening using CT imaging. Applied Sciences, vol. 12, no. 10, pp. 4825 (2022) doi: 10.3390/app12104825
8. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
9. 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, pp. 770–778 (2016)
10. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern Recognition, pp. 2818–2826 (2016)

ISSN 1870-4069

Breast Anomaly Detection Using YOLO-Based Deep Learning CAD System in Digital Mammograms: A Second Opinion Tool for Breast Cancer Diagnosis (1st Stage)

Johan Uriel García-Pérez[1], Moisés Márquez-Olivera[1],
Viridiana Hernández-Herrera[1], Laura Marrujo-García[2]

[1] Instituto Politécnico Nacional,
CIITEC,
Mexico

[2] Instituto Politécnico Nacional,
CECyT 2,
Mexico

johanuriel75@gmail.com,
{vhernándezhe,mvmarquez,lmarrujog}@ipn.mx

Abstract. Since digital mammograms is currently the most widely used tool worldwide for breast cancer detection, it is important to develop a system capable of supporting clinical decision making by detecting breast abnormalities, and thus, in conjunction with a radiologist, provide a more accurate diagnosis. Therefore, we propose to use an artificial intelligence (AI) algorithm based on convolutional neural networks (CNN) to detect breast abnormalities in digital mammograms. The analysis was performed on an integrated database of 1,213 digital mammograms obtained from a public database. Using a Hold-Out 70-30 validation method, the database was divided into two sections: training set (849 images) and validation set (364 images). Three training sessions were conducted, and each one was set up with 100 epochs: the first one consisted of batch size = 4, obtaining a maximum accuracy of 68 % in the semantic segmentation mode. The second training was performed under semantic segmentation, with batch size = 8, achieving an accuracy up to 91%. The third one was performed with batch size = 8 in segmentation mode, to the pre-processed database with a brightness and contrast enhancement filter, obtaining a maximum accuracy of 69 %. This allows us to conclude that the CNN is able to identify abnormalities in breast tissue. In addition, an increase in the accuracy and sensitivity of the CNN was observed when batch size increased, by making conditions under network was trained were factors that influenced the extraction of information from the mammograms.

Keywords: Breast cancer, abnormalities detection, computer aided detection, deep learning, you only look once, YOLO v8.

pp. 29–51; rec. 2023-10-27; acc. 2023-11-10

1 Introduction

Cancer is the leading cause of death in the world [1], and breast cancer is placed as most incidence as well as one of the most common death causes as far as cancer is concerned in women both in México, United states and worldwide [2, 3]. Is it estimated that, in 2020, about 2,261,419 new breast cancer cases in women were registered which represented approximately 24.5 % of the total new cancer cases worldwide in women (Fig. 1-a), furthermore, for this same year, of the total new registered breast cancer cases, 684,996 deaths were attributed to breast cancer, which represented 15.5 % of the total cancer deaths worldwide only in women (Fig. 1-b) [4, 5].

Breast cancer occurs when breast cells begin to grow out of control, forming a mass or conglomerate named tumour, which may be cancerous (malignant) or benign [6, 7]. Owing to early stages of breast cancer presents subclinically, currently within different clinical studies for breast cancer detection, screening mammography is one of the most common used tools for early breast cancer detection [8, 9].

Conventional screed-based mammography has given way to digital mammography, resulting in many benefits, including a simplified workflow and improved performance in certain patient subgroups, as well as the revolution in breast cancer care was witnessed by the introduction of mammography as an optimized radiographic imaging modality for the breast [10]. Mammography is a standard screening method for early breast cancer detection, however, it is really difficult for radiologist to provide accurate predictions for early detection, as it is complicated to interpret expertly due to various factors [11, 12].

In medicine, Artificial Intelligence (AI) has got two main application branches: physical and virtual, where virtual components are represented by Machine Learning (ML) or Deep Learning (DL), considering DL is developed by mathematical algorithms focused on improving learning through experience [13].

ML is dedicated to research and implement methods aiming to provide computers the ability to learn how to solve problems with explicit programming solutions, while DL is defined by multiple non-linear transformation modules combination, successfully modifying input information, achieving an internal data representation at multiple complexity and abstraction levels [14, 15].

This has made DL algorithms gain attention due to their considerable success because they can automatically learn feature representations and making feature extraction can be achieved from data without the need for prior definition by human experts, allowing this data-driven approach defining more abstract features, turning it more informative and generalisable [12, 16].

Deep learning remains within ML domain, and it is a special class of Artificial Neural Networks (ANN) that resembles the multi-layered human cognition system [8], besides, there is Convolutional Neural Network (CNN) another class of ANN that has become dominant in different computer vision tasks, including radiology [17], emphasizing in mammographic classification [18] because of its examination, recognition or image classification capability [19].

In recent years, CNNs have been applied in digital medical images classification for breast cancer detection and prediction because CNN application in breast cancer screening has gotten a significant advantage over traditional methods with respect to

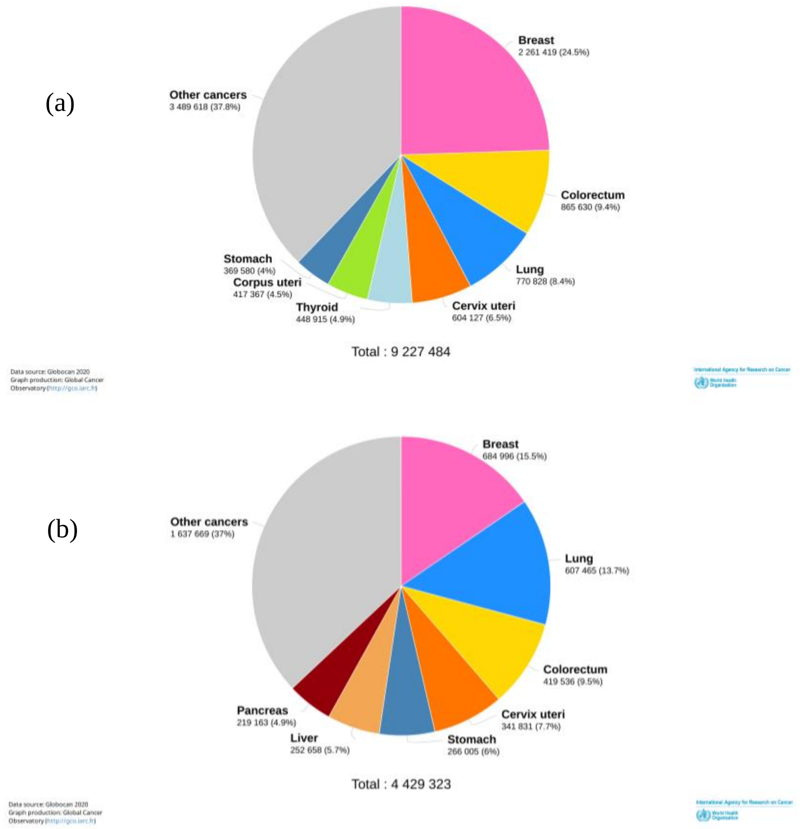


Fig. 1. Graphical schemes of 2020 mortality and incidence cancer estimations in female where a) shows the incidence number of new cases and b) the deaths number of breast cancer (GLOBOCAN 2020) [4].

the time taken to perform each test, where conventionally methods can take too long to analyse one piece of data at the time [19].

Because of breast cancer can be properly treated if an early diagnosis is correctly determined, making it feasible to have got screening methods for detecting early breast cancer signs [20], allowing AI to be an interesting factor for its application to support the detection of masses or microcalcifications present in breast tissue that can be visualized by mammography which can assist the physician in making a breast cancer diagnosis [21].

Initially in health care, computers were used in clinical image for administrative work such as image acquisition and storage, until now, they have become indispensable components in work environment, which includes the use of Computer Aided Diagnosis (CAD) systems [22], that is an AI for which is used to assist radiologist and cut back workload [23].

Nowadays, many AI algorithms are already being used in medical field (Figure 2) [22] but the potential use that can be given to AI in breast cancer diagnosis, extends to

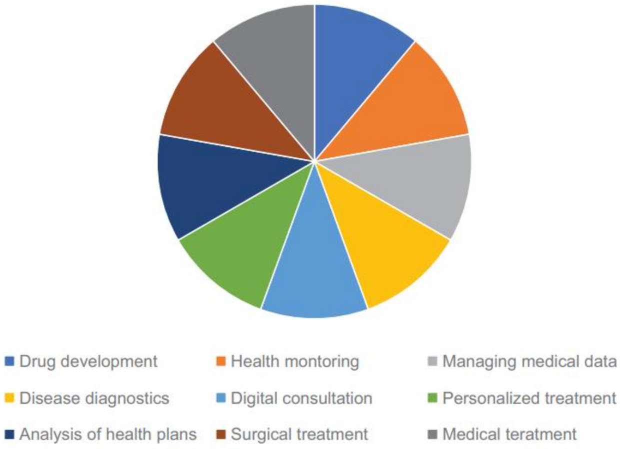


Fig. 2. Artificial Intelligence applications in health care [22].

support modalities in image interpretation and histopathology [21] because early detection can potentially improve the prognosis of breast cancer and significantly reduce mortality in women [24], therefore, CAD systems play an important role either for Computer Aided Detection (CADe), that focus on locating suspicious lesions; or Computer Aided Diagnosis (CADx), focusing on determining whether a previously detected lesion is benign or malignant [25].

Moreover, considering the current different techniques that provide digital results for breast cancer detection, DL opens a new way to be implemented in digital mammogram analysis, because it is able to integrate it for different tasks such as: injuries segmentation and classification; image generation and reconstruction; cancer risk prediction; and therapy response prediction and evaluation, where results have shown similar or better results by DL algorithm than radiologist results [23].

1.1 State-of-the-Art

Sundries researchers have proposed different CAD systems which may help for breast cancer detection or diagnosis using digital mammograms. Recent studies [26] implement an autonomous diagnosing cancer system using an integration method including CNN and image texture attribute extraction, applying a customised nine-layered CNN for categorizing in CNN stage, reaching a specificity and accuracy of 97.8 % and 98 % respectively for this method tested on MIAS (Mammographic Image Analysis Society) repository, moreover 98.8 % and 97.9 % were reached when tested on DDSM (Digital Database for Screening Mammography) repository.

On the other hand, a breast cancer image detection and a model based on convolutional and deconvolutional neural network (CDNN) was proved [27] testing the algorithm on a common dataset for ROI (Region Of Interest) segmentation, showing the model automatic classification performance improving of breast cancer which may provide a new idea about using medical diagnosis assisted by artificial intelligence. CNN method [20] was proposed to boost automatic breast cancer identification by analysing hostile ductal carcinoma tissue zones using a dataset of 275,000 images, founding the model successful due to 87 % accuracy achieved which is approximately 9 % more than ML reached, resulting as a probably option for reducing human mistakes.

Nevertheless, Rehman et al. [11] reached a 97 % score with a 2.35 and 99 % true positive ratio with 2.45 false positives per image by the Fully Connected Depthwise Separable Convolutional Neural Network (FC-DSCNN) computer-vision-based tested model on 35688 DDSM images and 2885 PINUM images respectively.

Ortíz-Rodríguez et al. [28] used image processing techniques to develop imaging biomarkers through mammographic analysis for breast cancer detection in early stages, training and testing a generalized regression ANN to classify malignant and benign tumours with a 95.83 % of accuracy reached.

Salama et al. [29] proposed a breast cancer image segmentation and classification framework including different pre-trained models, applied to MIAS, DDSM, and Curated Breast Image Subset of DDSM (CBIS-DDSM) for benign and malignant classification, showing U-Net model and InceptionV3 model the best results with 98.87 % of accuracy achieved.

Muduli et al. [30] tested a five learnable layers of four convolutional layers and a fully connected layer CNN model to facilitate automatically extraction of prominent features from mammograms and ultrasounds datasets, achieving a 96.55 %, 90.68 %, and 91.28 % accuracy in MIAS, DDSM and INbreast datasets respectively, moreover a 100 % and 89.73 % accuracy were achieved from BUS-1 and BUS-2 datasets respectively.

Agarwal et al. [31] state a patch-based CNN method for automated mass detection in Full Field Digital Mammograms (FFDM), training the model using CBIS-DDSM and INbreast datasets where InceptionV3 showed the best performance.

Since the detection and diagnosis of abnormalities in digital mammograms analysis is a still challenging task for radiologist because of necessity of analysing and identifying a “small” number of cancers that depends mostly on manual segmentation (which may take too long), computer equipment or operator [8], and even other different factors as: normal breast tissue variable appearance, overlapping tissue structures which may hide injuries in breast density tissues hindering mass detection [32, 33], breast radiographic complex structure and radiologist fatigue or distraction, that contribute to radiologist difficulties with misdiagnostic interpretations on mammograms [34, 35].

In this this paper we propose a semi-automatic system in which, the radiologist will be able to give a diagnosis by using CAD system as a support tool for abnormalities detection and then, just in few seconds, the radiologist can apply some different filters to the mammogram, analyse it, and finally, to get preliminary results that allows the radiologist to make a diagnosis, considering the AI algorithm provided results.

The use of AI algorithms in medical environment suggests that specialist can enhance the diagnosis 20 % more than a diagnosis made only by radiologist. This means the use of CAD systems will highly reduce misdiagnosis, improving diagnostic accuracy and sensitivity, decreasing radiologist workflow, visual fatigue due to mammography reading rates, and avoiding manual segmentation for greater productivity without impact diagnostic opinion [8, 36, - 39].

2 Materials and Methods

Convolutional Neural Networks employ the convolution operation as one of their layers, which perform similar operations to image processing filters [40, 41]. Convolutional layer consists of filters and image maps [42] taking an image and a small logistic regression, passing the logistic regression over the whole image [43]. Using a CNN with fewer parameters might improve significantly the time it takes to learn by reading the image "chunk-by-chunk" with the aim of allows to convolution to extract features from the input image preserving the spatial relationship between pixels [44, 42].

In CNN, the convolution operation is very similar to Gaussian and Sobel filters in image processing, because a kernel slides across an image analysing nearby pixels multiplying the weights with each aligned pixel, element-wise across the filter to finally add a bias value to the output [41, 44]. The amount the kernel shifts between pixels is called "stride" [44]. The develop procedure for the proposed CAD system for breast abnormalities detection was divided into 4 stages which included image and data acquisition, sorting and type approval; image pre-processing; model training; and finally, CAD system evaluation.

2.1 Dataset

In this study, digital mammograms were collected from Curated Breast Imaging Subset of Digital Database for Screening Mammography (CBIS-DDSM) [45] to train and validate proposed CAD system. CBIS-DDSM is an open access database which includes 10,239 images from 6,775 studies. CBIS-DDSM was analysed in order to rule out incomplete images (e.g. masks), or missing data corresponding to coordinates. Later, the resulting 1,213 images with respective data coordinates, were split according to *Hold-Out* method in a 70:30 ratios for training and validation packages, respectively.

Then, once split the dataset, is was standardized ensuring that 70 percent of training images contained the same number of images from Craniocaudal (CC) and Mediolateral (MLO) views, as well as the same number of images with and without anomalies, collecting masses according to Breast Imaging Reporting and Data System (BI-RADS), a system that allows to standardized terminology, systematize mammographic reports, and lesions categorizing, stablishing suspicion degree [46]. In addition, data number concerning to anomalies coordinates must correspond to training set of images number. Due to all images acquired from DDSM were stablished to 640x640 pixels, no more resizing processes were required. The same procedure was replicated for validation set.

2.2 Image Pre-Processing

After building the database, an adequate image pre-processing filter was carried out to improve contrast and brightness between the regions of interest (ROI) and the other sections of the mammography. In addition, performing an adequate segmentation of the mammograms will allow us to optimise the resources acquired, making image processing more efficient for the algorithm training. For the above reasons, an

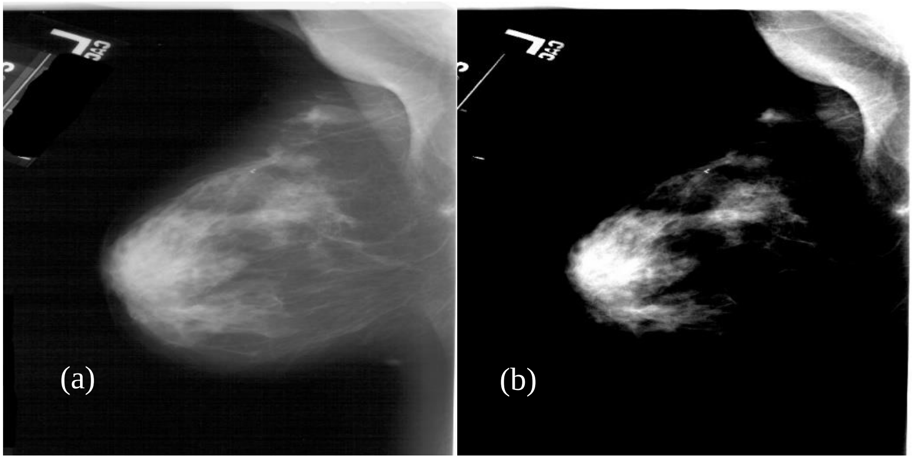


Fig. 3. Mammography comparison of left breast in MLO view where (a) is the raw mammogram from CBIS-DDSM, and (b) is the same mammogram with brightness and contrast enhancement filter.

enhancement filter image was developed in python to discard most of the fatty tissue, allowing fibrous tissue visualisation where any abnormality may be contained (Fig. 3).

2.3 You Only Look Once (YOLO) Model Training

Mass detection in breast tissue is a critical task for CAD systems [39], nevertheless, instead of developing our own deep model, an existing model will firstly use and adapted to solve our problem. Therefore, a Darknet-19 classifier model which forms the basis of real-time object detection system named YOLO [47, 48], was selected. YOLO is one of the state-of-the-art deep learning techniques [49, 47, 50] that uses a single convolutional network to whole image by dividing the input image into sub-regions and predicts multiple bounding boxes with their respective class probabilities for each region [49, 50].

YOLO is a unified model which original structure consist of 24 convolution layers, followed by 2 fully connected layers and trains on full images and directly optimizes detection performance [51, 50]. For this proposed CAD system, the YOLOv8 model released in 2023 by Ultralytics [52], was selected for abnormalities detection in breast tissue.

First, for running YOLOv8, a new environment was created to stablish a specific space to this model. The new environment was created using Anaconda prompt, where all Ultralytics packages and other needed libraries were installed. Moreover, other hyperparameters were modified within the model to focus the model on solving the abnormalities detection problem and use it the most efficient way. Once done, the resulted pre-processed Data Base of 6,065 digital mammograms (RDDSM) was used for model learning stage, where 70 % of RDDSM was used for training set realising a three-times training of 100 epochs and 0.5 confidence each.

Initial training consisted of batch size = 8 in detection mode. Second training was carried out in semantic segmentation [53] mode with batch size = 4, and final training

Table 1. Comparison between trained models to training time and testing time.

Model	Training time/epoch (s)	Testing time/image (s)
YOLO8-S4	600	4
YOLO8-S8	780	6

Table 2. Conditions for training YOLOv7 and YOLOv8 models.

Model	epoch	mode	confidence
YOLO7-S8	100	segment	0.5
YOLO8-D8	100	detect	0.5

was held in semantic segmentation mode with batch size = 8. This to visualise the difference between the performance of each model conditions in identifying abnormalities with different breast tissue conditions through mammography. The idea of using different bath sizes, it is to support the training stage, considering that a bigger batch size can accelerate it, but also requires more GPU memory which is limited to 4 GB provided by RTX 3050 Ti GPU from computer where model is trained. Using a lower bath size results in lower memory consumption but training speed could be affected.

2.4 CAD System Evaluation

Once all the previous stages were done, and proposed model were validated and tested, finally the CAD system was tested by running each of the proposed models over 20 randomly selected mammograms from mini-MIAS database to determine which of them have got the best performance on determining abnormalities trough breast density tissue. The criteria of sensitivity and accuracy evaluation are presented in equations 1-2 respectively [39, 49]:

$$Sensitivity\ (ST) = \frac{TP}{TP+FN}, \tag{1}$$

$$Accuracy\ (AC) = \frac{TP+TN}{TP+FN+TN+FP}, \tag{2}$$

where TP and FN, TN, and FP, correspond to true positive, false negative, true negative and false positive evaluated cases, respectively.

For determining TP, FN, TN, and FP classification, TP cases were considered when CAD system reached a score over 95 % on tested mammogram and it corresponded to mini-MIAS database selected coordinates; FN were considered when there were no tumours detected by CAD system, but mini-MIAS mammograms presented any tumour; TN were classified when CAD system no detect any tumours and the mini-MIAS mammograms no presented any tumour too; and FP were considered when CAD system predicted a tumour in a place that was discordant with mini-MIAS database coordinates or there were no present tumours in mini-MIAS database, to finally compare time taken while training and image prediction (Table 1).

