EDUCACIÓN
SECRETARÍA DE EDUCACIÓN PÚBLICA

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

Research in Computing Science

Vol. 151 No. 6
June 2022

Research in Computing Science

Research in Computing Science, Año 21, Volumen 151, No. 6, junio de 2022, 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 2022.

Research in Computing Science, year 21, Volume 151, No. 6, June 2022, is published monthly by the Center for Computing Research of IPN.

Volume 151(6)

Advances in Artificial Intelligence

Gilberto Ochoa-Ruiz (ed.)

Instituto Politécnico Nacional, Centro de Investigación en Computación
México 2022

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

Indexed in LATINDEX, DBLP and Periodica

Electronic edition

ISSN 1870-4069

Table of Contents

Page

ISSN 1870-4069

Problemas filosóficos y sociales en la implementación de sistemas opacos de IA en el área de la salud

Diego Vázquez Díaz

Universidad Nacional Autónoma de México
Facultad de Filosofía y Letras,
México

dikvaz@gmail.com

Resumen. Muchas de las guías éticas y regularizaciones en el uso de sistemas de Inteligencia Artificial (IA) han asumido el compromiso de asegurar la transparencia, explicabilidad e inteligibilidad de las tecnologías para sus usuarios. Sin embargo, esta búsqueda tiene como límite la opacidad que conlleva el alto nivel de complejidad de las tecnologías y los índices de analfabetismo tecnológico por parte de los agentes humanos que interactúan con los sistemas. Este problema se vuelve especialmente importante cuando hablamos de las aplicaciones clínicas de la IA si partimos del supuesto de que es necesario promover prácticas médicas centradas en el paciente que aseguren que estos pueden tomar decisiones sobre su salud de manera informada. En el presente trabajo se exploran algunas problemáticas filosóficas y sociales que conlleva la implementación de estas tecnologías en el área biomédica, así como algunas consideraciones éticas y epistemológicas para su resolución.

Palabras clave: Opacidad epistémica, práctica médica centrada en el paciente, transparencia, analfabetismo.

Philosophical and Social Problems of Medical Applications of Opaque AI Systems

Abstract. Many of the ethical guidelines and regulations on the use of AI systems have made a commitment to ensure the transparency, explainability and intelligibility of the technologies for their users. However, this search is limited by the opacity that comes with the high level of complexity of the technologies and the rates of technological illiteracy of those who interact with the systems. This problem becomes especially important when we talk about the clinical applications of AI if we state that it is necessary to promote patient-centered medical practices that ensure that patients can make informed decisions about their health. In the present work, some philosophical and social problems involved in the implementation of these technologies in the biomedical area are explored, as well as some ethical and epistemological considerations for their resolution.

Keywords: Epistemic opacity, patient-centered medical practices, transparency, analphabetism.

1. Introducción

En su guía para la Ética y Gobernanza de la Inteligencia Artificial para la Salud [1], publicada en el mes de junio de 2021, la Organización Mundial de la Salud (OMS) afirmó que la futura implementación de las tecnologías de Inteligencia Artificial (IA) en el área médica deberá considerar, al menos, seis factores de relevancia bioética a saber:

1. la protección de la autonomía,
2. la promoción del bienestar humano y de los intereses públicos,
3. el aseguramiento en la transparencia, explicabilidad e inteligibilidad de estas tecnologías,
4. el afincamiento de responsabilidades y rendición de cuentas,
5. la garantía en la inclusividad e igualdad,
6. la búsqueda de Inteligencias Artificiales sensibles y sostenibles.

En este artículo exploraré de manera específica las consideraciones sociales y filosóficas relevantes al tercer factor de interés bioético enlistado por la OMS para el uso de sistemas de Inteligencia Artificial en el área médica; a saber, el problema de la transparencia, explicabilidad e inteligibilidad de las tecnologías. En la guía, la OMS afirma que el aseguramiento de este factor, así como de los otros enumerados en el listado, dependerá en gran medida de un esfuerzo colectivo para diseñar e implementar políticas bioéticamente defendibles en favor de los intereses de los pacientes y de las comunidades.

Sin embargo, como afirman Bjerring y Busch, una práctica médica centrada en el paciente es incompatible con los usos de tecnologías de Inteligencia Artificial y, específicamente, de Aprendizaje Profundo (que serán referidas con la abreviatura IA/AP en este artículo) en el área de salud [2]. Esto se debe a que, generalmente, los usuarios finales de estas herramientas desconocen la operatividad de los sistemas utilizados. Esto limita la comprensión de la relación que se establece entre tratante y paciente, debido a que, en muchos casos, ninguno de los dos es consciente del funcionamiento de los dispositivos y, por tanto, existe una opacidad en el uso de los sistemas para la procuración de la salud.