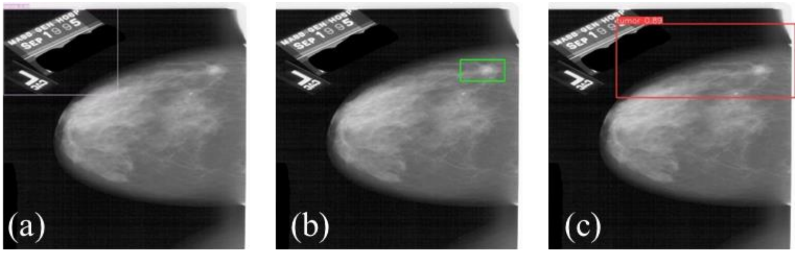


Fig. 4. Validation comparison where (a) is the segmentation mode with YOLOv7 model detecting the mammogram label inside pink bounding box; (b) is the original image with ROI visualized as a green bounding box; and (c) is the YOLOv8 detection where tumour is inside wide red bounding box.

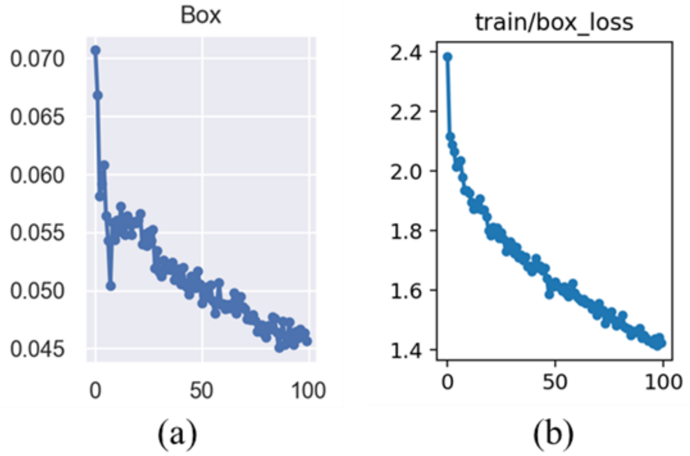


Fig. 5. Box loss training in 100 epochs for (a) YOLO7-S8 model and (b) YOLO8-D8 model.

YOLO8-D8, YOLO8-S4 and YOLO8-S8, correspond to detection mode with batch size = 8, segmentation mode with batch size = 4 and segmentation mode with batch size = 8, respectively.

3 Results

To present the results and compare the evolution between models, the YOLOv7 model was used to train under the same conditions as the YOLOv8 model, just changing to segment mode (Table 2).

Where YOLO7-S8 and YOLO8-D8 corresponds to YOLOv7 model in segmentation mode with batch size = 8, and YOLOv8 in detection mode with batch size = 8, respectively. The metric results are shown below for training and validation stages, this will allow to compare the accuracy obtained in every processed image by both models. In validation process is observed that YOLOv8 model has a higher accuracy than YOLOv7 model.

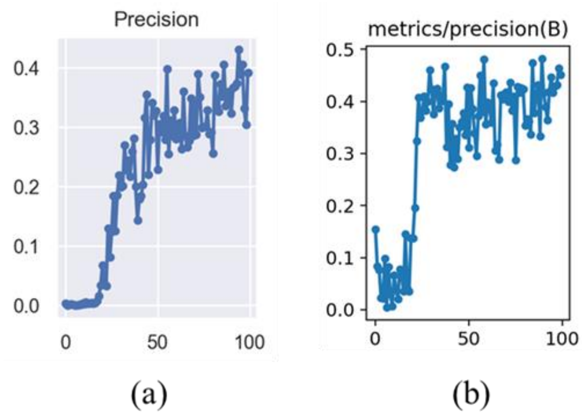


Fig. 6. Precision metrics over 100 epochs for each (a) YOLO7-S8, and (b) YOLO8-D8 models.

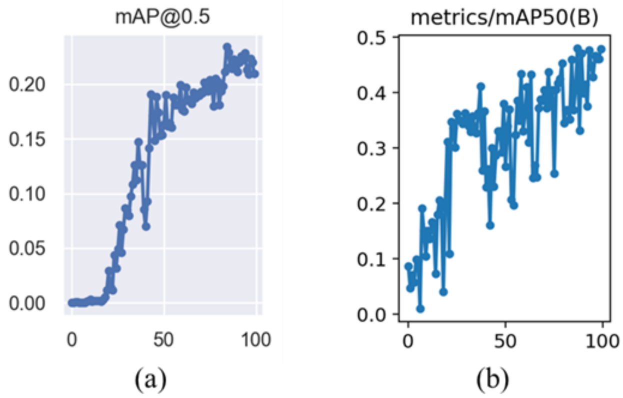


Fig. 7. mAP metrics over 100 epochs for each (a) YOLO7-S8 and (b) YOLO8-D8 models, showing the confidence increasing with every epoch realised.

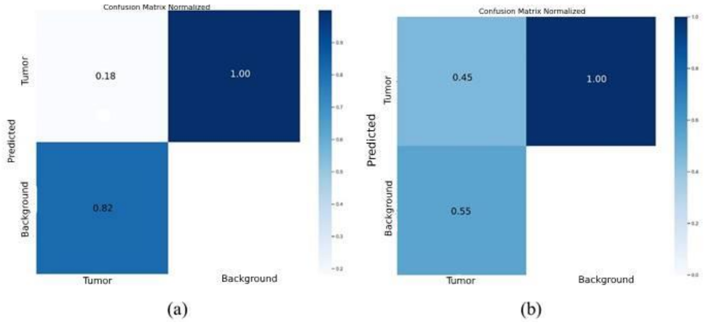


Fig. 8. Confusion matrices for (a) YOLO7-S8 and (b) YOLO8-D8 models, with 18 % accuracy and 45 % accuracy, respectively.

Even when both models have got a significant error when making prediction, the bounding box in YOLOv8 model is more accurate than YOLOv7 model (Fig. 4), as

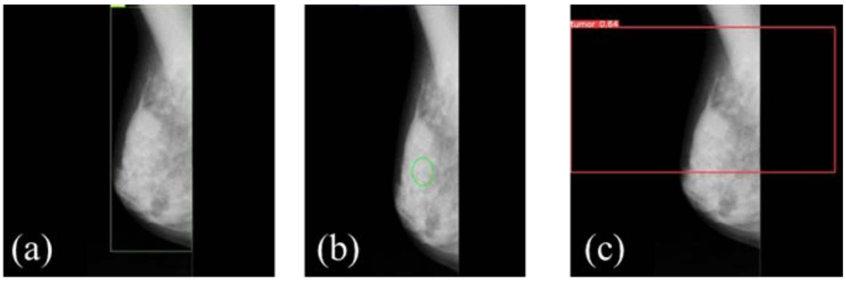


Fig. 9. Mini-MIAS database predictions for (a) YOLO7-S8 detecting all breast and pectoralis muscle inside green bounding box; (b) the original image with ROI visualized as a green bounding circle; and (c) the YOLO8-D8 prediction where tumour is inside a pretty wide red bounding box.

YOLOv7 model gets confused with mammography label, however, in YOLOv8 model although tumour is within bounding box, it is very wide, making less accurate as a support to radiologist. It was noted in training stage that errors decreased as the number of epochs performed increased (Fig. 5).

The box loss may be attributed to the model learning highly significant features through each iteration, whether it extracts values related to pixels within the segmented area, considers neighbouring pixels to compare with those found outside the segmentation boundaries, and considers features within the ROI boundary to inherit those features to the next layers.

The precision and mAP (mean Average Precision) were observed in both models, noting that both accuracy and average accuracy remain quite dispersed. Precision in YOLOv7 model, first epochs it keeps increasing, and as it approaches 50 epochs, it starts to disperse highly significantly, which may be attributed to low feature extraction trough convolution layers, which prevented it from determining where any tumours were located.

On the other hand, for model 8, initially the accuracy was mostly close, although it was low, but as the epochs were performed, a significant increase was obtained, although it was still quite sparse (Fig. 6). Nevertheless, mAP was slightly less dispersed in YOLOv7 model, however, it has quite significant increases and losses, contrary to YOLOv8 model, which has a totally dispersed mAP and even significant information loss after reaching its peak (Fig. 7).

It is observed how box and accuracy vary markedly from one model to another when certain hyperparameters are modified. Furthermore, in confusion matrices (Fig. 8) was noticed, even when YOLOv7 model was in segmentation mode, and epochs were the same, YOLOv8 model shown to be more efficient at screening in mammograms, reaching up to 0.45 % accuracy detection which is 27% more accurate than the 0.18 % reached by YOLOv7 model.

Finally, when testing for mini-MIAS database predictions, results were lower than expected, as both models presented slightly the same pattern at the validation stage, with YOLO8-D8 model being relatively closer to detecting the tumour, as it was within the delimited area, but this was too large, leaving a fairly wide error margin, as opposed to YOLO7-D8 model, which simply delimited the entire region belonging to the breast and pectoralis muscle (Fig. 9).

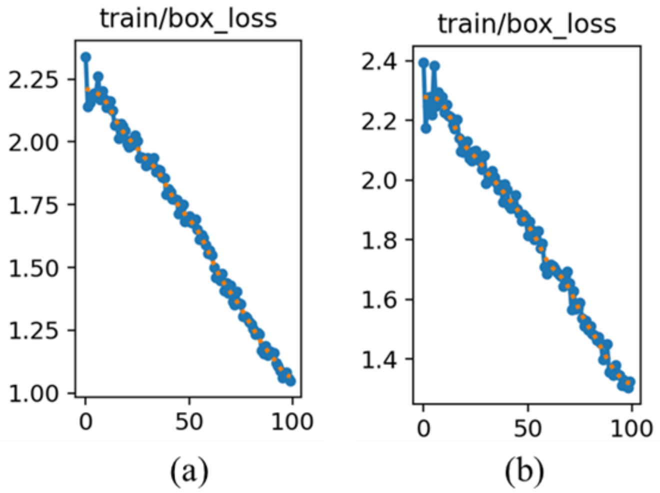


Fig. 10. Box loss during training in 100 epochs for (a) YOLO8-S4 model and (b) YOLO8-s8 model.

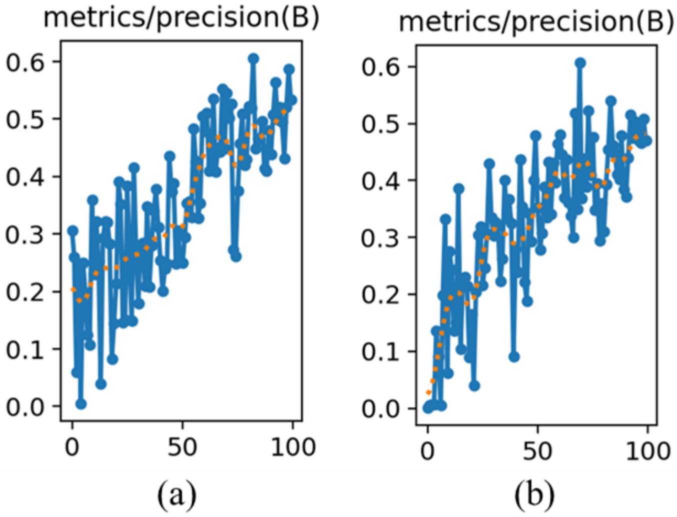


Fig. 11. Precision graphs over 100 epochs for each (a) YOLO8-S4, and (b) YOLO8-S8 models. Orange dots show the increasing accuracy of each model, being lower the YOLO8-S8 accuracy at starting point.

3.1 Comparison of Training Results

Analysing the previous results, it is possible to observe a significant efficiency with respect to the YOLOv8 model, which indicates that this model is able to extract features that allow it to differentiate the anomaly from the breast tissue.

From the resulting box loss graphs of each training for segmentation task, it can be noticed that, in both cases, the YOLO8-S4 model and the YOLO8-S8 model showed a

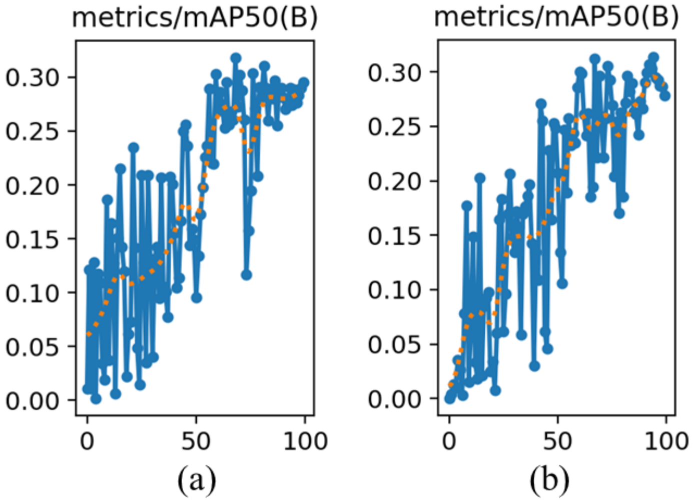


Fig. 3. mAP metrics over 100 epochs for each (a) YOLO8-S4 and (b) YOLO8-S8 models. Orange dots show the increasing confidence with every epoch realised and accuracy behaviour.

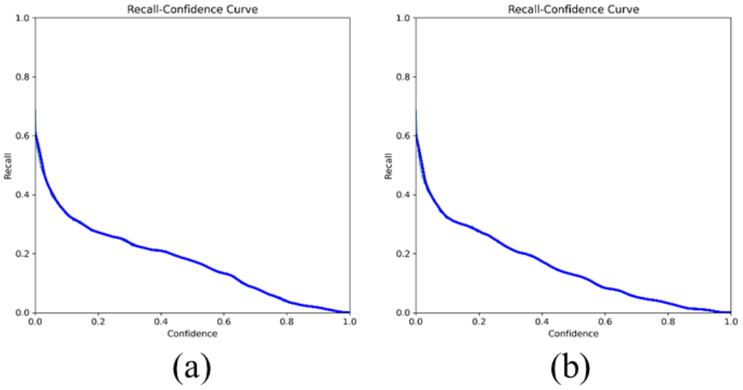


Fig. 4. Confidence function with (a) recall approximately to 60 % and confidence up to 65 % for YOLO8-S4 model and (b) 60 % recall and 65 % confidence for YOLO8-S8 model.

very low dispersion, moreover, this low dispersion behaves in a constant way during the 100 training epochs (Fig. 10), although, initially, the YOLO8-S8 model showed a higher dispersion in the first epochs.

This could imply the loss information was much lower for both models when training was carried out in segmentation mode.

As for the accuracy results (Fig. 11), it could be noted a quite significant dispersion in both models, denoting that their maximum accuracy is even below their maximum peak accuracy. In both models a slight dispersion decreasing can be seen when they are close to 100 epochs, but in the YOLO8-S8 model, the dispersion is much more noticeable. It can be observed that at beginning YOLO8-S8 has got an even lower

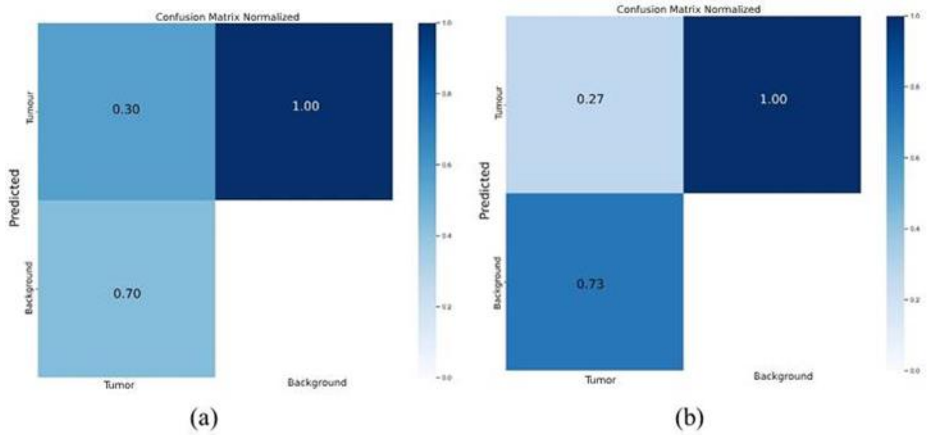


Fig. 14. Confusion matrices for (a) YOLO8-S4 and (b) YOLO8-S8 models, with 30 % accuracy and 27 % accuracy, respectively.

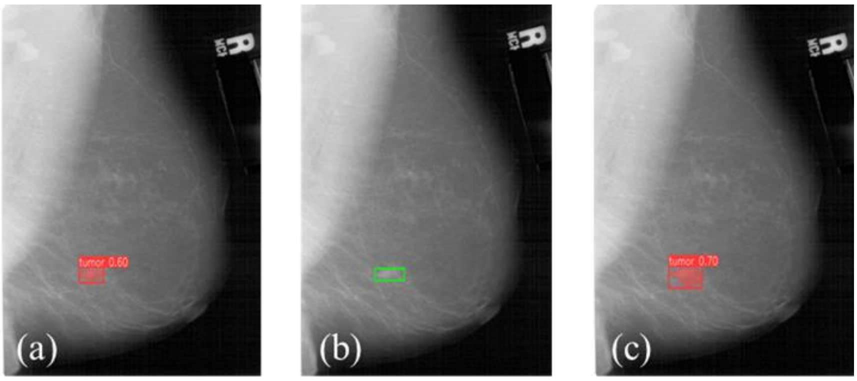


Fig. 155. Score validation comparison where (a) YOLO8-S4 reaching 60 % of detection showed inside red bounding box; (b) is the original image with ROI visualized as a green bounding box; and (c) YOLO8-S8 with 70 % of detection where tumour is inside red bounding box.

accuracy than the YOLO-S4 model (orange dots). This would imply that the extraction of highly significant features is quite complicated, which could be due to the complexity of the interpretation of the anomaly through the superimposition of the breast tissue on the mammograms. For the case of mAP, in the YOLO8-S4 model, the dispersion that exists during the learning process is observable, however, when approaching epoch 50, there is a more uniform gradual increase, but subsequently there is a quite noticeable decrease in accuracy approximately between epochs 70-80.

In the YOLO8-S8 model, the increase, although also very dispersed, is more uniformly increasing, with no significant decrease in accuracy over the 100 epochs (Fig. 12). Comparing the confidence-recall curves (Fig. 13), it is seen that difference between one model and other are too low (only 1 %), which could reflect a similar behaviour at the time of making detections in the mammograms.

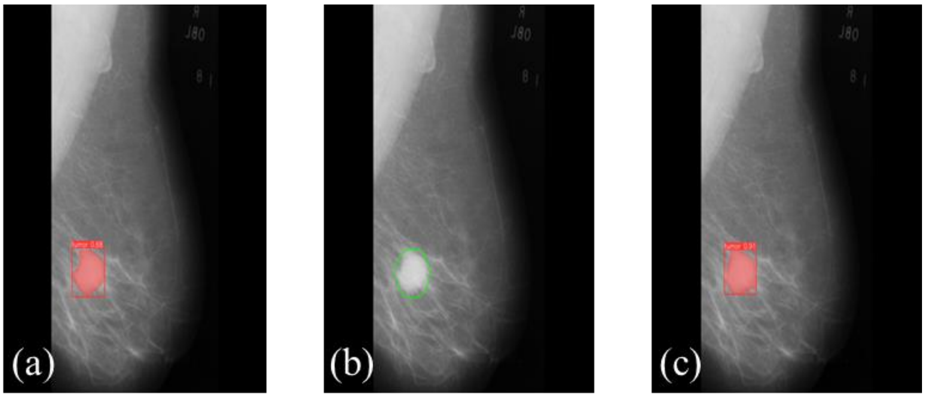


Fig. 16. Evaluation test for (a) YOLO8-S4 reaching 68 % of detection showed inside red bounding box; (b) is the original image with ROI visualized as a green bounding circle; and (c) YOLO8-S8 with 91 % of detection where tumour is segmented inside red bounding box.

Table 3. ST and AC comparison results for YOLO8-S4 and YOLO8-S8 models trough TP, FN, TN, and TP calculations.

Model	TP	FN	TN	FP	ST	AC
YOLO8-S4	15	2	1	1	0.88	0.84
YOLO8-S8	18	1		1	0.95	0.90

Furthermore, we can note the confidence of both models is moderately strong in learning, which implies both models can continue learning and could be potential candidates for future applications.