En el primer apartado de este trabajo propondré una definición de esta falta de transparencia, a la que me referiré como «opacidad epistémica». Esta opacidad ha llamado especialmente la atención de los organismos reguladores y de las agencias defensoras de los Derechos Humanos. La organización Artículo 19 ha afirmado, por ejemplo, que existe y debe promoverse una libertad de conocimiento que consiste en "el derecho a demandar y recibir información de los ostentadores del poder para la transparencia en la buena Gobernanza y para el desarrollo sostenible" [3].

La misma organización ha hecho evidente que esta opacidad e inescrutabilidad ha dado pie a diversos esfuerzos para hacer a las tecnologías de IA transparentes. Sin embargo, como afirma Vidushi Marda [4], en realidad no existe un consenso respecto a las formas de transparencia que debemos buscar. En el primer apartado de este trabajo caracterizaré algunos tipos de opacidad que limitan nuestra comprensión del

funcionamiento de estos sistemas. Posteriormente, argumentaré que estas limitantes imposibilitan el aseguramiento de la transparencia, explicabilidad e inteligibilidad en el uso de estas tecnologías, lo cual contraviene las aspiraciones en la regulación de la IA en el área de salud.

Para finalizar y como respuesta a este problema, presentaré algunas estrategias basadas en la buena gobernanza digital que pueden ayudar a promover la toma de decisión informada por parte de los pacientes a pesar de la existencia de esa opacidad, y presentaré algunas consideraciones finales.

2. Marco filosófico-conceptual

Podemos definir a los sistemas de Inteligencia Artificial como sistemas que pueden emular, aumentar o competir con el desempeño de humanos inteligentes en tareas específicas. Por su parte, podemos hablar de Aprendizaje Profundo (AP) cuando las tecnologías de aprendizaje automático son capaces de procesar información compleja (como imágenes o sonidos) mediante transformaciones múltiples para hacer más efectivo el entrenamiento y aprendizaje del sistema. La particularidad de estos sistemas es que no solamente sirven para el procesamiento de la información de acuerdo con valores ingresados por los agentes humanos, sino que son capaces de aprender y modificar sus propios valores para la obtención de resultados cada vez más precisos y exitosos.

A través del procesamiento iterativo de información, estos sistemas tienen la habilidad para extraer sus propias variables y asignarles pesos específicos para el análisis e interpretación de los datos. Al procesar amplios volúmenes de información en espacios de tiempo cada vez más reducidos, estas tecnologías representan importantes bastiones del aumento en la complejidad de los sistemas computacionales que, en múltiples sentidos, han superado las capacidades de los agentes humanos y que, por tanto, obstaculizan un derecho al conocimiento de las tecnologías que sirven a la gestión de la salud pública e individual.

Debido en parte al analfabetismo digital, a las políticas de privacidad de las empresas tecnológicas y a las capacidades cognitivas de los agentes humanos, actualmente nos encontramos rodeados de tecnologías que al estar condicionadas por *opacos y crípticos principios y mecanismos* [5] impiden que, en muchos casos, sus usuarios las comprendamos a cabalidad. Si bien es común hacer referencia a estas limitantes mediante la denominación de las tecnologías de IA/AP como “cajas negras”, para los fines de este trabajo llamaremos a este fenómeno el *problema de la opacidad epistémica*. Podemos afirmar que un proceso es epistémicamente opaco en relación con un agente X en un momento t en caso de que X no conozca todos los elementos epistemológicamente relevantes del proceso [6].

Esta definición de opacidad epistémica sería planteada por Paul Humphreys en su estudio sobre la necesidad por pensar filosóficamente los problemas que conlleva la relación entre humanos y máquinas de computo en la producción de conocimiento. De acuerdo con Jenna Burrell [7], es necesario distinguir, al menos, tres tipos de opacidad que se presentan en los sistemas de Inteligencia Artificial y, particularmente, de Aprendizaje Profundo: (1) la opacidad que surge de las características de los algoritmos computacionales y de la escala de información requerida para su efectiva utilización;

(2) la opacidad de los sistemas debido al analfabetismo tecnológico; y (3) la opacidad de los algoritmos deliberadamente creada por las corporaciones (o por los mismos desarrolladores) debido a políticas de seguridad, propiedades intelectuales o códigos comerciales.