3.2 Validation Results

Contrary to what would be expected, considering the results provided by the YOLO8-D8 model, this occasion both models presented a similar behaviour during the training stage, however, in the confusion matrices (Fig. 14), it can be observed that the accuracy of both models is very close to each other, but in both cases, it is below the 45 % previously obtained with the YOLO8-D8 model, which would mean an information loss during learning stage.

As YOLO8-S4 and YOLO8-S8 models managed to obtain 30 % and 27 % of accuracy, respectively, they are apparently less efficient, even YOLO8-S8 a little bit less.

Considering the results previously observed, the YOLO8-S4 model stood out slightly in confidence and accuracy terms, with respect to the curves analysed, however, a totally opposite performance was achieved when the validation of the models in RDDSM was carried out, being the YOLO8-S8 model which managed to obtain an average of 10 % higher detecting tumours, than the other model, reaching up to 70 % of detection, with respect to the 60 % achieved by the YOLO8-S4 model (Fig. 15).

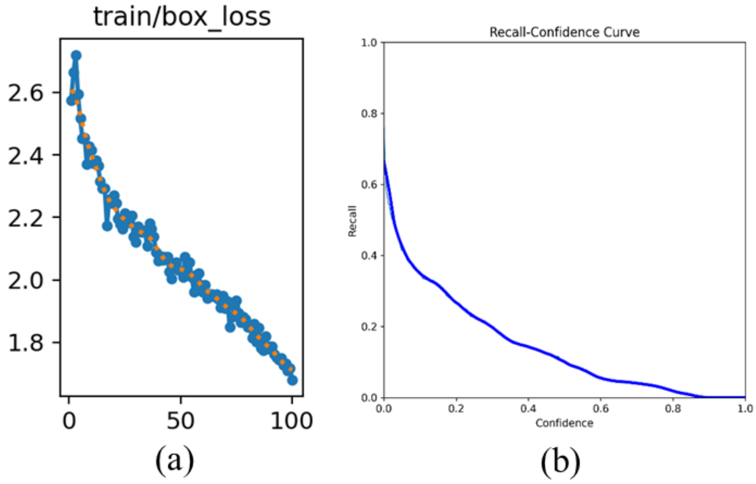


Fig. 17. YOLO8-S8 results trained on RSDDM with the enhancement filter applied where (a) represents the box loss metrics and orange dots shows the loss behaviour, and (b) represents the recall-confidence curve for the same model applied to RDDSM filtered.

3.3 CAD System Evaluation

Once previous processes were finalized, best weights were selected and used for further predictions in mini-MIAS database. YOLO8-S8 model was even higher than in the validation stage, as it achieved a score of 91 % on the mini-MIAS database, 23 % higher than the 68 % achieved by the YOLO8-S4 model (Fig. 16). Even though both models showed a lower accuracy than the previously YOLO7-S8 and YOLO8-D8 tested models, the YOLO8-S8 model, which was even less accurate, scored the highest of all models.

Finally, the comparison between the proposed models in segmentation mode is shown below (Table 3). Where TP, FN, TN, and FP, correspond to true positive, false negative, true negative and false positive evaluated cases, respectively. Moreover, St and AC corresponds to sensitivity and accuracy respectively.

As in the previous results, the YOLO8-S8 model showed a better performance during the 20 runs on mini-MIAS database, achieving a ST and AC of 95 % and 90 %, respectively. This difference can also see when individually evaluating TP, FN, TN, and FP, where YOLO8-S8 model had 18 hits in TP, one error in FN and one error in FN.

RDDSM brightness and contrast enhancement filter assessment. After testing the proposed models on the selected databases and comparing the results between each one, the model with the best performance during the training and evaluation stages was selected to be, subsequently, trained under the same conditions on the RDSSM, now with the brightness and contrast enhancement filter.

Initially, with YOLO8-S8, training stage was set at 100 epochs, with 0.5 confidence and batch size = 8. Each epoch had an approximately duration of 11 minutes. Once finished training stage, some troubles were found. In the first training results, it could be seen that box loss curve decreased steadily and with little dispersion, although in the

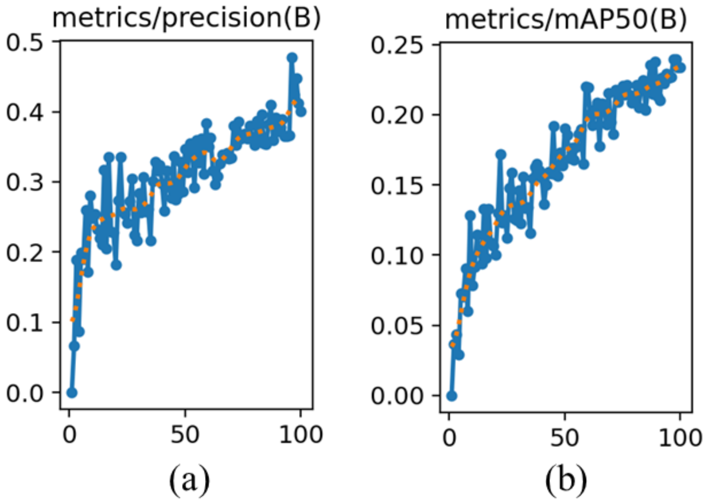


Fig. 18. (a) Precision graph and (b) mAP metrics for model applied to RDDSM filtered. Orange dots shows the increasing of accuracy foe every epoch.

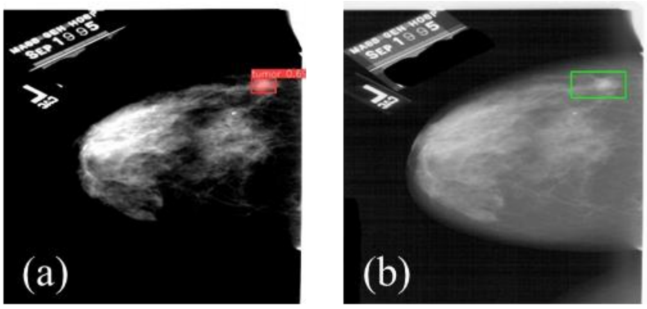


Fig. 19. Comparison results between (a) model validation in filtered mammogram from RSDDM with tumour detected inside rex bounding box with 69 %, and (b) original mammogram with tumour identified inside green bounding box.

initial periods, the decrease was slightly faster (Fig. 17-a). The confidence-recall curve (Fig. 17-b) shows similar confidence to the previous models (Fig. 13), although it is slightly higher, with a 1 % difference, reaching 66 % compared to 65 % for the previously tested models.

Subsequently, for the accuracy and mAP metrics, in both cases the initial accuracy was considered low, although in the early stages it increases rapidly. In the case of precision (Fig 18-a), there is a mostly significant dispersion before the first 50 stages, later it starts to decrease, but there is no stability when approaching the final epochs.

As for mAP, when approaching the final epochs, there seems to be a smaller dispersion, however, during the rest of the learning process, the dispersion is notorious, especially before the first 50 epochs (Fig. 18-b). With these results, it could be predicted that the model would behave similarly to that applied to the unfiltered RDDSM. The

Table 4. Metrics comparison between each model.

Model	Confusion Matrix	Box loss	Precision	mAP
YOLO7-S8	0.18	0.045	0.4	0.23
YOLO8-D8	0.45	1.4	0.48	0.48
YOLO8-S4	0.30	1.00	0.58	0.28
YOLO8-S8	0.27	1.4	0.49	0.29
YOLO8-S8F	0.24	1.4	0.4	0.24

result in the evaluation stage is shown below. Initially, system was not able to detect any tumours, but as more tests were performed on different mammograms, the system was able to identify some tumours (Fig. 19), reaching up to 69 % in detected mammograms, which could be a low accuracy compared to previous models, (Fig. 15). Even when model was tested on mini-MIAS database, this had pronounced difficulty in screening, achieving only 6 hits out of the 20 mammograms chosen for evaluation.

With the analysed results of the models applied to the raw RDDSM, the expectations increased for the evaluation of the model applied to the same database with the brightness and contrast enhancement filter, since in the images, the fibrous tissue was mostly visible, which would mean that the neural network would present fewer difficulties by not having to process irrelevant information such as areas of fatty tissue. However, the results showed that the model extracted fewer features, as in some cases, it failed to make a detection in the images, and in others, the percentage achieved was considered low.

Finally, in order to further show the performance comparison between the models, the metrics resulting from each training for each model are shown below. Where YOLO7-S8 is the YOLOv7 model in segmentation task with batch size = 8; YOLO8-D8 is YOLOv8 model in detection task with batch size = 8; YOLO8-S4 is YOLOv8 model in segmentation task with batch size = 4; YOLO8-S8 is YOLOv8 model in segmentation task with batch size = 8; and YOLO8-S8F is YOLOv8 model in segmentation task with batch size = 8 trained in filtered images.

4 Discussion

The performances obtained in each of the models during the training stages are found to be variable and, in some cases, quite scattered in terms of accuracy and learning rate. Although during the evaluation stage they showed a better performance, and even in the test phase where up to 91 % of assertiveness was achieved, this would imply that the model can be optimised by improving the images with pre-processing to avoid the overlapping of the breast tissue [49], as it was observed, the application of a brightness and contrast filter was counterproductive in terms of the observer and what was interpreted by the neural network.

Another method is by modifying the hyperparameters of the CNN and performing an in-depth analysis on the use of neurons within the network, in order to turn off those that are not being used and reduce the processing time during training and predictions.

Considering the RDDSM results, and analysing the learning curves of the models, it is possible to realise an increase in learning by augmenting data with which the network will be trained, as DL models need a considerable amount of data [49, 39] in order to be able to extract the most features through each iteration.

One of the differences most remarkable are between YOLO7-S8 and YOLO8-D8, where YOLO8-D8 model show relative better results than another model. This may due to YOLO topology, since in YOLOv8 CSPLayer was changed to C2f module, combining high-level features with contextual information to improve detection accuracy, moreover, the anchor-free used model which allows each branch to focus on its task improving model overall accuracy [52], and even YOLO does not require a complex pipeline once looked the image [49].

Instance segmentation showed better results than instance detection. Instance segmentation works by separating an example from belonging class by separating individually and comparing each labelled pixel values between segmented classes to non-segmented classes and it is useful where too many objects of same class are present and need to be differentiated. This probably affects the way model learn features, because object detection uses the pixel space given to a specific object.

Another drawback observed was the confusion of the model tested with the filtered database with the model tested with the raw images. This confusion could be due to the intensity of the brightness in the pixels related to the tumour, the dense breast tissue and the pectoral muscle [49], causing the comparison between pixel values to be depreciated as they could contain the same values and prevented the neural network from extracting any highly significant features.

5 Conclusion

In this paper, the evaluation of the Darknet-19 YOLO V8 model is presented, testing different possibilities that a CAD system can offer for mass detection. The selected model achieved up to 91 % of maximum assertiveness when tested on a publicly available database, however, it is necessary to test it on a properly constructed database in order to evaluate the model on more recent cases using more recent digital mammograms.

While it is considerable that the model needs to be trained on a much more robust database, what has been demonstrated so far provides a guideline for using CAD systems as medical assistants that can provide a second opinion, or function as medical decision support, and with the detection times determined, a decrease in workflow is foreseen, moreover, with the detection times determined, a reduction in the workflow is foreseen, as well as a reduction in the workload in hospitals whose staff capacity is affected and an improvement in the radiologist's performance by reducing visual fatigue, as a specific region is previously obtained that can be subjected to analysis, as a result of the prediction of the proposed system, avoiding the specialist in question from performing a complete reading of the mammography

6 Future Work

Future work will initially focus on data augmentation for better training of the proposed model, as well as adequate pre-processing of the images to deal with the problems of intensity and overlapping of the breast tissue. The development of a graphical interface for a doctor-computer interface is also foreseen, facilitating access to the diagnostic tool when performing mammography analysis in the medical imaging section.

7 Conflict of Interest

The authors declare no conflicts of interest.

Acknowledgments. We are pleased to mention the Centro de Investigación e Innovación Tecnológica from Instituto Politécnico Nacional (CIITEC-IPN), where we have been able to make use of the facilities, allowing us to carry out this research. We would also like to thank the Consejo Nacional de Ciencia y Tecnología (CONACyT) for the financial support granted, making possible the development of this research and future work.

References

1. Organización Mundial de la Salud.: Cólera. Boletín epidemiológico, Sistema nacional de vigilancia epidemiológica, Sistema único de información, Dirección general de epidemiología, vol. 39, no. 41, pp. 3–9 (2022)
2. American Cancer Society: How common is breast cancer? Breast Cancer Statistics (2024) https://www.cancer.org/cancer/types/breast-cancer/about/how-common-is-breast-cancer.html
3. Instituto Mexicano del Seguro Social: Epidemiología del cáncer de mama. IMMS (2022) https://www.gob.mx/imss/articulos/epidemiologia-del-cancer-de-mama-318014
4. Global Cancer Observatory: Cancer today (2024) gco.iarc.fr/today/en
5. Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., Bray, F.: Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer Journal for Clinicians, vol. 71, no. 3, pp. 209–249 (2021) doi: 10.3322/caac.21660
6. American Society of Clinical Oncology: Cáncer de mama: Introducción. American Society of Clinical Oncology Journals (2012) https://www.cancer.net/node/18093
7. Instituto Nacional del Cáncer: Prevención del cáncer de seno (mama) (PDQ®) – versión para pacientes (2013) cancer.gov/espanol/tipos/seno/paciente/prevencion-seno-pdq
8. Ting, F. F., Tan, Y. J., Sim, K. S.: Convolutional neural network improvement for breast cancer classification. Expert Systems with Applications, vol. 120, pp. 103–115 (2019) doi: 10.1016/j.eswa.2018.11.008
9. Fuchsjäger, M., Morris, E., Helbich, T.: Breast imaging: Diagnosis and intervention, Medical Radiology (2022) doi: 10.1007/978-3-030-94918-1
10. Rehman, K. U., Li, J., Pei, Y., Yasin, A., Ali, S., Mahmood, T.: Computer vision-based microcalcification detection in digital mammograms using fully connected depthwise separable convolutional neural network. Sensors, vol. 21, no. 14, pp. 4854 (2021) doi: 10.3390/s21144854

11. Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L. H., Aerts, H. J.: Artificial intelligence in radiology. Nature Reviews Cancer, vol. 18, no. 8, pp. 500–510 (2018)
12. Hamet, P., Tremblay, J.: Artificial intelligence in medicine. Metabolism, vol. 69, pp. S36–S40 (2017) doi: 10.1016/j.metabol.2017.01.011
13. Galveia, J. N., Travassos, A., Quadros, F. A., da-Silva-Cruz, L. A.: Computer aided diagnosis in ophthalmology: deep learning applications. Classification in BioApps: Automation of Decision Making, pp. 263–293 (2017) doi: 10.1007/978-3-319-65981-7_10
14. Cottrell, S. S.: A simple method for finding the scattering coefficients of quantum graphs. Journal of Mathematical Physics, vol. 56, no. 9 (2015) doi: 10.1063/1.4931082
15. Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M. A., Al-Amidie, M., Farhan, L.: Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. Journal of Big Data, vol. 8, no. 1 (2021) doi: 10.1186/s40537-021-00444-8
16. 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
17. Abdelrahman, L., Al-Ghamdi, M., Collado-Mesa, F., Abdel-Mottaleb, M.: Convolutional neural networks for breast cancer detection in mammography: A survey. Computers in biology and medicine, vol. 131, p.p. 104248 (2021) doi: 10.1016/j.compbiomed.2021.10424
18. Desai, M., Shah, M.: An anatomization on breast cancer detection and diagnosis employing multi-layer perceptron neural network (MLP) and convolutional neural network (CNN). Clinical eHealth, vol. 4, pp. 1–11 (2021) doi: 10.1016/j.ceh.2020.11.002
19. Alanazi, S. A., Kamruzzaman, M. M., Islam-Sarker, M. N., Alruwaili, M., Alhwaiti, Y., Alshammari, N., Siddiqi, M. H.: Boosting breast cancer detection using convolutional neural network. Journal of Healthcare Engineering, vol. 2021, pp. 1–11 (2021) doi: 10.1155/2021/5528622
20. Houssami, N., Kirkpatrick-Jones, G., Noguchi, N., Lee, C. I.: Artificial intelligence (AI) for the early detection of breast cancer: A scoping review to assess AI’s potential in breast screening practice. Expert Review of Medical Devices, vol. 16, no. 5, pp. 351–362 (2019) doi: 10.1080/17434440.2019.1610387
21. Amisha, Malik, P., Pathania, M., Kumar-Rathaur, V.: Overview of artificial intelligence in medicine. Journal of Family Medicine and Primary Care, vol. 8, no. 7, pp. 2328 (2019) doi: 10.4103/jfmpc.jfmpc_440_19
22. Balkenende, L., Teuwen, J., Mann, R. M.: Application of deep learning in breast cancer imaging. Seminars in Nuclear Medicine, vol. 52, no. 5, pp. 584–596 (2022) doi: 10.1053/j.semnuclmed.2022.02.003
23. Vállez, N., Bueno, G., Déniz, O., Dorado, J., Seoane, J. A., Pazos, A., Pastor, C.: Breast density classification to reduce false positives in cade systems. Computer Methods and Programs in Biomedicine, vol. 113, no. 2, pp. 569–584 (2014) doi: 10.1016/j.cmpb.2013.10.004
24. Sechopoulos, I., Teuwen, J., Mann, R.: Artificial intelligence for breast cancer detection in mammography and digital breast tomosynthesis: State of the art. Seminars in Cancer Biology, vol. 72, pp. 214–225 (2021) doi: 10.1016/j.semcancer.2020.06.002
25. Melekoodappattu, J. G., Dhas, A. S., Kandathil, B. K., Adarsh, K. S.: Breast cancer detection in mammogram: Combining modified CNN and texture feature based approach. Journal of Ambient Intelligence and Humanized Computing, vol. 14, no. 9, pp. 11397–11406 (2022) doi: 10.1007/s12652-022-03713-3
26. Wang, Y., Yang, F., Zhang, J., Wang, H., Yue, X., Liu, S.: Application of artificial intelligence based on deep learning in breast cancer screening and imaging diagnosis. Neural Computing and Applications, vol. 33, no. 15, pp. 9637–9647 (2021) doi: 10.1007/s00521-021-05728-x

27. Ortiz-Rodriguez, J. M., Guerrero-Mendez, C., Martinez-Blanco, M. R., Castro-Tapia, S., Moreno-Lucio, M., Jaramillo-Martinez, R., Solis-Sanchez, L. O., Martínez-Fierro, M. L., Garza-Veloz, I., Moreira-Galvan, J. C., Barrios-Garcia, J. A.: Breast cancer detection by means of artificial neural networks. Advanced Applications for Artificial Neural Networks (2017) doi: 10.5772/intechopen.71256
28. Salama, W. M., Aly, M. H.: Deep learning in mammography images segmentation and classification: Automated CNN approach. Alexandria Engineering Journal, vol. 60, no. 5, pp. 4701–4709 (2021) doi: 10.1016/j.aej.2021.03.048
29. Muduli, D., Dash, R., Majhi, B.: Automated diagnosis of breast cancer using multi-modal datasets: A deep convolution neural network-based approach. Biomedical Signal Processing and Control, vol. 71, pp. 102825 (2022) doi: 10.1016/j.bspc.2021.102825
30. Agarwal, R., Diaz, O., Lladó, X., Yap, M. H., Martí, R.: Automatic mass detection in mammograms using deep convolutional neural networks. Journal of Medical Imaging, vol. 6, no. 03, pp. 1 (2019) doi: 10.1117/1.jmi.6.3.031409
31. Muralidhar, G. S., Haygood, T. M., Stephens, T. W., Whitman, G. J., Bovik, A. C., Markey, M. K.: Article commentary: Computer-aided detection of breast cancer — have all bases been covered? Breast Cancer: Basic and Clinical Research, vol. 2, pp. BCBCR.S785 (2008) doi: 10.4137/bcbcr.s785
32. Mert, A., Kılıç, N., Bilgili, E., Akan, A.: Breast cancer detection with reduced feature set. Computational and Mathematical Methods in Medicine, vol. 2015, pp. 1–11 (2015) doi: 10.1155/2015/265138
33. Warren-Burhenne, L. J., Wood, S. A., D'Orsi, C. J., Feig, S. A., Kopans, D. B., O'Shaughnessy, K. F., Sickles, E. A., Tabar, L., Vyborny, C. J., Castellino, R. A.: Potential contribution of computer-aided detection to the sensitivity of screening mammography. Radiology, vol. 215, no. 2, pp. 554–562 (2000) doi: 10.1148/radiology.215.2.r00ma15554
34. Touami, R., Benamrane, N.: Microcalcification detection in mammograms using particle swarm optimization and probabilistic neural network. Computación y Sistemas, vol. 25, no. 2 (2021) doi: 10.13053/cys-25-2-3429
35. Taylor, P.: Computer aided detection. Breast Cancer Research, vol. 2, no. S2 (2000) doi: 10.1186/bcr227
36. Khoo, L. A. L., Taylor, P., Given-Wilson, R. M.: Computer-aided detection in the United Kingdom national breast screening programme: prospective study. Radiology, vol. 237, no. 2, pp. 444–449 (2005) doi: 10.1148/radiol.2372041362
37. Aslam, M. A., Aslam, Cui, D.: Breast cancer classification using deep convolutional neural network. Journal of Physics: Conference Series, vol. 1584, no. 1, pp. 012005 (2020) doi: 10.1088/1742-6596/1584/1/012005
38. Al-antari, M. A., Al-masni, M. A., Choi, M. T., Han, S. M., Kim, T. S.: A fully integrated computer-aided diagnosis system for digital x-ray mammograms via deep learning detection, segmentation, and classification. International Journal of Medical Informatics, vol. 117, pp. 44–54 (2018) doi: 10.1016/j.ijmedinf.2018.06.003
39. Ketkar, N., Moolayil, J.: Convolutional neural networks. Deep Learning with Python: Learn Best Practices of Deep Learning Models with PyTorch, pp. 197–242 (2021) doi: 10.1007/978-1-4842-5364-9_6
40. Teoh, T. T., Rong, Z.: Convolutional neural networks. Artificial Intelligence with Python, pp. 261–275 (2022) doi: 10.1007/978-981-16-8615-3_16
41. Manaswi, N. K.: Convolutional neural networks. Deep Learning with Applications Using Python: Chatbots and Face, Object, and Speech Recognition with TensorFlow and Keras, pp. 91–96 (2018) doi: 10.1007/978-1-4842-3516-4_6
42. Skansi, S.: Convolutional neural networks. Introduction to Deep Learning: from logical calculus to artificial intelligence, pp. 121–133 (2018) doi: 10.1007/978-3-319-73004-2_6
43. Prakash-Kolla, B., Kanagachidambaresan, G. R.: Programming with TensorFlow: Solution for Edge Computing Applications pp. 45–51 (2021) doi: 10.1007/978-3-030-57077-4

44. Heath, M., Bowyer, K., Kopans, D., Kegelmeyer, P., Moore, R., Chang, K., Munishkumaran, S.: Current status of the digital database for screening mammography. Digital Mammography: Nijmegen, pp. 457–460 (1998) doi: 10.1007/978-94-011-5318-8_75
45. Aibar, L., Santalla, A., Criado, M. L., González–Pérez, I., Calderón, M., Gallo, J., Parra, J. F.: Clasificación radiológica y manejo de las lesiones mamarias. Clínica e Investigación en Ginecología y Obstetricia, vol. 38, no. 4, pp. 141–149 (2011) doi: 10.1016/j.gine.2010.10.016
46. Ozturk, T., Talo, M., Yildirim, E. A., Baloglu, U. B., Yildirim, O., Acharya, U. R.: Automated detection of COVID-19 cases using deep neural networks with X-ray images. Computers in Biology and Medicine, vol. 121, pp. 103792 (2020) doi: 10.1016/j.compbiomed.2020.103792
47. Redmon, J., Farhadi, A.: YOLO9000: Better, faster, stronger. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7263–7271 (2017)
48. Al-masni, M. A., Al-antari, M. A., Park, J., Gi, G., Kim, T., Rivera, P., Valarezo, E., Choi, M., Han, S., Kim, T.: Simultaneous detection and classification of breast masses in digital mammograms via a deep learning yolo-based cad system. Computer Methods and Programs in Biomedicine, vol. 157, pp. 85–94 (2018) doi: 10.1016/j.cmpb.2018.01.017
49. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016)
50. 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
51. Terven, J., Córdova-Esparza, D., Romero-González, J.: A comprehensive review of yolo architectures in computer vision: From YOLOv1 to YOLOv8 and YOLO-NAS. Machine Learning and Knowledge Extraction, vol. 5, no. 4, pp. 1680–1716 (2023) doi: 10.3390/make5040083
52. Li, B., Shi, Y., Qi, Z., Chen, Z.: A survey on semantic segmentation. In: IEEE International Conference on Data Mining Workshops (2018) doi: 10.1109/icdmw.2018.00176

ISSN 1870-4069

Automatic System for the Interpretation of the Vickers Hardness Test Using Artificial Vision

Ricardo Labrada-Lara, Moisés Márquez-Olivera,
Viridiana Hernández-Herrera, David Jaramillo, Christian García

Instituto Politécnico Nacional,
CIITEC,
Mexico

jlabradal1300@alumno.ipn.mx,
{mvmarquez, mxvhernándezhe}@ipn.mx

Abstract. Hardness is defined as the opposition that materials present to being penetrated or permanently deformed by another harder body. The Vickers type hardness test, also known as microhardness test, is a test used to determine the hardness of the surface of a material at a microscopic level. Due to the size of the imprint left by the indenter, parallax errors are usually common among operators. This, added to the vision fatigue caused by long testing sessions, is problematic when obtaining the values necessary to calculate the hardness of the material. With the implementation of intelligence and artificial vision, an algorithm can be developed that is capable of detecting and measuring the imprint left on the material at the time of carrying out the test and performing the calculation corresponding to the hardness of the material, in addition, to identify between a correct fingerprint and an incorrect one. For this reason, this research project proposes the development of an automatic system that allows the interpretation of the Vickers hardness test using artificial vision through the implementation of a convolutional neural network (CNN) trained with a database. which will be created from positive and negative images obtained from Vickers tests, resulting in a total of 3000 images using a 70/20/10 hold out validation method, having 1470 for the class called “correcta” and 1530 for the “incorrecta” class, resulting in an assertiveness index of 98 %.

Keywords: Indentation, Vickers hardness test, Computer-aided detection, Deep Learning, YOLO v8.

1 Introduction

All materials have characteristics in common, one of them is hardness, which is defined as the opposition that materials present to being penetrated or deformed by another harder body [1, 2]. This is why hardness is an important parameter in many industrial applications. To know this, artificial various methods were developed which were called hardness tests [3, 4]. In the area of materials, hardness testing is a tool used to ensure that welding, heat treatment and manufacturing methods have not altered the original material or in other cases serves as a quick method to determine that forming

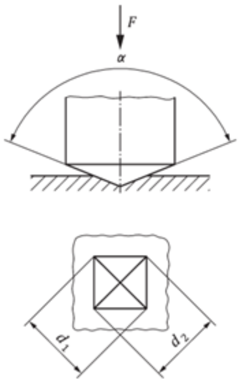


Fig. 1. Indentation left in a Vickers type hardness test [11].

and manufacturing techniques have not made the material too soft or hard, in addition to that, for some metals and polymers, there are empirical correlations between hardness and the resistance (or modulus) of the material [5, 6].

Some of the tests responsible for measuring hardness are the Brinell, Vickers and Rockwell tests, the latter is characterized by knowing the hardness based on the depth of the indentation while the Brinell and Vickers scales, also known as micro hardness tests, are based on measuring the length of the indentations [7, 8].

Particularly, for this research project we will focus on the Vickers type test. Initially, this method was carried out with the help of micro durometers which made an indentation (imprint) with a diamond-tipped indenter, in the shape of a straight pyramid with a square base and with a specific angle between opposite faces at the vertex (generally 136 °), said indenter exerted a force on the surface of a test piece followed by measuring the length of the diagonals of the indentation left on the surface after removing the applied force or applied load, as shown in Fig. 1 [7, 9,10].

It is worth mentioning that both the execution of the indentation and the measurement of the diagonals are carried out by an operator who, based on his perception, opinion and experience, determines the distance of the latter [12]. These parameters, both to generate a hardness test and to extract data, are determined by the ISO 6507-1 standard, which represents the Vickers hardness with the following equation:

$$HVN = \frac{2F\sin(\frac{\alpha}{2})}{g_n d^2}, \tag{1}$$

where: α is equal to the mean angle between the opposite faces at the vertex of the pyramidal indenter (generally it is 136°), F represents the load applied to the indenter, g_n is the value of gravity 9.8 m/s^2 and d^2 is the value obtained from the diagonals. When substituting the known values, the following equation is obtained:

$$HVN33 = \frac{F * 0.1891.}{d^2} \tag{2}$$

The next step was to make micro durometers that will manually help the operator measure the diagonals using knobs to position a mesh over the indentation so that the same durometer can later apply the formula with the values obtained and return the value of the hardness of the material, it is worth mentioning that these micro durometers are calibrated under the ISO 6507-2 standard [13, 14]. This evolution was carried out with the intention of eliminating human error when measuring the diagonals, however, this posed another problem which is that it continues to depend on an operator to perform the calibration of the mesh in charge of measuring the diagonals [15 - 17].

1.1 Artificial Intelligence (AI)

The field of artificial intelligence is in constant growth and according to the World Economic Forum, the Future of Jobs Report 2023 concludes that almost a quarter of all jobs (23%) will change in the next five years, being artificial intelligence the seventh technological skill on the rise during this period [18, 19]. Almost 75% of companies surveyed are expected to adopt artificial intelligence, second only to the field of robotics (humanoid and industrial robots).

This makes it almost mandatory for students to acquire the ability to analyze and interpret various problems/situations, as well as the ability to understand and work with emerging technologies, regardless of the field they choose, they should try to develop these generalist competencies to be prepared for a future that is changing by leaps and bounds [20]. This is why the most current micro-hardness testers make use of specialized software to obtain an image of the indentation and calculate the hardness automatically.

However, the acquisition of this type of equipment, as well as its calibration, can be generate considerable cost. This type of software can be replicated using artificial intelligence, which is defined as all those algorithms that seek to make a program intelligent or rational, trying to emulate human intelligence and natural language, by subjecting the algorithms to learning phases with methods. varied (Machine Learning or Deep Learning), but all with the same purpose, which is to collect the greatest amount of information and make it functional [21 - 23].

This is achieved with the help of a database so that said data can be converted into knowledge, the greater the amount of information presented, the greater the speed with which it will identify the desired parameters and/or patterns to combine them with the requirements provided by the developer.

1.1.1Introduction Background and Scope of this Study

As mentioned above, the processes of identification and/or extraction of information can be given by various methods, whether Machine or Deep Learning, by analyzing the characteristics of the object of interest it is not possible to obtain good results through data analysis structured since the latter is presented in the form of an image, which is why the model that best fits this situation is a Convolutional Neural Network (CNN), this type of networks are characterized by having the ability to process and extract information from images [24, 25], these being the input values, as shown in Fig. 2.

This is achieved by “decomposing” the image so that the network is able to identify the most characteristic features of the object of interest and allows it to classify them into previously established categories [26-28], the result of this learning process calculates its degree of accuracy as soon as an image is presented to him and he must

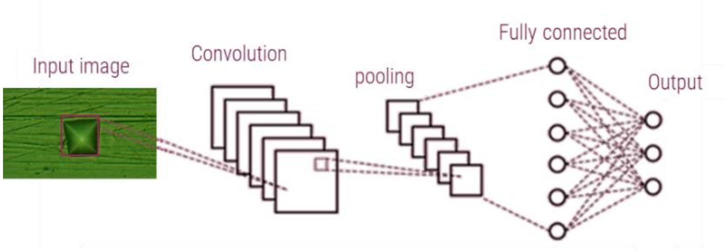


Fig. 2. Scheme of a convolutional neural network (CNN).

look for the characteristics that were taught to him in order to classify it into one of the categories he knows.

As an example of projects where not only this type of neural networks but other artificial intelligence models have been applied to solve problems similar to the one proposed or even that can create a union between different areas, such as the mechanical area and artificial intelligence, as one of these projects is the research of Zexian Li [26].

This article presents a proposal in which artificial intelligence is used to provide a solution for obtaining the values necessary to know the hardness of a material from the image left when performing a hardness test using a fully convolutional neural network encoder-decoder (FCN-ED). As shown in the materials and methods section, Zexian Li generated a database with images of indentations made on his own.

Once the database was compiled, it was subjected to digital image processing to eliminate minor elements. interest to subsequently be entered into the proposed algorithm (FCN-ED) and initiate a training phase and then underwent a validation phase to corroborate the percentage of learning during the previous phase. Once the necessary adjustments have been made, we proceed to the testing phase where the algorithm had to calculate the hardness from new images, giving satisfactory results when comparing them with the results obtained by traditional methods.

Another article with a similar process is that of Denis Privezentse [29] who presents it as a project focused on the premise that the most current equipment for carrying out hardness tests is very expensive, so the aim is to automate said process through artificial vision.

Even though this was the real intention of the research, it was not possible to complete this objective by passing a different premise, now being an image process where 6 different types of filters are applied (starting with a smoothing filter, a simple filter, a binomial filter, a Gaussian filter and at the end a simple filter) to achieve better quality in the image obtained by the analog microscope to which a digital camera was adapted instead of the lens, once the image passes through the different filters the result is given to an expert to calculate the hardness manually, resulting in a hardness that is not very far from that obtained in a traditional way, that is, without the use of any external equipment apart from the microhardness tester.

On the other hand A. P. Fedotkin [30] presents a different solution to CNNs for extracting mechanical data obtained from an image of a hardness test, although there are different methods, and therefore different algorithms, capable of determining the area of the indenter which is necessary to know the hardness of a material in hardness tests, Fedotkin proposed an algorithm based on the comparison of two-dimensional

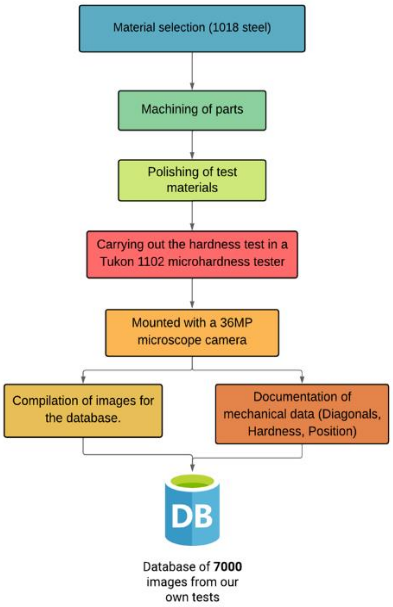

Material selection (1018 steel)
Machining of parts
Polishing of test materials
Carrying out the hardness test in a Tukon 1102 microhardness tester
Mounted with a 36MP microscope camera
Compilation of images for the database.
Documentation of mechanical data (Diagonals, Hardness, Position)
DB
Database of **7000** images from our own tests

Fig. 3. Process of creating test material/database.

diagrams of the optical properties of the exterior and interior parts of a fingerprint on the image obtained. For example, in the first case, several circles with radii in the range of 5 µm to 90% of the frame height are constructed from the center of the frame. Subsequently, the radius was increased while continuing to compare the properties inside and outside the area of interest until a clear difference was found between both properties (region within the indentation and the rest of the surface).

2 Materials and Methods

In order to train any artificial intelligence algorithm, a database of the subject to be analyzed is needed, which is why the first step of the methodology is based on the collection of images of indentations from Vickers hardness tests, Not finding a sufficiently robust database, we opted to carry out our own database based on Vickers hardness tests in a Tukon™ 1102 micro durometer with specimens made of 1018 steel also made in our own way, and to capture The images were supported by a 36 MP microscope camera that was mounted on the micro durometer while a x50 objective was selected, this entire process can be better observed in Fig. 3.

2.1 Database

As the tests were carried out, the conditions and parameters under which the tests were carried out were documented, as well as the images of the latter and their respective results were captured, all in order to have a reference regarding the results presented by the test algorithm in its final phase.

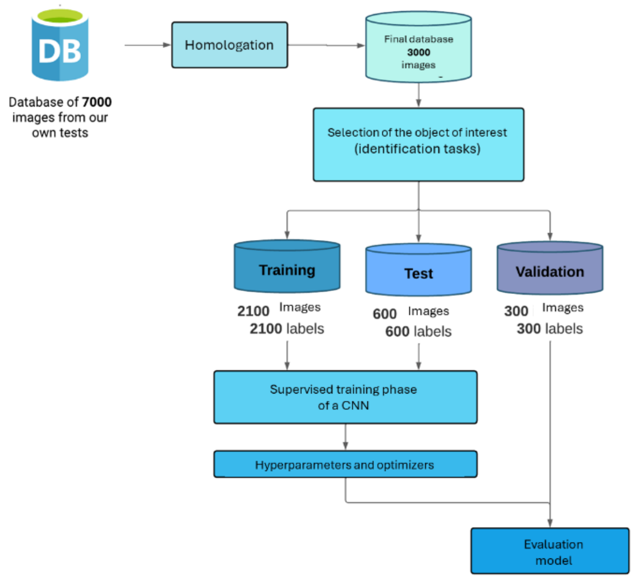

DB
Database of 7000 images from our own tests
Homologation
Final database 3000 images
Selection of the object of interest (identification tasks)
Training
2100 Images
2100 labels
Test
600 Images
600 labels
Validation
300 Images
300 labels
Supervised training phase of a CNN
Hyperparameters and optimizers
Evaluation model

Fig. 1. Database division.

Table 1. Sets that make up the Database divided by classes.

Train		**Test**		**Val**	
correct	incorrect	correct	incorrect	correct	incorrect
1,041	1,059	287	313	142	158

The result of the above concluded with a data bank of 7,000 images, as seen in Fig. 4, which make up the data bank, but of course these images are raw, the next step being the approval of the bank to preserve only those images that have the object of interest and that highlight in them the most significant characteristics for the next steps, the result of this point was the formation of a database of 3,000 images that range from indentations to different sizes, with indentation blurred images, correct and incorrect indentations, images with filters, indentations seen from different angles, all with the intention of improving training results.

Having the approved database, the region of interest is determined by means of labels that contain the coordinates of the indentation in the image. This is achieved by exporting the latter in text format to be used in the training of YOLO neural networks. Subsequently, the database must be divided into three sets: a training set (Train), a validation set (Val) and a test set (Test), dividing the images, occupying 70% of the data for the set of training, 20% for validation set and 10% for test data, applying this approach results in 2,100 images for Train, 600 images for Val and 300 images for Test, this selection was achieved manually by evaluating the total of images by class and performing the aforementioned division, as seen in Table 1 showing the number of images by sets and class.

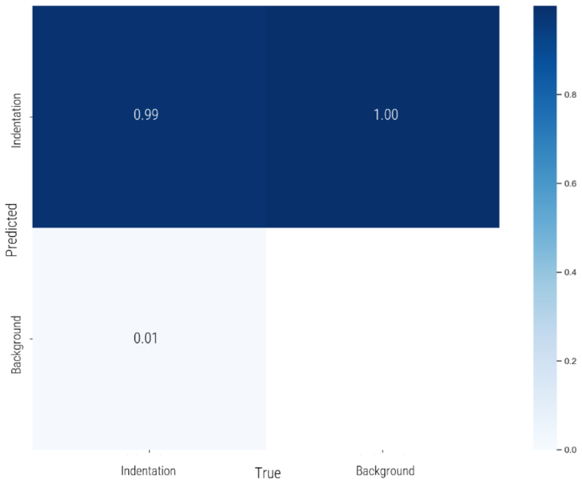


Fig. 5. YOLOV8-D4 confusion matrix (YOLOv8 with 100 epochs and one class).

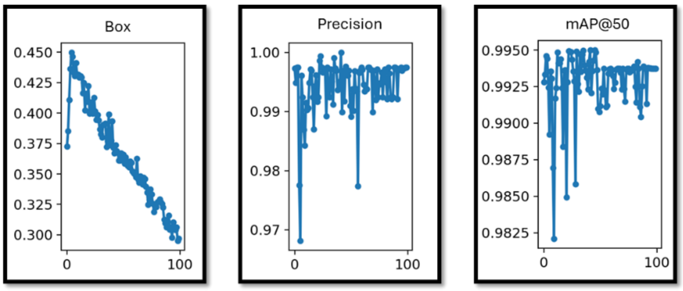


Fig. 6. YOLO8-D4 results graphs.

2.2 Image Preprocessing

As the next step of the methodology, there is the preprocessing and segmentation of images, the objective of this stage is the improvement of the final images in the database if required, this in order to enhance the essential characteristics of the object of interest that the algorithm must learn, for these different types of filters are applied that allow modifying characteristics of an image such as sharpness, greater contrast, among many other characteristics.