Tomando como punto de partida la clasificación de Burrell, partiré de la hipótesis de que la creación de códigos bioéticos y regulaciones que promuevan la transparencia, explicabilidad e inteligibilidad en el uso de sistemas de Inteligencia Artificial en el área de la salud requiere una exploración de las características específicas de cada uno de estos tipos de opacidad epistémica en relación con los miembros de las comunidades tecnológicas en las que serán utilizadas.

Para ello, partiré de que el conjunto de los elementos epistémicamente relevantes de un proceso o sistema son contingentes y relativos a las necesidades de diferentes agentes humanos; por tanto, argumentaré que la transparencia buscada en defensa del derecho al conocimiento también debe considerar hacia quiénes se dirige y de qué modo se puede promover.

Afirmaré que, de acuerdo con esta hipótesis, las características de la opacidad epistémica de cada contexto tecnológico conllevan limitantes para la búsqueda de transparencia, explicabilidad e inteligibilidad que deben ser consideradas para la formulación de principios bioéticos. En última instancia, defenderé que estas consideraciones para el uso de sistemas de IA/AP en el área de la salud no pueden ni deben comprometerse con una absoluta transparencia, explicabilidad e inteligibilidad en su uso si quieren promover al mismo tiempo una práctica médica centrada en el paciente en la que este sea capaz de tomar decisiones informadas respecto a su salud.

3. La opacidad epistémica como producto de las limitantes cognitivas

Los agentes humanos de las comunidades digitales contemporáneas nos encontramos en medio de una contradicción que parece ser infranqueable. Por un lado, en nuestras culturas hemos aceptado la idea de que el derecho al conocimiento es absoluto e ilimitado, mientras que, por el otro, nos enfrentamos con la creciente incapacidad de conocer a cabalidad las tecnologías con las que día a día convivimos.

Si bien esta incapacidad no es un problema exclusivo del siglo XXI, podemos afirmar con suficiente seguridad que la emergencia de las tecnologías digitales y, específicamente, de Inteligencia Artificial y de Aprendizaje Profundo, ha ensanchado la brecha existente entre las capacidades humanas de comprensión del mundo y el estado real de cosas en él.

De acuerdo con Humphreys, la opacidad epistémica no es un problema exclusivo de las ciencias computacionales, sino que es una cuestión que compete a la filosofía de la ciencia que se ha preguntado por cómo conocemos a través de instrumentos científicos. No obstante, esta acotación, la opacidad sí es un problema particularmente comprometedor para la informática computacional ya que, contrarios a los instrumentos de medición y representación analógicos, “ningún humano puede examinar y justificar cada elemento de los procesos computacionales que producen un valor de salida o de otros artefactos de las ciencias de la computación” [6].

El argumento central del problema de la opacidad epistémica radica en que los algoritmos computacionales tienen tantos pasos y los sistemas procesan tan amplias cantidades de información que resultan inaccesibles e impenetrables para los agentes cognitivos humanos y, por tanto, que la creencia en sus resultados termina siendo imposible de justificar.

Durán y Formanek [8], por el contrario, han afirmado que, en realidad, sí contamos con diferentes recursos para generar confianza en los sistemas computacionales que, en consecuencia, podrían justificar nuestras creencias. De acuerdo con los autores podemos identificar al menos cuatro fuentes que nos permiten atribuir fiabilidad a los sistemas computacionales:

1. los métodos de verificación y validación,
2. el análisis en la robustez,
3. la historia de sus implementaciones exitosas/no exitosas y
4. el conocimiento experto.

Estas cuatro fuentes han sido exploradas como medios para asegurar la transparencia en la utilización de tecnologías de IA/AP. En contraposición a esta idea, en lo consecutivo afirmaré que estas fuentes no contribuyen realmente a superar la limitante que representa la opacidad epistémica en el uso de los sistemas.

Esto dejará en evidencia, además, que, partiendo de la búsqueda por defender el derecho al conocimiento y la información, las estrategias que buscan asegurar la transparencia, explicabilidad e inteligibilidad de las tecnologías no son compatibles con una práctica médica centrada en el paciente.

En lo consecutivo, exploraré cómo estas cuatro fuentes pueden y han sido abordadas en relación con los sistemas de Inteligencia Artificial; esto tendrá como objetivo indicar la insuficiencia que tienen para, de hecho, promover la transparencia de los sistemas.