Once the images have been improved, the search begins for the different characteristics contained in each image that belongs in this case to the indentations. For this purpose, different identification methods are used which highlight the coordinates where the object of interest is located using detection at an early stage of the project and later moving on to using segmentation and declaring the corresponding labels in text format: “indentacion”, “correcta” and “incorrecta”.

Having the elements classified by tags and separated into their respective sets, the next step is to make use of YOLOv8 obtained directly from the authors repository [31], By using the tools provided in it, you can extract the features previously highlighted by

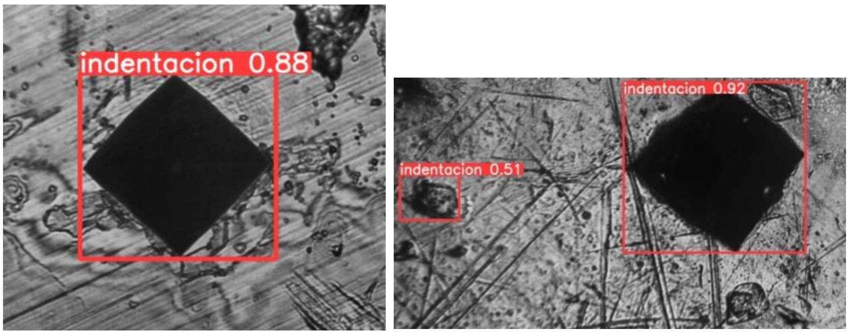


Fig. 7. Detected indentations result of the YOLO8-D4 model with 100 epochs and one class.

the segmentation task. This is possible since YOLOv8 works with Convolutional Neural Networks that allow you to view the image as three-by-three matrices (height, length and depth). The parameters established for the training that achieved the best results, with respect to the characteristics of both the database and the equipment used, were 100 epochs with a batch of 4 applied to the previously explained base.

3 Results

As part of the results, the graphs obtained from Yolov8 will be shown under the previously set parameters, as well as the images of the validation set where it is confirmed whether it really learned in the expected way. Initially, training was proposed with the database applying a detection type identification method, but maintaining the epoch parameters at 100 and a batch of 4, of course this was carried out in an early process of the project so the database did not It had the same number of images of the current database, in addition to the fact that at that time only a single class had been proposed and limiting the identification process only to detecting indentation as “indentacion”, shown in the following Fig. 5.

As can be seen in Fig. 5, a null error is shown when recognizing an indentation, which may mean that the network was overtrained, in addition to the fact that the background is practically zero when it should coincide with itself or at least the latter should be confused with what is proposed as an indentation. In addition to the confusion matrix, it is also possible to obtain metrics that show the learning process of the algorithm in this case, the graphs in Fig. 6.

In this image you can see how in the "box" graph, at first the forgetting factor began at a relatively low point, but in the following epochs it increased until as the epochs continued to increase it decreased again, doing so in this occasion constantly until the end of the 100 epochs, something similar happens in the precision and mAP@50 graphs, in both the graph starts with a high assertiveness index, being at 99%; however as the epochs pass the algorithm begins to present a drop in learning, varying greatly between epochs despite having a percentage of between 99% and only in very specific cases reaching 100%.

Thanks to these graphs, it can be verified that the hypothesis of over-training by the algorithm is erroneous, since if this had been the case, in the mAP@50 graph a value

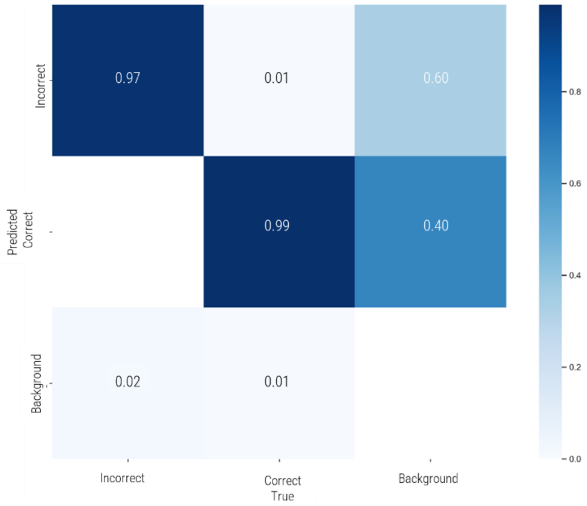


Fig. 2. YOLOV8-S4 confusion matrix (YOLOv8 with 100 epochs and two classes).

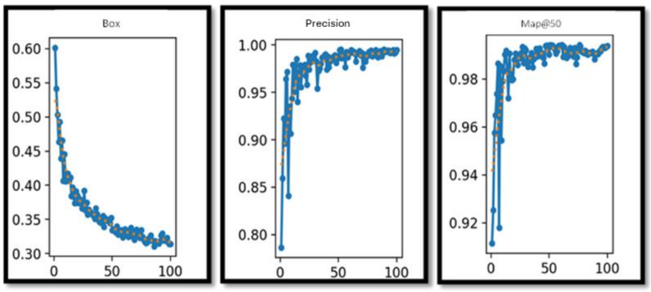


Fig. 3. YOLO8-S4 results graphs.

equal to 100% should have first been reached, and from that point on, the following data should have been consistent, reflecting that the algorithm stopped learning to simply memorize the information taught.

However, not only did it not reach that value, but as the times passed, the learning capacity it presented in Initially, it began to stabilize at a relatively lower value than what it initially reached.

Added to the above, when performing tests on the validation set, although it was capable of detecting an indentation with a relatively accurate index as shown in Fig. 7, it often confused imperfections in the image, passing them off as a possible indentation, although with a lower assertiveness index.

The way to improve the accuracy was by strengthening the database up to the value shown in the methodology, in addition to changing the identification task from detection to segmentation, this in order to better delimit the indentation causing the algorithm to be capable of extracting more significant elements and also increasing the number of classes to two, no longer just identifying What is an indentation? If you do not differentiate between an incorrect indentation and a correct one, the training

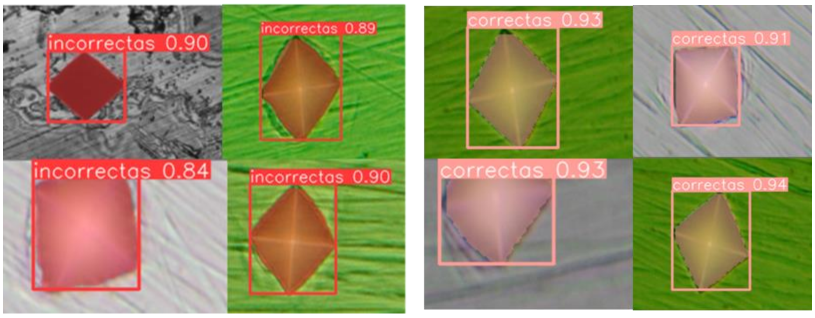


Fig. 4. Detected indentations resulting from a YOLO8-S4 model with 100 epochs and two classes.

parameters used were the same as in the step case, obtaining on this occasion the following confusion matrix.

In comparison to Fig. 5 with Fig. 8, it can be noted that by increasing the database and increasing the number of classes, the network does not seem to have been overtrained since none of its values are completely 1.0. and the percentage of error presented when confusing both classes is very low, in addition in this training the background already has some value, it is not the most promising thing to be confused with the classes, but compared to the training presented at the beginning it is an advance, In addition to the fact that the images in the validation set also increased, there is more variety of images when checking the results.

As part of the metrics obtained from the training shown in Fig. 9, it can be seen in the Map@50 graph how in the first epochs the algorithm does not present constant learning, this may be because the images selected for those epochs presented difficulty in determining the object of interest, however, before reaching epoch #50 there is constant learning that remains above 98%, this is reinforced with the precision graph where it can be observed that the oscillations of learning are reduced as the periods increase until they reach the point where their pressure does not vary greatly and without completely reaching 100%.

As a result of the above, when testing, the algorithm is able to satisfactorily identify between correct and incorrect indentations taken from the validation set, this can be seen in the following Fig. 10.

4 Discussion

This article aims to detect between an indentation carried out correctly and an incorrect one through an image of it, this by making use of convolutional neural networks, obtaining as a result an algorithm capable of interpreting said images and selecting which one of these two classes belong. To reach this result and as shown in the results part, when training with a reduced database, compared to the size of the most recent version, and a single “indentation” class, an assertiveness index of 99% was obtained.

However, this fact does not necessarily have to be good either since such a high assertiveness index, which almost reaches 100% or, failing that, reaches 100%, can reflect that the algorithm experienced overtraining, which indicates that it stops learning and is only remembering based on the indentation set. This has the effect that if it is presented with images outside of said set, perhaps the algorithm will not be able to correctly identify the indentation.

The way to corroborate this hypothesis is to review the metrics obtained from the training, in which it is observed how the algorithm presents a dispersion in its data, this means that the way in which it is learning is not constant despite obtaining a high index per which the algorithm was not considered overtrained, however, the dispersion in the data can be caused by two reasons, the first is that the images used in the training set present irregularities that make data extraction more difficult, most significant of the object of interest, linked to the above is the second reason, which is speculated to be the identification method is not correctly extracting the most significant data.

With this first, it was proposed to strengthen the database in order to expand the number of images, that is, more indentations had to be made and each of their data had to be documented (length of both diagonals and hardness obtained). This was in order for the algorithm to have a greater variety of data from which to extract significant characteristics and for this last process the identification method was changed from detection to segmentation, this with the aim of delimiting as much as possible the region where the object of interest is located.

At the same time that the database was strengthened, the classes were increased, going from “indentation” to “correct” and “incorrect”, considering that although the algorithm presented problems with learning, it still had an assertiveness index of 99 %, as the training had been planned under the new parameters (database and identification task), they initially generated a confusion matrix where the classes, although they already showed an assertiveness index of 99%, showed equally good values (97% for the incorrect class and 98% for the correct class) in addition to reviewing and comparing the training graphs, a more constant learning was shown compared to the previous training, corroborating that increasing the number of images and changing the identification task was positive since the dispersion in the data only occurs in the first epochs, but before reaching epoch 50 it not only stops showing drops in learning, but also that the latter also becomes more constant with the passing of the epochs until reaching a point where the learning curve “stabilizes”, added to the above, the training characteristics were varied to verify if by increasing the number of epochs, reducing or increase the batch, considerable differences were found or that called into question the experimentation process as it was being carried out.

However, the variation in the aforementioned parameters did not show a notable improvement in either the assertiveness index or the learning behavior, so it was decided that the parameters established in the materials and methods section (100 training epochs, batch of 4 and segmentation task) would be the parameters that showed better performance when identifying correct and incorrect indentations.

5 Conclusion

This article presents the evaluation of the comparison between two YOLOV8 models using different validation methods to determine which one gives better results when identifying not only an indentation, but also the latter can be classified between a test performed positively and one negative with an assertiveness index of 98% under the established hyperparameters, however, it is necessary to work on how to extract the area of interest, that is, the indentation, thus delimiting the object of study, applying a detector of vertices and be able to measure the diagonals in the indentation, in this way having the necessary values to apply the formula and obtain the hardness of the material from an image of the indentation, in this way, the aim is to reduce the workload of those in charge of carrying out the Vickers hardness tests as well as the difference in the measurements of the diagonals resulting from a parallax error added to the visual fatigue resulting from these tests.

6 Future work

As part of the future implementations, it is considered to finalize the part corresponding to the measurement of the diagonals in the image of an indentation and consequently the extraction of the mechanical value of the hardness, as well as the development of a graphic interface that allows the user to interact with the algorithm, view the result of the indentation classification, perform the hardness measurement and document the results.

7 Conflict of Interests

The authors declare that there is no conflict of interest.

Acknowledgment. We especially thank the technological research and innovation center (CIITEC) and its teachers for the support and trust in my work as well as the time given to the project.

References

1. Zhang, C., Li, F., Wang, B.: Estimation of the elasto-plastic properties of metallic materials from micro-hardness measurements. Journal of Materials Science, vol. 48, no. 12, pp. 4446–4451 (2013) doi: 10.1007/s10853-013-7263-3
2. Wang, Z., Sha, A.: Micro hardness of interface between cement asphalt emulsion mastics and aggregates. Materials and Structures, vol. 43, no. 4, pp. 453–461 (2009) doi: 10.1617/s11527-009-9502-2
3. Shinohara, K.: Relationship between work-hardening exponent and load dependence of vickers hardness in copper. Journal of Materials Science, vol. 28, no. 19, pp. 5325–5329 (1993) doi: 10.1007/bf00570084
4. Moss, D. R., Basic, M.: Pressure vessel design manual. Fourth Edition, Oxford, Butterworth-Heinemann, pp. 719–742 (2013) doi: 10.1016/C2010-0-67103-3
5. Wang, M., Wang, C.: Bulk properties of biomaterials and testing techniques. Reference Module in Biomedical Sciences: Encyclopedia of Biomedical Engineering, pp. 53–64 (2019) doi: 10.1016/b978-0-12-801238-3.99861-1
6. Chen, H., Fu, Z., Chen, D., Peng, H., Li, W., Meng, Z., Fan, Z.: A unified sharp indentation method for obtaining stress-strain relations, strength and vickers hardness of ductile metallic materials. Materials Today Communications, vol. 33, pp. 104652 (2022) doi: 10.1016/j.mtcomm.2022.104652
7. ISO.: Metallic materials, In: Vickers Hardness Test. Part 1: Test Method. pp. 1-11 (2018)

8. Ma, D. J., Wang, J. L., Sun, L., Huang, Y.: Method for identifying vickers hardness by instrumented indentation curves with Berkovich/Vickers indenter. Experimental Mechanics, vol. 56, no. 5, pp. 891–901 (2016) doi: 10.1007/s11340-016-0136-3
9. Lai, M. O., Lim, K. B.: On the prediction of tensile properties from hardness tests. Journal of Materials Science, vol. 26, no. 8, pp. 2031–2036 (1991) doi: 10.1007/ bf00549163
10. Fabijanić, T. A., Franz, M., Alar, Ž.: Influential factors on hardness uniformity of vickers hardness blocks for high hardness range. Measurement, vol. 78, pp. 358–365 (2016) doi: 10.1016/j.measurement.2015.07.030
11. Elssner, G. H., Hoven, H., Kiessler, G., Wellner, P.: Ceramics and ceramic composites: materialographic preparation. Elsevier Science, pp. 144–158 (1999)
12. Broitman, E.: Indentation hardness measurements at macro-, micro-, and nanoscale: A critical overview. Tribology Letters, vol. 65, no. 1, pp. 23 (2016) doi: 10.1007/s11249-016-0805-5
13. Kadiyan, S., Dehiya, B. S., Garg, R. K., Kamiya, P., Saini, M.: A statistical method to predict the hardness and grain size after equal channel angular pressing of AA-6063 with intermediate annealing. Arabian Journal for Science and Engineering, vol. 46, no. 3, pp. 2055–2070 (2020) doi: 10.1007/s13369-020-04999-1
14. Fernandes, T. E., Ferreira, M. A., de-Miranda, G. P. C., Dutra, A. F., Antunes, M. P., da-Silva, M. V., de-Aguiar, E. P.: Classification of lathe's cutting tool wear based on an autonomous machine learning model. Journal of Control, Automation and Electrical Systems, vol. 33, no. 1, pp. 167–182 (2021) doi: 10.1007/s40313-021-00819-5
15. Hu, X., Li, J., Wang, Z., Wang, J.: A microstructure-informatic strategy for vickers hardness forecast of austenitic steels from experimental data. Materials & Design, vol. 201, pp. 109497 (2021) doi: 10.1016/j.matdes.2021.109497
16. Gao, Y. X., Fan, H.: A micro-mechanism based analysis for size-dependent indentation hardness. Journal of Materials Science, vol. 37, pp. 4493–4498 (2002) doi: 10.1023/A:1020662215932
17. Jain, A., Razdan, A. K., Kotru, P. N., Wanklyn, B. M.: Load and directional effects on microhardness and estimation of toughness and brittleness for flux-grown LaBO3 crystals. Journal of Materials Science, vol. 29, no. 14, pp. 3847–3856 (1994) doi: 10.1007/bf00357358 S
18. Di-Battista, A., Grayling, S., Hasselaar, E., Leopold, T., Li, R., Rayner, M., Zahidi, S.: Future of jobs report 2023. In: World Economic Forum, Geneva, weforum.org/reports/the-future-of-jobs-report-2023
19. Ciancarini, P., Succi, G.: In: Proceedings of 4th international conference in software engineering for defence applications. CRIS Current Research Information System, pp. 1–330 (2015) doi: 10.1007/978-3-319-27896-4
20. Pimenov, D. Y., Bustillo, A., Wojciechowski, S., Sharma, V. S., Gupta, M. K., Kuntoğlu, M.: Artificial intelligence systems for tool condition monitoring in machining: Analysis and critical review. Journal of Intelligent Manufacturing, vol. 34, no. 5, pp. 2079–2121 (2022) doi: 10.1007/s10845-022-01923-2
21. Shan, H., Jia, X., Yan, P., Li, Y., Paganetti, H., Wang, G.: Synergizing medical imaging and radiotherapy with deep learning. Machine Learning: Science and Technology, vol. 1, no. 2, pp. 021001 (2020) doi: 10.1088/2632-2153/ab869f
22. Soleymani, M., Khoshnevisan, M., Davoodi, B.: Prediction of micro-hardness in thread rolling of St37 by convolutional neural networks and transfer learning. The International Journal of Advanced Manufacturing Technology, vol. 123, no. 9-10, pp. 3261–3274 (2022) doi: 10.1007/s00170-022-10355-4
23. Lipiński, D., Ratajski, J.: Modeling of microhardness profile in nitriding processes using artificial neural network. In: Advanced Intelligent Computing Theories and Applications. With Aspects of Artificial Intelligence: Third International Conference on Intelligent Computing, pp. 245–252 (2007) doi: 10.1007/978-3-540-74205-0_27

24. Ji, Q.: Probabilistic graphical models for computer vision. Academic Press, pp. 191–297 (2020)
25. De-Oliveira-Baldner, F., Bastos-Costa, P., Santos-Gomes, J. F., Rodriguez-Leta, F.: A review on computer vision applied to mechanical tests in search for better accuracy. Advances in Visualization and Optimization Techniques for Multidisciplinary Research, pp. 265–281 (2019) doi: 10.1007/978-981-13-9806- 3_9
26. Li, Z., Yin, F.: Automated measurement of vickers hardness using image segmentation with neural networks. Measurement, vol. 186, pp. 110200 (2021) doi: 10.1016/j.measurement.2021.110200
27. Tanaka, Y., Seino, Y., Hattori, K.: Vickers hardness measurement by using convolutional neural network. Journal of Physics: Conference Series, vol. 1065, pp. 062001 (2018) doi: 10.1088/1742-6596/1065/6/062001
28. Tanaka, Y., Seino, Y., Hattori, K.: Automated Vickers hardness measurement using convolutional neural networks. The International Journal of Advanced Manufacturing Technology, vol. 109, no. 5-6, pp. 1345–1355 (2020) doi: 10.1007/s00170-020-05746-4
29. Privezentsev, D., Zhiznyakov, A., Kulkov, Y.: Analysis of the microhardness of metals using digital metallographic images. Materials Today: Proceedings, vol. 11, part 1, pp. 325–329 (2019) doi: 10.1016/j.matpr.2018.12.152
30. Fedotkin, A. P., Laktionov, I. V., Kravchuk, K. S., Maslenikov, I. I., Useinov, A. S.: Automatic processing of microhardness images using computer vision methods. Instruments and Experimental Techniques, vol. 64, no. 3, pp. 357–362 (2021) doi: 10.1134/s0020441221030180
31. Jocher, G., Chaurasia, A., Qiu, J.: YOLO by Ultralytics (2023)

ISSN 1870-4069

Parkinson's Detection Using Convolutional Neural Networks on Handwritten Wave Images

Carolina Rosas-Alatriste, Noé Oswaldo Rodríguez-Rodríguez,
Paola Itzel Delena-García, Antonio Alarcón-Paredes

Instituto Politécnico Nacional,
Centro de Investigación en Computación,
Mexico

{crosasa2023, nrodriguezr2023, pdelenag2023,
aalarcon}@cic.ipn.mx

Abstract. Parkinson's disease is a neurodegenerative condition for which the early detection is a very challenging activity for the medical community. Although traditional methods for Parkinson's disease diagnosis involve the use of EEG (electroencephalographic) activity, previous works have proposed to analyze sketches of guided spirals and waves drawn by a patient versus those drawn by healthy people. In this work, we made use of the same dataset, employing data augmentation techniques for enriching the diversity of the images. Besides, architectures such as ResNet50 and VGG19 demonstrated promising results using transfer learning. Results reported in this manuscript are comparable with those of the stateof-the-art, but also have the potential to improve the accuracy in the near future.

Keywords: Parkinson, CNN, deep learning, machine learning, classification, supervised learning.

1 Introduction

Parkinson's disease is a common neurological condition that can significantly disrupt a patient's ability to lead a normal life. It is a progressive neurodegenerative disorder that is often challenging to detect in its early stages. Traditional methods of diagnosing Parkinson's disease using EEG (electroencephalogram) data involve laborious and time-consuming manual feature extraction. To address this issue, in this article, we propose a diagnostic method that can be conducted in a medical office assisted by convolutional neural networks (CNN).

Taking into consideration the clinical presentation of the disease, as discussed in previous references, it is possible to diagnose Parkinson's disease by analyzing the dynamics of sketching guided spirals and waves drawn by a patient on a sheet of paper. We used the database published by Zham et al. (2017) [12], which is composed by a set of images of sketches labeled by healthcare professionals into categories of healthy patients and patients with Parkinson's disease.

Table 1. Summary table of the results found in the literature for the detection of Parkinson's disease. The table specifies the techniques and specific tasks reported in each case.

Reference	Tasks	Methodology	ACC
Vatsaraj and Nagare [11]	Spiral and wave analysis	CNN	96.7%
Zham et al. [12]	Guided spiral	Naïve Bayes with features in handwriting	93.3%
Gallicchio et al. [3]	Spirals and stability movement	Deep Echo State Networks	89.3%
Gil Martín et al. [4]	Spirals and stability movement	CNN	96.5%
Khatamino et al. [7]	Spirals and stability movement	CNN	72.5%
Chakraborty et al. [1]	Spirals and wave analysis	CNN	93.3%

1.1 Related Work

The prevailing method extensively employed by medical professionals in clinical settings for diagnosing Parkinson's Disease involves assessing and evaluating patients through a review of their medical history. This assessment often leads to the assignment of a rating scale based on the patient's performance. The predominant rating system in use to date is the Unified Parkinson Disease Rating Scale (UPDRS), as introduced by Goetz and Stebbins in 2004. [5]. Based on the work done by Goets and Stebbins, Sa, W. et al. in 2003 [9] argued that bradykinesia and other motor symptoms play a fundamental roll in opportune clinic Parkinson´s diagnosis.

Numerous studies have highlighted the direct connection between Parkinson's disease and symptoms related to motor function disorders, such as rigidity, tremors, and bradykinesia. Rigidity and bradykinesia are often evident in the early stages of the disease and affect a patient's ability to write and sketch. Research indicates that an individual's handwriting is influenced by factors like education, knowledge, and language proficiency (Zham et al., 2017) [12].

In contrast, the sketching of spiral and wave drawings serves as independent and noninvasive measures. Extracting features from handwritten sketches can be dynamic, facilitating real-time and dependable analysis. This approach also allows for the development of applications capable of extracting these features through online patient assessments.

Some related work that makes use of Machine Learning related techniques are referenced in Table 1 1. One of the most outstanding results found in the state-of-the-art is the research made by Gil Martín et al. (2019) [4]. This study contributes to this endeavor by examining the application of a convolutional neural network (CNN) for the detection of Parkinson's disease based on drawing movements. The CNN comprises two key components: feature extraction, which involves convolutional layers, and classification, implemented through fully connected layers.

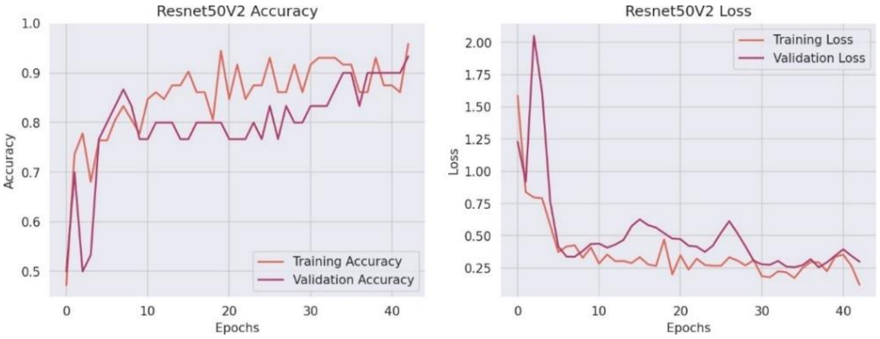


Fig. 1. Results for ResNet50.

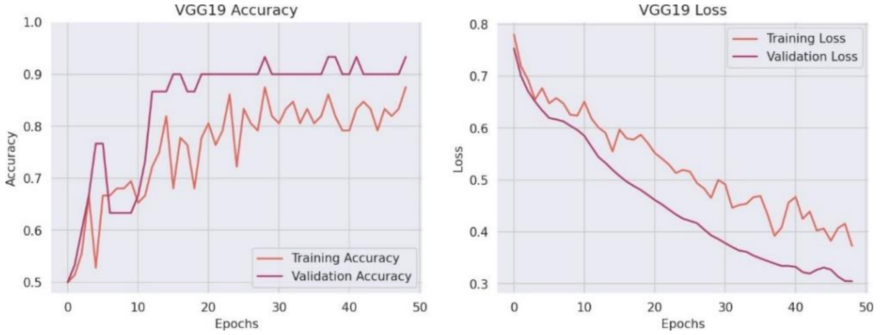


Fig. 2. Results for VGG19.

The inputs to the CNN consist of the Fast Fourier Transform module, focusing on frequencies within the 0 Hz to 25 Hz range.We assessed the discriminative capacity of various directions during drawing movements and found that the X and Y directions yielded the most favorable outcomes. This analysis was conducted using a publicly available dataset: the Parkinson Disease Spiral Drawings Using Digitized Graphics Tablet dataset. The most noteworthy results obtained in this study demonstrate an accuracy of 96.5%. In other hand, the top result found in literature is obtained by Vatsaraj and Nagare (2021) [11].

In this research, the team explore this biomarker by scrutinizing the sketching patterns evident in spiral and wave drawings produced by both healthy subjects and Parkinson's disease patients. Additionally, this study introduces optimizations to algorithms for feature extraction and classification. Notably, the proposed model exhibits an accuracy of 96.67%, alongside a precision of 93.33% and a recall of 100%. Additionally, Parkinson's disease (PD) is a neurodegenerative condition characterized by frequently changing motor symptoms.

The effective monitoring of these symptoms is crucial for tailoring treatment to individual patients. Donié et al. (2023) [9] mention that traditional time series classification (TSC) and deep learning methods have limitations when applied to PD symptom monitoring using data from wearable accelerometers. This is primarily due to the complexity of PD movement patterns and the limited size of datasets. In the state-of-the-art, we will find some achievements boarding various techniques. The

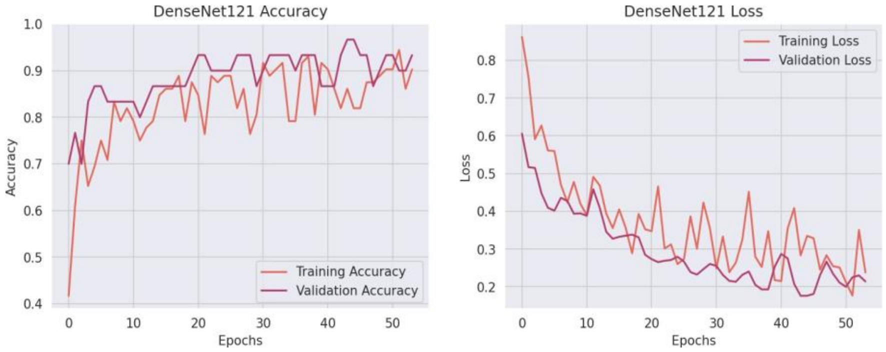


Fig. 3. Results for DenseNet121.

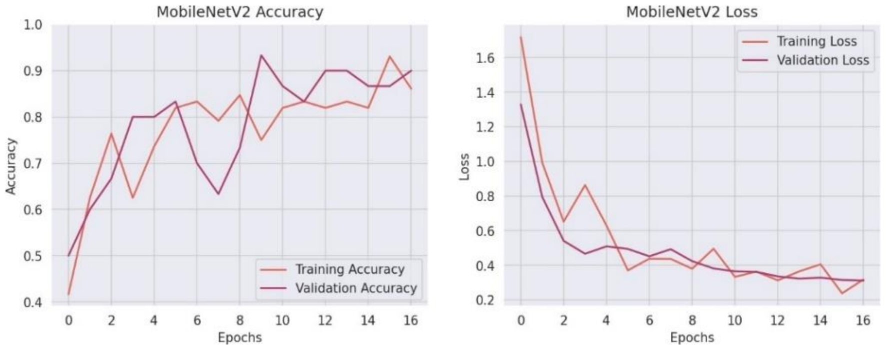


Fig. 4. Results for MovileNetV2.

mentioned techniques take advantage of specific aspects seen on the Parkinson’s clinical picture. For instance, Taylan et al. (2020) [10], mention that there has been a growing promise in the use of various statistical regression models for diagnosing neurodegenerative conditions such as Parkinson’s disease.

Nevertheless, when experimental data includes outlier observations that significantly deviate from the rest of the data points, traditional and widely recognized statistical regression models can yield inaccurate results for neurodegenerative disease diagnosis. In that sense, Dabbabi et al. (2023) [2], considered vocal cord disorders that are often considered a prominent contributor to Parkinson’s disease in many individuals, with speech impairments serving as one of the initial indicators of this condition.

In their study, they propose a model using VOT-MFCC as the primary feature and employs a Fully-Connected Deep Neural Network (FC-DNN) as the classifier; with encouraging results. Also, in the matter of Parkinson’s detection via Machine Learning, in the state-of-the-art we might find promising results. Kumar et al. (2023) [8] reported Deep Learning techniques proposed as a means to streamline the Parkinson’s detection process and enhance accuracy.

YAMNet, a computationally efficient deep-learning model designed for audio categorization, was employed to extract features from a speech signals dataset related to Parkinson’s disease. The study assessed the effectiveness and precision of

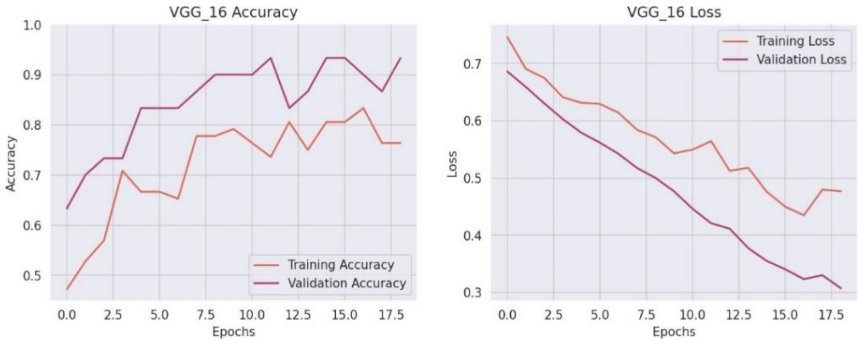


Fig. 5. Results for VGG 16.

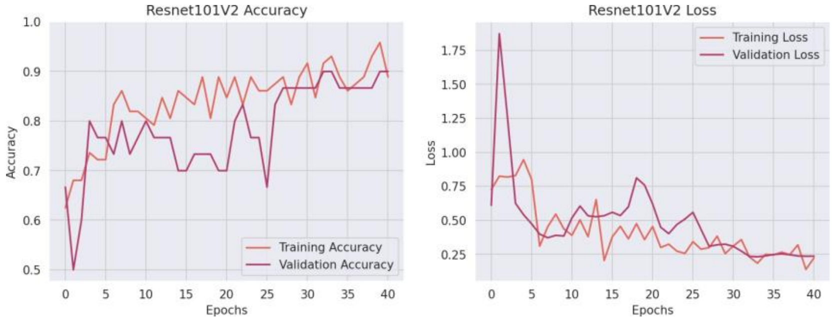


Fig. 6. Results for ResNet 101.

the predictions. By analyzing speech signals, the research aimed to develop a precise and efficient tool for early detection and management of the disease, achieving an accuracy rate near to 82%. This underscores the potential of using speech signals as a diagnostic tool for Parkinson’s disease.

In the same way, Gomez et al. (2023) [6], mentioned that Patients afflicted with Parkinson’s disease (PD) commonly exhibit reduced facial movements, so in their study, they use three distinct approaches are explored to model the facial expressions of individuals with PD: (i) facial analysis using single images as well as sequences of images, (ii) employing transfer learning from facial analysis to recognize action units, and (iii) implementing triplet-loss functions to enhance the automated classification of PD patients and healthy subjects. The investigators also reported 82% as their best accuracy.

2 Methods

2.1 Data Augmentation

The dataset used in this work corresponds to the results achieved by Zhan et al. (2017) [12]. In this research, the study introduces a novel approach by utilizing the Composite Index of Speed and Pen-pressure (CISP) from sketching as a potential feature for

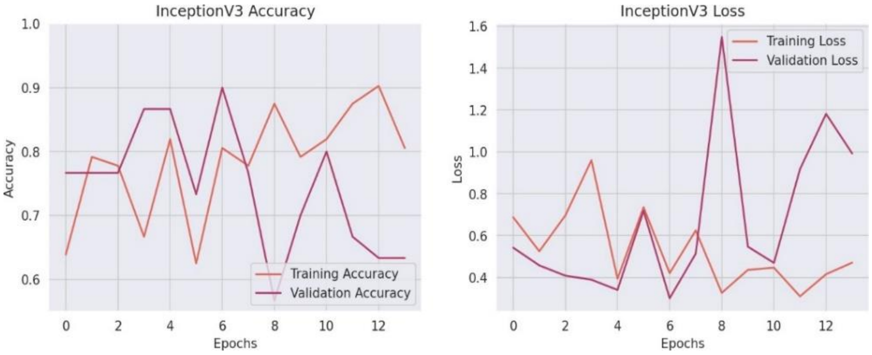


Fig. 7. Results for Inception V3

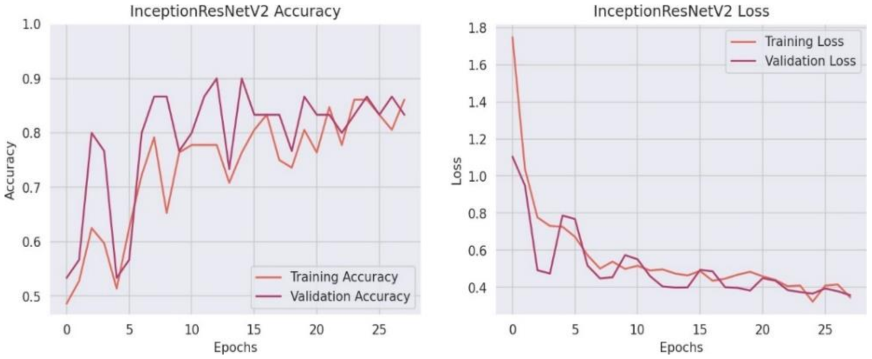


Fig. 8. Results for Inception ResNet.

assessing the severity of Parkinson’s disease (PD). The study involved a total of 55 participants, comprising 28 individuals in the control group (CG) and 27 PD patients. The results include a public database composed of 102 images of sketched waves categorized in two classes: Parkinson and healthy. Given the limited size of the database, data augmentation techniques in the context of image data preprocessing for deep learning were applied.

The first step involves rescaling the pixel values of the images to a standardized range of [0, 1], a routine pre-processing step to ensure that the neural network receives data in a consistent format. The rotation range parameter allows each image to be randomly rotated within a range of 40 degrees in both clockwise and counterclockwise directions. This augments the dataset by introducing variations in object orientations, simulating the real-world diversity of image capture.

Furthermore, the width shift range and height shift range parameters permit horizontal and vertical shifts of up to 20% of the total image width and height, respectively. These mimics the effects of different object placements within the frame, making the model more adaptable to such changes.

Shear transformations are introduced with the shear range parameter, allowing for slanting along the horizontal axis within a 20% range. This emulates perspective changes and adds to the dataset’s diversity. Zooming in and out, a common source of

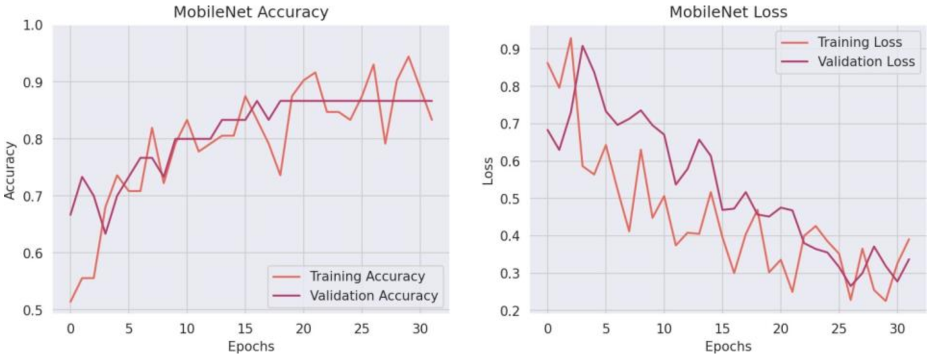


Fig. 9. Results for MobileNet.

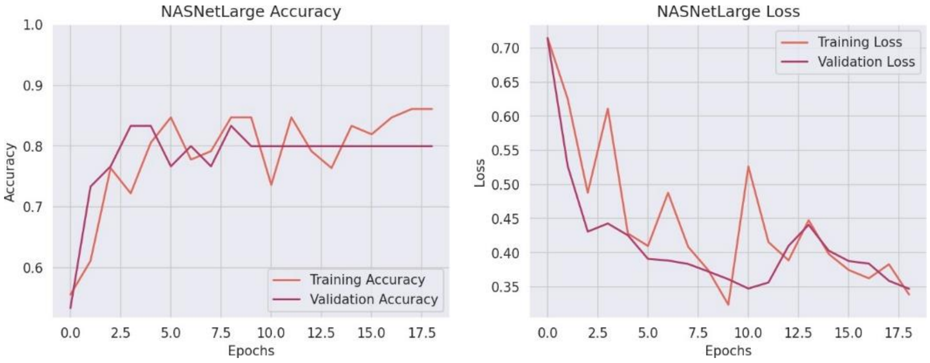


Fig. 10. Results for NasNet large.

variation in real-world images, is achieved with the zoom range parameter set to 20%. This helps the model handle variations in object size and distance. Finally, horizontal flip is enabled, enabling random horizontal flipping of images.

This is useful for scenarios where objects can appear in a mirrored orientation. In summary, these data augmentation techniques are invaluable for enhancing the performance and robustness of deep learning models, particularly when faced with limited training data. By simulating various real-world conditions and image variations during training, the model becomes better equipped to generalize its learning to unseen data, leading to more accurate and reliable predictions in practical applications.

2.2 Transfer Learning with Early Stop Implementation for Binary Classification

A deep learning model tailored for binary classification has been carried out using transfer learning. It leverages the architectures ResNet50, VGG19, DenseNet121, MobileNetV2, VGG16, ResNEt101, InceptionV3, InceptionResNEtV2, MobileNEt, NasNetLarge and ConvNeXtLarge. Those arquitectures come pre-equipped with weights from ImageNet. In all cases, the last 30% layers of the respective architecture were unfrozen for the training stage, while the other layers remained frozen. Additionally, "top" layers were added at the end of the network for binary

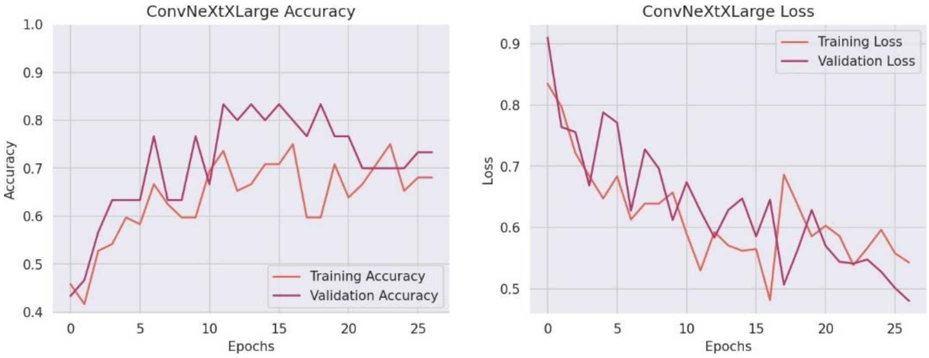


Fig. 11. Results for ConvNext large.

classification. The purpose of this model is to serve as a feature extractor. The base model is then subjected to a process known as fine-tuning, wherein a selective set of its last 30% layers is unfrozen for further training.