3.1. Métodos de verificación y validación

El problema de la opacidad epistémica en sistemas complejos, como los de Aprendizaje Profundo, tiene un carácter necesario. Esto es que no importa el nivel de pericia de un agente cognitivo humano X, siempre habrá un grado de opacidad respecto a sus principios y mecanismos. Los agentes humanos están, por tanto, sometidos necesariamente a la indescifrabilidad de estos sistemas. Por ello, en los últimos años, los desarrolladores de sistemas de Aprendizaje Profundo hicieron notar la necesidad por crear vías para hacer descriptibles los procesos internos de estas tecnologías.

Ante el problema de la incapacidad por monitorear en su totalidad a los sistemas, la industria e investigación en materia computacional recurrió a la creación de una rama de la Inteligencia Artificial específicamente destinada a hacer explicables a los sistemas. De acuerdo con la Agencia de Proyectos de Investigación Avanzados de Defensa (DARPA por sus siglas en inglés) la Inteligencia Artificial Explicable (XAI), tiene por misión "entender, confiar de manera apropiada y administrar de manera efectiva una emergente generación de máquinas acompañantes artificialmente inteligentes".

El XAI pretende crear un conjunto de técnicas de aprendizaje de máquinas que contribuya a crear modelos más explicables manteniendo un alto de nivel de rendimiento. Así, se esperaría que se pudieran desarrollar técnicas que solucionen el

conflicto entre explicabilidad-contra-desempeño de los sistemas. En diversas áreas, como el diagnóstico asistido computarizado, existe una necesidad de que los sistemas sean transparentes, entendibles y explicables para ganar la confianza de los médicos, reguladores e, incluso, de los pacientes.

Idealmente, como afirma Singh [9], un sistema de diagnóstico médico debería ser capaz de explicar completamente y a todas las partes involucradas la lógica a través de la cual toma una decisión. Para cumplir con esta misión, diversas estrategias para la validación de los resultados de un sistema han sido propuestos. Esta validación requiere, por un lado, la verificación en la correspondencia entre el resultado del sistema y el estado de cosas del mundo y, por el otro, la verificación del correcto funcionamiento operativo del mismo de acuerdo con los principios que lo determinan.

Para el equipo de Singh, la explicabilidad es un elemento necesario para una utilización segura, ética, justa y confiable de la Inteligencia Artificial en el mundo real; por tanto, el XAI podría fungir como un método para desmitificar el carácter de "cajas negras" de estos sistemas. En el área médica se han implementado diversas estrategias para reducir el nivel de incertidumbre que provoca la operatividad de estos sistemas.

Tal es el caso de los métodos de interpretabilidad sensible, como los *saliency maps* y los métodos de detección de señal, que crean mapas de características detectadas por el sistema; en el caso médico estas herramientas permiten, por ejemplo, visualizar mapas de regiones salientes en imágenes radiológicas que han servido a las tecnologías para identificar tumores o signos de una enfermedad y que justifican su diagnóstico.

Asimismo, otros recursos como la difusión exitosa de significados capa a capa o los métodos contrafactuales permiten identificar la relevancia de un valor de entrada en la obtención de un resultado. Por el contrario, como afirmó David Ritscher en el taller público organizado por la FDA en 2020 en torno al rol de la IA en imagenología médica, la Inteligencia Artificial Explicativa tiene un problema nominal de gran relevancia: el XAI no explica nada. Lo que hacen estos métodos, por el contrario, es simplemente dar indicios acerca de lo que está ocurriendo dentro del sistema.

Naturalmente, estos métodos pueden generar una mayor transparencia en la operatividad del sistema, probar sus comportamientos o, incluso, encontrar fallas dentro de él. Sin embargo, los sistemas de XAI entran en un círculo de opacidad que Ritscher diagnosticaría como "tener una Inteligencia Artificial envuelta en otra Inteligencia Artificial huésped".

El problema con los métodos de XAI es que no validan los modelos de Aprendizaje Profundo, sino que permiten, acaso, visibilizar o interpretar algunas partes del sistema, haciéndolo mayormente entendible a los agentes cognitivos.Por lo anterior, podemos afirmar que los métodos de Inteligencia Artificial Explicable o bien aumentan la opacidad epistémica del sistema en su conjunto o bien permiten solamente disminuir el índice de incertidumbre sobre la operatividad de partes del sistema mismo.