This selective approach allows the upper layers to adapt to the task's specific requirements while keeping the remaining layers frozen, preserving the knowledge learned from ImageNet. To form the complete classification model, the base model is integrated into a Sequential architecture. This is achieved through a series of operations, commencing with a Flatten layer, which reshapes the feature maps obtained from the base model into a one-dimensional vector. This is followed by the introduction of two.

Dense layers; the first one incorporates 256 units and a ReLU activation function, while the final layer consists of a single unit with a sigmoid activation function, suitable for binary classification. The Adam optimizer was set with a specified learning rate, the utilization of binary cross-entropy as the loss function, and the adoption of accuracy as the evaluation metric. Finally, an early stopping mechanism is implemented, serving as a safeguard against overfitting by monitoring the validation loss and terminating the training process if the loss fails to improve over a predetermined number of epochs. The mechanism ensures that the model is restored to its most optimal state during training.

3 Results

In Table 2, a summary of the numerical results is found. The performance graphs for loss and accuracy during training and validation stages are presented. The obtained results evaluate the performance of various Convolutional Neural Network (CNN) algorithms for a specific task. The task involved the classification of data, and the metrics used for evaluation included both training and validation accuracy, as well as training and validation loss, along with the corresponding best epoch.

4 Discussion

In table 2, we find the results obtained for train accuracy, validation accuracy, train loss, validation loss, and best epoch. We observe that the best results for validation

Table 2. Results obtained after the model implementation for different model architectures.

CNN Algorithms	Train ACC	Validation ACC	Train Loss	Validation Loss	Best Epoch
ResNet50	0.9538	0.9333	0.1202	0.2997	38
VGG19	0.8750	0.9333	0.3729	0.3049	29
DenseNet121	0.8194	0.9333	0.4092	0.2418	24
MobileNetV2	0.7500	0.9333	0.4959	0.3821	10
VGG16	0.7361	0.9333	0.5642	0.4207	12
ResNet101	0.8889	0.9000	0.2230	0.2359	34
Inceptionv3	0.8056	0.8999	0.4205	0.3009	7
InceptionResNetV2	0.7778	0.8999	0.4959	0.4038	13
MobileNet	0.7917	0.8666	0.4480	0.4726	17
NastNetLarge	0.7222	0.8333	0.6110	0.4428	4
ConvNeXtXLarge	0.7361	0.8333	0.5299	0.6267	12

accuracy were achieved after implementing the model in conjunction with the ResNet50 architecture and the specifications described in the methodology section (2). Regarding the state-of-the-art, Table 1 shows that the best result obtained for the wave analysis task is reported by Vatsaraj and Nagare (2021) [11] with a 96.7% accuracy.

In comparison to the results achieved in this study, the absolute difference between the two results corresponds to 3.27%. Additionally, the same absolute difference is observed after the implementation of the model combined with the architectures VGG19, DenseNEt121, MobileNetV2 and VGG16. Considering the dataset's size, the results reveal consistency in the model. Nevertheless, it is necessary to highlight the fact that the best results in this study were obtained after implementing fewer complex architectures, which is consistent with what was reported by Vatsaraj and Nagare (2021) [5].

Therefore, for future research, it is suggested to use or build networks with fewer convolutional layers in order to get closer to the 96.7% reported in the state-of-the-art. ResNet50, which is known for its depth, demonstrated the highest training accuracy at 95.38%. This suggests that the model learned the training data effectively and can capture complex patterns. However, it exhibited a slightly lower validation accuracy of 93.33%, indicating that it may have encountered some overfitting, as the validation accuracy is slightly lower than the training accuracy.

The training and validation losses of ResNet50 were 0.1202 and 0.2997, respectively, and the best epoch was achieved at 38. These results indicate a good balance between model complexity and generalization, making ResNet50 a strong candidate for this task. VGG19, another deep architecture, demonstrated a relatively high training accuracy of 87.50% but reached an even higher validation accuracy of 93.33%. This suggests that VGG19 achieved good generalization performance. The training and validation losses were 0.3729 and 0.3049, respectively, and the best epoch occurred at 29.

These results indicate that VGG19 managed to generalize well without overfitting, making it a promising choice for this task. DenseNet121 exhibited an 81.94% training accuracy and a 93.33% validation accuracy. Its training and validation losses were 0.4092 and 0.2418, and the best epoch was at 24. DenseNet121 demonstrated robust generalization performance with a slightly lower training accuracy but consistent validation accuracy, suggesting its effectiveness in capturing relevant features.

MobileNetV2, VGG16, and ResNet101 showed varying performance with training and validation accuracies of 75.00% to 88.89%. MobileNetV2 had the highest training loss, indicating room for improvement in its ability to capture features effectively. On the other hand, VGG16 and ResNet101 showed better balance in terms of losses. Inceptionv3, InceptionResNetV2, MobileNet, NastNetLarge, and ConvNeXtXLarge exhibited performance below the 90% accuracy threshold. These models may require further optimization or modifications to improve their classification capabilities.

The results provide insights into the suitability of different CNN architectures for the given task. ResNet50 and VGG19 demonstrated strong performance, while other models showed varying degrees of success. The choice of the best model depends on the specific trade-off between training and validation accuracy, as well as considerations of overfitting and generalization. Further investigations and fine-tuning may be necessary to enhance the performance of some models.

5 Conclusion

Parkinson’s disease is a common neurological condition that can significantly disrupt a patient’s ability to lead a normal life. It is a progressive neurodegenerative disorder that is often challenging to detect in its early stages. Traditional diagnostic methods, particularly those using electroencephalogram (EEG) data, are often time-consuming and challenging to apply in the early stages of the disease. Parkinson’s disease is associated with distinctive motor symptoms, we introduced an approach that analyzes sketching patterns in guided spirals and waves drawn by patients. By applying CNNs, we aimed to facilitate a more efficient and accessible diagnostic process.

This work was based on a dataset provided by Zham et al. (2017) [12] that included sketches categorized as Parkinson’s or healthy. To address the limited size of the dataset, we employed data augmentation techniques, preparing the images for deep learning analysis. Transfer learning was then applied using pre-trained CNN architectures, where the top layers were tailored for binary classification. Regarding the results obtained, ResNet50 and VGG19 exhibited strong performance.

ResNet50 achieved the highest training accuracy. This suggests the model’s ability to capture complex patterns yet, the validation accuracy was slightly lower, indicating the possibility of overfitting. On the other hand, VGG19 showed a good balance between training and validation accuracy.

By another hand, DenseNet121 demonstrated robust generalization performance, making it a promising choice for the Parkinson´s diagnosis. Also, the experimentation showed that MobileNetV2, VGG16, and ResNet101 showed varying performance but indicated potential for improvement. Finally; Inceptionv3, InceptionResNetV2, MobileNet, NastNetLarge, and ConvNeXtXLarge had accuracies below the desired threshold. This suggests that these models may require further optimization or modifications to enhance their classification capabilities.

Considering the previous and comparing the found results to the state-of-theart, we found a discrepancy of 3.27% in accuracy. Notably, Vatsaraj and Nagare (2021) [11] achieved an accuracy of 96.7%. To approach this level of accuracy in future research, it is suggested that models with fewer convolutional layers be explored.

References

1. Chakraborty, S., Aich, S., Jong-Seong-Sim, Han, E., Park, J., Kim, H.: Parkinson’s disease detection from spiral and wave drawings using convolutional neural networks: A multistage classifier approach. In: Proceedings of the 22nd International Conference on Advanced Communication Technology, pp. 298–303 (2020) doi: 10.23919/icact48636.2020.9061497
2. Dabbabi, K., Kehili, A., Cherif, A.: Parkinson detection using VOT-MFCC combination and fully-connected deep neural network (FC- DNN) classifier. In: Proceedings of the IEEE International Conference on Advanced Systems and Emergent Technologies, pp. 1–6 (2023) doi: 10.1109/ic_aset58101. 2023.10150791
3. Gallicchio, C., Micheli, A., Pedrelli, L.: Deep echo state networks for diagnosis of parkinson's disease. In: Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, pp. 1–6 (2018) doi: 10.48 550/ARXIV.1802.06708
4. Gil-Martín, M., Montero, J. M., San-Segundo, R.: Parkinson’s disease detection from drawing movements using convolutional neural networks. Electronics, vol. 8, no. 8, pp. 907 (2019) doi: 10.3390/electronics8080907
5. Goetz, C. G., Stebbins, G. T.: Assuring interrater reliability for the UPDRS motor section: utility of the UPDRS teaching tape. Movement Disorders, vol. 19, no. 12, pp. 1453–1456 (2004) doi: 10.1002/mds.20220
6. Gomez, L. F., Morales, A., Fierrez, J., Orozco-Arroyave, J. R.: Exploring facial expressions and action unit domains for parkinson detection. PLOS ONE, vol. 18, no. 2, pp. e0281248 (2023) doi: 10.1371/journal.pone.0281248
7. Khatamino, P., Canturk, I., Ozyilmaz, L.: A deep learning-CNN based system for medical diagnosis: an application on parkinson’s disease handwriting drawings. In: Proceedings of the 6th International Conference on Control Engineering and Information Technology, pp. 1–6 (2018) doi: 10.1109/ceit.2018.8751879
8. Kumar, S. A., Sasikala, S., Arthiya, K. B., Sathika, J., Karishma, V.: Parkinson’s speech detection using YAMNet. In: Proceedings of the 2nd International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation, pp. 1–5(2023) doi: 10.1109/icaeca56562.2023. 10200704
9. Williams, S., Wong, D., Alty, J. E., Relton, S. D.: Parkinsonian hand or clinician’s eye? Finger tap bradykinesia interrater reliability for 21 movement disorder experts. Journal of Parkinson’s Disease, vol. 13, no. 4, pp. 525–536 (2023) doi: 10.3233/jpd-223256
10. Taylan, P., Yerlikaya-Özkurt, F., Uçak, B. B., Weber, G.: A new outlier detection method based on convex optimization: application to diagnosis of parkinson’s disease. Journal of Applied Statistics, vol. 48, no. 13-15, pp. 2421–2440 (2020) doi: 10.1080/02664763. 2020.1864815
11. Vatsaraj, I., Nagare, G.: Early detection of Parkinson’s disease using contrast enhancement techniques and CNN. International Journal of Engineering Research and Technology, vol. 10, no. 5, pp. 295–298 (2021) doi: 10.17577/IJERTV 10IS050187
12. Zham, P., Kumar, D. K., Dabnichki, P., Arjunan, S. P., Raghav, S.: Distinguishing different stages of Parkinson’s disease using composite index of speed and pen-pressure of sketching a spiral. Frontiers in Neurology, vol. 8 (2017) doi: 10.3389/fneur.2017.00435

ISSN 1870-4069

Modelo computacional para la estimación de crecimiento del Ambystoma mexicanum empleando redes neuronales artificiales

Christian Axel Vera-Cortes, José Juan Carbajal-Hernández,
Luis Pastor Sánchez-Fernández

Instituto Politécnico Nacional,
Centro de Investigación en Computación,
México

cverac2021@cic.ipn.mx

Resumen. En México habitan 17 de las 38 especies reconocidas del género Ambystoma, 15 de ellas se encuentran en la lista de especies amenazadas. Esto se debe a la degradación del hábitat natural de estas especies que ha llevado a la disminución de sus poblaciones y a la amenaza de su supervivencia. Este trabajo propone la creación de un modelo computacional para obtener estimaciones de parámetros biométricos relacionados de la especie Ambystoma mexicanum. Este modelo correlaciona una red neuronal artificial con los parámetros fisicoquímicos con el incremento de peso. Asimismo, se utilizó un modelo autorregresivo para predecir valores medioambientales y alimentar con ellos a la red neuronal. Los resultados del modelo permiten establecer tendencias para determinar el adecuado manejo de la calidad del agua y del alimento, permitiendo obtener mejores tasas de crecimiento del organismo.

Palabras clave: Redes neuronales artificiales, ambystoma mexicanum, acuacultura.

Computational Model for Estimating Growth of Ambystoma Mexicanum Using Artificial Neural Networks

Abstract. Mexico is home to 17 of the 38 recognized species of the genus Ambystoma, 15 of which are on the list of threatened species. This is due to the degradation of the natural habitat of these species that has led to the decline of their populations and the threat to their survival. This study proposes the creation of a computational model to obtain estimates of biometric parameters of the species Ambystoma mexicanum. The model correlates an artificial neural network with physicochemical parameters with weight gain. Likewise, an autoregressive model was used to predict environmental values and feed them to the neural network. The result of the model establishes trends to determine the adequate management of water and food quality, allowing better growth rates of the organism.

Keywords: Artificial neural networks, ambystoma mexicanum, aquaculture.

1. Introducción

1.1. Antecedentes

En México habitan 17 de las 38 especies reconocidas del género de anfibios *Ambystoma* [1]. A las especies de este género se les conoce como ajolotes o achoques, según la región. Se encuentran distribuidas en la zona montañosa que abarca la Sierra Madre Occidental y se une con el Eje Volcánico Transversal [2]. Dentro las especies que habitan en México, 15 se encuentran en alguna categoría de riesgo según la norma NOM-059-SEMARNAT-2010 [3].

La disminución de las poblaciones de *Ambystoma* se debe a la degradación de su ambiente producto del dragado de lagos, la contaminación y la introducción de especies invasoras [2]. La especie más emblemática del género *Ambystoma* es el subgénero *mexicanum*, conocida comúnmente como ajolote de Xochimilco. Esta especie se encuentra en peligro de extinción según la norma NOM-059-SEMARNAT-201.

El rescate de ésta y otras especies de ajolotes se realiza mediante la Secretaría de Medio Ambiente y Recursos Naturales (SEMARNAT) en conjunto con la Comisión Nacional de Áreas Naturales Protegidas (CONANP), quienes ha lanzado un plan de acción para su estudio y preservación [3]. Además, existen proyectos locales para preservar las especies de cada región [3].

Entre estos últimos se encuentra el proyecto Chinampa Refugio de la Universidad Nacional Autónoma de México (UNAM) para la preservación del *A. mexicanum* mediante la creación de áreas seguras al interior del lago de Xochimilco [4]. Por otra parte, para la preservación del *A. altamirani*, o ajolote de arroyo de montaña, la Secretaría de Educación, Ciencia, Tecnología e Innovación (SECTEI) ha impulsado un proyecto para la conservación *ex situ* de la especie trasladando ejemplares de su hábitat natural a una unidad de manejo para la conservación de la vida silvestre (UMA) en el Parque Nacional del Desierto de los Leones [5].

Otro esfuerzo de preservación se encuentra en el municipio de Tecámac, en el Estado de México, donde recientemente se descubrió la presencia del ajolote atigrado, especie que se creía extinta [6]. Finalmente, para la preservación del achoque de Pátzcuaro *A. dumerilii*, el Instituto Nacional de Pesca (INAPESCA) a través de Centro Regional de Investigación Acuícola y Pesquera (CRIAP) Pátzcuaro mantiene un programa para el estudio de la especie en condiciones de laboratorio [7].

Adicionalmente, existen manuales para el cuidado de ajolotes en cautiverio. No obstante, estos escasamente incorporan tecnologías de vanguardia para su estudio. Por lo anterior, es evidente que existe un área de oportunidad en la cual desde las ciencias de la computación se pueden hacer aportaciones para la preservación de la especie.

1.2. Estado del Arte

En el área de acuacultura, predecir el crecimiento de organismos acuáticos tiene gran importancia [8]. La forma más común de predecir el crecimiento de una especie es tomar los datos de un cultivo y ajustarlos a una curva como la función de crecimiento de Gompertz o el modelo de crecimiento de von Bertalanffy. No obstante, estos modelos no consideran la variabilidad de condiciones entre grupos que han crecido en distintas condiciones de cultivo [8]. Por lo anterior, nuevos modelos han sido propuestos que tomen en consideración otras variables.

Entre los modelos propuestos se encuentra uno basado en redes neuronales artificiales que predicen el crecimiento de camarón a partir de datos de alimentación y temperatura [9]. Ampliando el modelo anterior, se desarrolló un modelo que, además de la alimentación y la temperatura, incorpora la edad del camarón, la temperatura del agua y la densidad larvaria como datos de entrenamiento [10]. De manera general, para cualquier especie acuática criada mediante acuacultura, se ha propuesto un modelo de crecimiento basado en redes neuronales que utilicen los parámetros fisicoquímicos de calidad del agua como entrada al sistema [11].

Otros trabajos similares son: el uso de un procedimiento empírico bayesiano para predecir el crecimiento de camarón [12]; el uso de un modelo jerárquico bayesiano para modelar la variabilidad en modelos de crecimiento de camarón [8]; redes neuronales para predicción del crecimiento de peces [13] y para correlacionar peso y longitud en langostas [14].

Finalmente, es importante señalar que al investigar sobre trabajos similares realizados para alguna especie del género Ambystoma, sólo se encontró el uso del modelo de von Bertalanffy para estimar la edad de los ajolotes cuando se hacían estudios en especímenes silvestres [15], pero no se encontró ninguna investigación enfocada en su totalidad en desarrollar un modelo de crecimiento.

1.3. Contribución

En este trabajo se presenta un modelo para la predicción de crecimiento del *A. mexicanum*, el cual se pretende sea utilizado por los criadores de la especie para tomar decisiones respecto a su cuidado y crianza. El modelo se hizo a partir de los parámetros fisicoquímicos de calidad del agua, así como datos sobre alimentación. Para generar el modelo anterior se utilizó una red neuronal artificial del tipo perceptrón multicapa entrada mediante el algoritmo de propagación hacia atrás. Por otra parte, para hacer predicciones a futuro se emplearon modelos autorregresivos para hacer una predicción de los parámetros fisicoquímicos.

2. Requerimientos

Para criar al *A. mexicanum* en cautiverio, es necesario contar con las instalaciones adecuadas que incluyan: un acuario o estanque donde habitará el ajolote y que debe contar un buen filtrado, proporcionar una buena iluminación y sustrato [16]. Además, al ser el ajolote una especie acuática, es de vital importancia mantener una calidad del agua adecuada para el correcto desarrollo de la especie [16].

La calidad del agua se mide a través de los parámetros fisicoquímicos de la misma. Los parámetros más importantes que se deben considerar son: potencial de hidrógeno (pH), dureza general (GH), dureza en cuanto a carbono (KH), concentración de nitritos y nitratos, concentración de amonio, concentración de CO_2, porcentaje de oxígeno disuelto, concentración de cloro y temperatura. En la literatura se reportan rangos dentro de los cuales se considera que el agua es segura para la especie en su etapa adulta. Dichos rangos pueden observarse en la Tabla 1.

Tabla1. Rangos recomendados de valores de los parámetros fisicoquímicos de calidad del agua para la crianza del Ambystoma mexicanum.

Parámetro	**Unidad**	**Rango**
pH	---	6.5 – 8
Cloro	mg/l	0
Dureza General (GH)	dh	6 – 16°
Dureza en cuanto a carbono (KH)	dh	3 – 10°
Nitritos (NO_2-)	mg/l	< 1
Nitratos (NO_3-)	mg/l	< 20
Amonio	mg/l	< 0.1
Concentración de CO_2	mg/l	< 0.5
Saturación de oxígeno disuelto (OD)	%	≥ 80
Densidad	---	1
Temperatura	Celsius	10° – 18°

La temperatura es considerada una de las variables de calidad del agua más importantes [16]. Se reporta que, en especímenes jóvenes, temperaturas de hasta 25°C aceleran el desarrollo, aunque a medida que alcanzan la edad adulta este efecto desaparece [16] y es necesario mantener la temperatura en los rangos mostrados en la Tabla 1. El pH afecta el metabolismo de todos los organismos acuáticos [17].

El amonio no ionizado es el principal compuesto excretado por los animales acuáticos como producto de su proceso metabólico [18]. Un exceso de amonio en el agua hace más difícil para los organismos su excreción, lo que puede llevar a la reducción o paralización de la actividad alimenticia [17]. También hace que los ajolotes sean más susceptibles a enfermedades e inhibe el crecimiento [18].

Los niveles de pH y amonio están, además, estrechamente relacionados entre sí y con los niveles de nitritos y nitratos, ya que estos últimos se forman como consecuencia de las reacciones químicas de los dos primeros. Finalmente, se sabe que el porcentaje de oxígeno disuelto también influye en el metabolismo de los animales acuáticos.

3. Preprocesamiento

3.1. Adquisición de datos

Para este estudio se tomaron los datos reportados en “Mantenimiento en cautiverio de *Ambystoma mexicanum* con dietas enriquecidas con selenio” [19]. Dicho reporte proporciona los datos de un experimento que se realizó durante tres meses. En éste se comparó el incremento de peso de tres grupos de ajolotes sometidos a tres regímenes de alimentación distintos. Todos los grupos tuvieron una dieta a base de tubifex, pero

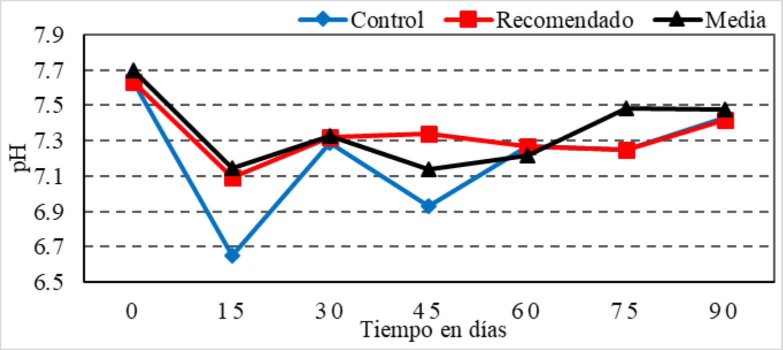


Fig.1. Datos reportados de pH durante un periodo de 90 días.

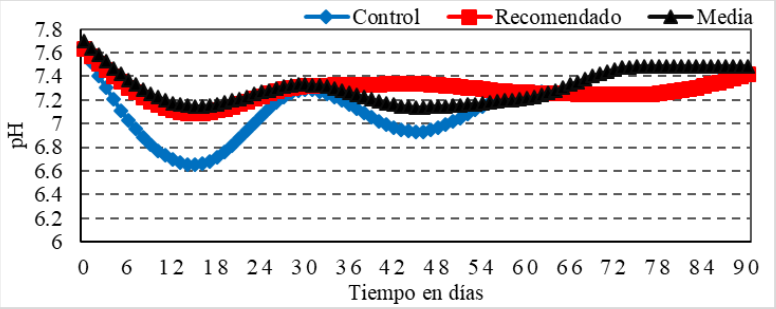


Fig. 2. Resultado de la interpolación para la variable pH durante un periodo de 90 días.

a dos de ellos se les añadió un suplemento de selenio. A uno de estos grupos se les proporcionó la dosis del suplemento recomendada y al otro la mitad de la dosis recomendada.

Los parámetros fisicoquímicos reportados fueron: pH, amonio, nitritos, nitratos y temperatura. Mientras que las variables biométricas reportadas son: peso, incremento de peso, biomasa y longitud total. No obstante, de esta última solo se reportaron los valores iniciales y finales. Para todas las variables se reportó el promedio quincenal por grupo, lo que da un total de 7 muestras por cada variable. Aunque el conjunto de parámetros es un poco grande, en la Fig. 1 se muestra la gráfica del pH para ejemplificar el tratamiento de los datos efectuados en este trabajo.

3.2. Interpolación

Debido a que la cantidad de datos reportados por parámetro es pequeña, se realizó un remuestreo para estimar los datos por día. Para este fin, se empleó una interpolación por trazadores cúbicos. Se eligió esta técnica después de probar otros métodos [20, 21] y encontrar que era la que arrojaba los mejores resultados para todas las variables.

La interpolación por trazadores cúbicos consiste en obtener un polinomio de tercer grado para cada intervalo entre dos nodos [20] como se muestra a continuación:

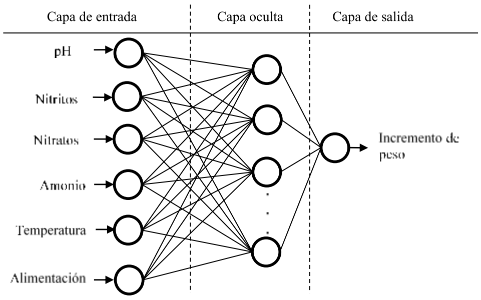


Fig. 3. Arquitectura de la red neuronal.

$$f(x) = ax^3 + bx^2 + cx + d. \tag{1}$$

La interpolación se hizo con el software MATLAB. Dicho software implementa el conjunto subrutinas PCHIP [21]. En la Fig. 2 se puede apreciar el resultado de la interpolación para la variable de pH.

3.3. Normalización

Las redes neuronales requieren de un conjunto de datos con la misma escala, debido a que los parámetros por su naturaleza presentan valores a diferentes escalas, se realizó un proceso de normalización de la siguiente forma:

$$I_i = (L_M - L_m) \times ((p_i - p_{mín})/(p_{máx} - p_{mín})) - L_{m,} \tag{2}$$

donde I es el vector de parámetros normalizados, L_M y L_m son los límites superior e inferior del rango en el que se quiere normalizar, p_i es el i-ésimo parámetro a normalizar y $p_{máx}$ y $p_{mín}$ son los valores máximo y mínimo de los parámetros a normalizar.

Para este trabajo se optó por una normalización de [-1, 1], debido a que es la comúnmente utilizada [24]. Sustituyendo en la ecuación (2) se obtiene:

$$\mathrm{I_i} = (1 - (-1)) \times ((p_i - p_{mín})/(p_{máx} - p_{mín})) - 1. \tag{3}$$

La ecuación (3) se aplicó a cada una de las variables empleadas en el estudio, tanto a los parámetros fisicoquímicos como a las variables biométricas.

4. Modelo

4.1. Diseño de la red neuronal artificial

En el presente trabajo, se utilizaron redes neuronales artificiales para correlacionar los parámetros fisicoquímicos de calidad del agua con el incremento de peso del *Ambystoma mexicanum*. Se seleccionaron las RNA debido a que son capaces de aproximar cualquier función continua [22]. Además, resultan útiles para resolver problemas en los que las relaciones entre variables no son evidentes [22].

Las RNA se componen de un conjunto de unidades interconectadas conocidas como neuronas artificiales [23]. Cada una de estas neuronas realiza una suma ponderada de

Tabla 2. Orden de los modelos AR obtenido con el criterio FPE para cada parámetro fisicoquímico.

Parámetro	**Grupo**	**Orden**
	Control	17
pHa	Recomendada	6
	Media	6
	Control	18
Nitritos	Recomendada	18
	Media	18
	Control	17
Nitratos	Recomendada	18
	Media	18
	Control	19
Amonio	Recomendada	18
	Media	21
	Control	23
Temperatura	Recomendada	6
	Media	6

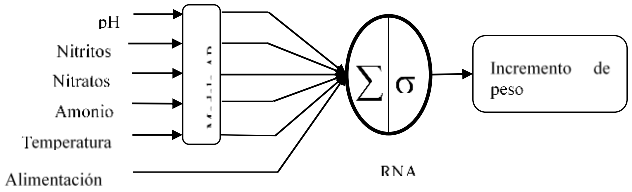


Fig. 4. Modelo computacional para la predicción del incremento de peso.

los valores de entrada multiplicándolos por coeficientes conocidos como pesos *w* y sumando un valor adicional, conocido como sesgo *b*. El resultado de la suma es introducido a una función matemática conocida como función de activación.

Las RNA están organizadas en capas: en la primera, llamada capa de entrada, se reciben los valores de entrada. En este trabajo los valores de entrada son los parámetros fisicoquímicos y el tipo de régimen de alimentación.

Después, se encuentra una capa oculta. El número de neuronas de esta capa será seleccionado tras una serie de pruebas que se detallarán más adelante. Para esta red se ha escogido una tangente hiperbólica como función activación dada por la siguiente ecuación:

$$f(x) = 2 / (1 + e^{-2x}) - 1. \tag{4}$$

Tras la capa oculta se encuentra la capa de salida. Dicha capa entregará el valor de incremento de peso estimado. Debido a que se entrega un único valor a la salida sólo se necesita una neurona. Ésta tiene una función de activación lineal, comúnmente usada en la capa de salida para problemas de regresión [24], dada por la siguiente expresión:

Tabla 3. Error cuadrático medio para los conjuntos de entrenamiento, validación y prueba de las RNA para la estimación del incremento de peso.

Variable	Número de Neuronas	Entrenamiento	Validación	Prueba
Incremento de peso	2	0.002466276	0.005598149	0.007224309
	3	1.049401056	0.994811983	2.1967806
	4	0.004231471	0.111931298	0.0063346
	5	0.019030662	0.019389994	0.019100361
	6	0.001663586	0.002664131	0.007169914
	7	0.00200717	0.004276588	0.008812153
	8	0.009628086	0.015219814	0.01630441
	9	0.009355713	0.013654767	0.014383261
	10	0.000539841	0.001046942	0.00069038

$$f(x) = x. \tag{5}$$

La arquitectura utilizada se puede observar en la Fig. 3. A este tipo de arquitectura se le conoce como perceptrón multicapa.

4.2. Entrenamiento de la red neuronal artificial

Para determinar los pesos *w* y sesgos b de la RNA del perceptrón multicapa se utiliza el algoritmo de propagación hacia atrás [24]. El objetivo del algoritmo es minimizar el error cuadrático medio (MSE) entre el valor deseado *t* y la salida de la red *a* [25]. El MSE está dado por la ecuación (6):

$$E = (t - a)^{2.} \tag{6}$$

Para minimizar el MSE se modifican los pesos *w* y sesgos *b* de las neuronas. Existen distintos algoritmos para optimizar el entrenamiento. Debido a que el número de parámetros y datos a emplear es reducido, se utilizó la optimización por Levenberg Marquardt. Ésta es más rápida que otros algoritmos para una cantidad moderada de parámetros, aunque tiene como desventaja los requerimientos de almacenamiento [24].

Para el entrenamiento se separó el conjunto de entrada en tres: entrenamiento, prueba y validación. El primer subconjunto contiene el 70% de los datos, es decir, 63 muestras por cada variable, mientras que los dos últimos tienen 15% de los datos cada uno, es decir, entre 13 y 14 muestras por variable. La división se hace aleatoriamente.

Para determinar el número de neuronas de la capa oculta se realizaron una serie de pruebas que consistieron en entrenar una RNA con un número de neuronas de 2 a 10 y seleccionar la que mostrase un mejor desempeño, o bien aquella a partir de la cual el desempeño no mostrase mejorías.

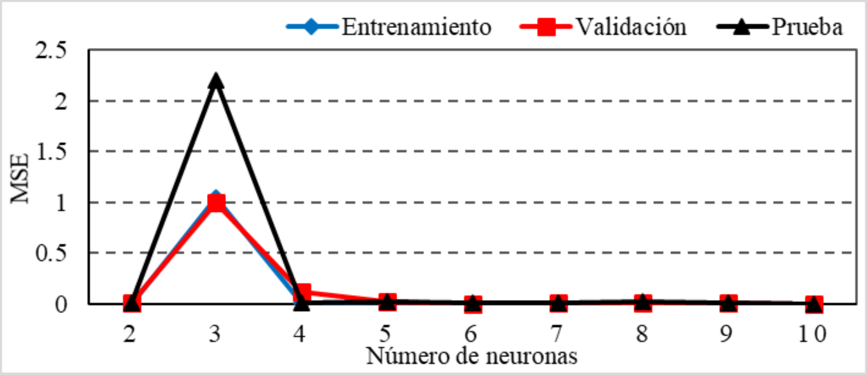


Fig. 5. Evolución del MSE según incrementa el número de neuronas de la RNA para la estimación del incremento de peso.

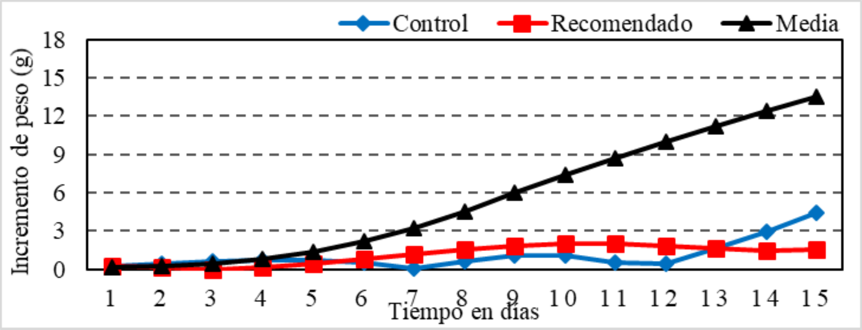


Fig. 6. Error absoluto de la predicción del incremento de peso.

4.3. Predicción

Para hacer predicciones a futuro de los parámetros fisicoquímicos se desarrollaron modelos autorregresivos para cada uno de los parámetros. Un modelo autorregresivo (AR) es una representación de una señal aleatoria en la cual los valores actuales dependen linealmente de los valores pasados. La función general de un modelo AR está dada por la ecuación (7):

$$p_t = \theta_1 p_{t-1} + \theta_2 p_{t-2} + \ldots + \theta_i p_{t-i} + \ldots + \theta_n p_{t-n} + \varepsilon_{t}, \tag{7}$$

donde p es el valor de la serie de tiempo de cada uno de los parámetros fisicoquímicos, θ_i es el i-ésimo coeficiente, ε_t es ruido blanco, t es el tiempo de la predicción y n es el orden del modelo.

El orden del modelo es el número de valores previos que se usarán para el cálculo de los coeficientes. Se obtuvo mediante el criterio del error final de predicción (FPE) dado por la ecuación (8):

$$\text{FPE} = V_n\,(1 + 2n/(\text{L} - n)\,), \tag{8}$$

donde L es el número de datos en la serie de tiempo, n es el orden del modelo y V_n es el error de predicción. Los coeficientes obtenidos se pueden observar en la Tabla 2.

El cálculo de los coeficientes AR se realizó mediante el método Yule-Walker implementado en LabVIEW. Una vez obtenidos los modelos AR, se hicieron predicciones de los parámetros que posteriormente fueron introducidos a la RNA. La Fig. 4 ilustra el proceso.

5. Resultados

Los resultados del entrenamiento de la red pueden observarse en la Tabla 3. Se puede apreciar que, para los tres conjuntos, el MSE es menor que 1 en la mayoría de los casos. Para apreciar más claramente la evolución del error, los resultados se muestran en la Fig. 5. En esta se observa que, a partir de 5 neuronas el error deja de disminuir significativamente.

En la Fig. 6 se puede observar el error absoluto de la salida de la RNA al introducir predicciones de los modelos AR y los valores reales para cada uno de los tres regímenes alimenticios. Como se puede observar el error incrementa a medida que aumentan el número de predicciones. Especialmente el grupo de dosis media muestra un aumento significativo después de 7 días de predicción. Por lo que el modelo puede hacer predicciones para este periodo de tiempo como máximo.

6. Conclusiones

En este trabajo se mostró que una red neuronal artificial es capaz de correlacionar los parámetros fisicoquímicos con el incremento de peso. Como se pudo observar, el error de la RNA decrece rápidamente aun cuando el número de neuronas en la capa oculta es pequeño. Esto se debe a que la muestra de datos es pequeña. Además, sólo fue necesaria una única capa oculta por la misma razón. No obstante, es importante mencionar que los datos utilizados fueron tomados durante un periodo corto de tiempo, además, hay parámetros fisicoquímicos importantes que no se tomaron en consideración debido a la carencia de datos disponibles.

Por lo anterior, es importante hacer estudios con datos tomados durante períodos más prolongados que abarquen todo el ciclo de vida del ajolote. Además, es necesario que se tomen medidas de variables importantes que no se consideraron en la realización de los modelos debido a la falta de información. Finalmente, es importante recordar que la mayoría de las especies del género *Ambystoma* que existen en México están amenazadas y que es posible adaptar este modelo a otras especies.

Referencias

1. Amphibian Species of the World: An online reference. Version 6.1, https://amphibiansoftheworld.amnh.org/index.php, último acceso 2023/03/06
2. Shaffer, H. B.: Natural history, ecology and evolution of the Mexican "axolotls.". Axolotl Newsletter, vol. 18, no. 5, pp. 5–11 (1989)

3. SEMARNAT: Programa de acción para la conservación de las especies ambystoma spp, SEMARNAT/CONANP (2018)
4. Olguín-Lacunza, M. A., Torres, R.: El ajolote de Xochimilco, a punto de la extinción, UNAM Global (2023)
5. Secretaría de Educación, Ciencia, Tecnología e Innovación, https://ciencia.sectei.cdmx.gob.mx/2023/10/12/proyecto-8164/, último acceso 2023/11/14
6. Periódico La Jornada: Buscan con IA contribuir al rescate del ajolote atigrado. La Jornada, (2023)
7. Secretaría de Agricultura y Desarrollo Rural, https://www.gob.mx/agricultura/articulos/inapesca-al-cuidado-del-achoque?idiom=es
8. Yu, R., Leung, P.: A Bayesian hierarchical model for modeling white shrimp (Litopenaeus vannamei) growth in a commercial shrimp farm. Aquaculture, vol. 306, no. 1–4, pp. 205–210 (2010) doi: 1016/j.aquaculture.2010.04.028.
9. Yu , R., Leung, P., Bienfang, P.: Predicting shrimp growth: Artificial neural network versus nonlinear regression models. Aquacultural Engineering, vol. 34, no. 1, pp. 26–32 (2006) doi: 10.1016/j.aquaeng.2005.03.003
10. Esmaeili, A., Tarazkar, M. H.: Prediction of shrimp growth using an artificial neural network and regression models. Aquaculture International, vol. 19, pp. 705–713 (2011) doi: 10.1007/s10499-010-9386-8
11. Deng, C., Gao, Y., Gu, J., Miao, X., Li, S.: Research on the growth model of aquaculture organisms based on neural network expert system. In: 2010 Sixth International Conference on Natural Computation, vol. 4, pp. 1812–1815 (2010) doi: 10.1109/ICNC.2010.5584492
12. Whiting, D. G., Tolley, H. D., Fellingham, G. W.: An empirical Bayes procedure for adaptive forecasting of shrimp yield. Aquaculture, vol. 182, no. 3–4, pp. 215–228 (2000) doi: 10.1016/S0044-8486(99)00263-X
13. Benzer, R.: Population dynamics forecasting using artificial neural networks. Fresenius Environmental Bulletin, vol. 24, no. 2, pp. 460–466 (2015)
14. Benzer, S., Benli, Ç. K., Benzer, R.: The comparison of growth with length-weight relation and artificial neural networks of crayfish, Astacus leptodactylus, in Mogan Lake. Journal Black Sea/Mediterranean Environment, vol. 21, no. 2, pp. 208–223 (2015)
15. Zambrano-González, L., Reynoso, V. H., Herrera, G.: Abundancia y estructura poblacional del axolotl (Ambystoma mexicanum) en los sistemas dulceacuícolas de Xochimilco y Chalco. Informe final SNIB-CONABIO proyecto No. AS004 (2004)
16. Servín-Zamora, E.: Manual de mantenimiento en cautiverio y medicina veterinaria aplicada al ajolote de Xochimilico (Ambystoma mexicanum) en el zoológico de Chapultepec. Tesis de Licenciatura, Universidad Nacional Autónoma de México (2011)
17. Vinatea-Arana, L.: Principios químicos de calidad del agua en acuicultura: una revisión para peces y camarones. UAM-Xochimilco (2006)
18. Corona-Salto, A.: Calidad del agua tratada por un humedal artificial para su uso en cultivo del ajolote Ambystoma Mexicanum shaw (Amphibia urodela) en Xochimilco D.F. Universidad Autónoma Metropolitana (2012)
19. Espinosa-Román, O.: Mantenimiento en cautiverio de Ambystoma mexicanum con dietas enriquecidas con selenio. UAM-Xochimilco (2019)
20. Chapra, S. C., Canale, R. P.: Métodos numéricos para ingenieros. 5ta edn. McGraw-Hill Interamericana, México (2007)
21. The MathWorks, Inc. https://la.mathworks.com
22. Zhang, G., Patuwo, B. E., Hu, M.: Forecasting with artificial neural networks: The state of the art. International Journal of Forecasting, vol. 14, no. 1, pp. 35–62 (1998)
23. Yáñez-Márquez, C., López-Leyva, L. O., Aldape-Pérez, M.: Neurona artificial de McCulloch & Pitts. Instituto Politécnico Nacional, Centro de Investigación en Computación (2007)

24. Hagan, M. T., Demuth, H. B., Beale, M. H., De-Jesús, O.: Neural network design. 2nd edn. Martin Hagan, Oklahoma (2014)
25. Christiansen, N. H., Voie, E. T., Winther, O., Høgsberg, J.: Comparison of Neural Network Error Measures for Simulation of Slender Marine Structures, Journal of Applied Mathematics vol. 4,pp. 1–11 (2014) doi: 10.1155/2014/759834

Electronic edition
Available online: http://www.rcs.cic.ipn.mx

http://rcs.cic.ipn.mx

Centro de Investigación
en Computación