3.2. Análisis de robustez

De mano con los métodos propuestos por el XAI, los análisis de robustez han sido asumidos como medios para verificar el correcto funcionamiento de los sistemas de Aprendizaje Profundo. Esta estrategia funge como respuesta al problema de la cantidad de información que un sistema debe procesar en relación con la compleja estructura operativa que le es propia. Podemos afirmar que el análisis en la robustez es aquello

que permite aprender sobre los resultados de un modelo dado, para saber si son artefactos del sistema o si están relacionados con características centrales del mismo. Esto es, que un análisis de robustez permite distinguir los errores en la construcción de la arquitectura del sistema a partir de una comprobación de que en situaciones imprevistas se comportará de manera correcta.

Respecto a la Inteligencia Artificial, podemos encontrar un proceso de confirmación de la robustez del sistema en el paso de validación y prueba del aprendizaje. Cuando un sistema es construido para el reconocimiento de voz o para la detección de objetos en una imagen, es necesario que pase por un proceso de entrenamiento que contemple el procesamiento de una amplia base de datos en relación con la tarea que debe realizar. Si, por ejemplo, un sistema tiene una arquitectura diseñada para la detección de tumores en la mama, resultará necesario que el sistema sea alimentado con una cantidad amplia de imágenes tomográficas de mamas sanas y de mamas con tumores.

Solo a través de un proceso de entrenamiento el sistema será capaz de extraer las características relevantes de este conjunto de imágenes para producir un mapa de características que sirva como parámetro para la evaluación de los nuevos valores de entrada. Tras ello y antes de enfrentarse con casos "reales", el sistema debe pasar por un proceso de validación en el que imágenes no antes procesadas serán utilizadas para medir el nivel de éxito del sistema y para perfeccionar sus parámetros. Solo después de este proceso el sistema es puesto a prueba con un pequeño número de imágenes que confirman el nivel de efectividad de este para clasificarlas.

En este punto se esperaría que el sistema tuviera la capacidad de asignar en las correctas categorías las tomografías de mamas sanas y de mamas con tumores. Así, la robustez partiría de una generalización inductiva sobre el funcionamiento del sistema. A mayor cantidad de casos en los que el sistema se comporta de manera correcta en situaciones imprevistas, tendría un carácter más robusto. A pesar de que este método de falsación es común en la evaluación de tecnologías de todo tipo, podría cuestionarse cómo se pueden establecer estos parámetros de éxito estadístico y si los resultados de estas evaluaciones proveen suficiente seguridad en el empleo de los sistemas. Hacia el final de este artículo trataremos algunas posibilidades que permiten aprovechar estas evaluaciones de robustez en relación con la gestión de riesgos.

3.3. Historia de las implementaciones exitosas/no exitosas

Sería un error suponer que un sistema computacional es un sistema acabado. En este sentido, es necesario reconocer que en el estado actual de cualquier sistema computacional hay una historia de sus exitosas y no exitosas ejecuciones, implementaciones y funcionamientos. En lo inmediato, podemos comprometernos con que el desarrollo tecnoevolutivo de los sistemas de IA/AP ha estado determinado, en gran medida, por la aspiración de superación de otras tecnologías en competencia.

Por ejemplo, en el caso de la detección de patologías por medio de sistemas de visión computacional, el perfeccionamiento de los sistemas parte de los resultados de otras arquitecturas y la utilización de estos conocimientos para el afinamiento de sus resultados. Trasladando el argumento de Durán y Formanek, podemos presumir que este proceso sería una justificación para confiar en el sistema y, dentro del marco de nuestra discusión, para asumir que sabemos cómo funciona la tecnología.

El problema fundamental con este recurso es que asume que los individuos conocen y comprenden la historia de las tecnologías. Si embargo, esta información generalmente no forma parte del saber colectivo. A pesar de que esto no es una fuente directa de la opacidad, sí implica una desventaja epistémica para ciertos actores. Esto nos conecta con la cuarta fuente de transparencia que suele asumirse como medio para asegurar el funcionamiento de las tecnologías y, en muchos casos, la fiabilidad en su uso: el conocimiento experto.

3.4. Conocimiento experto

De acuerdo con Durán y Formanek, el conocimiento experto es la cuarta fuente que provee justificaciones suficientes para confiar en un sistema computacional y en muchos casos ha sido explorado como medio para asegurar la transparencia epistémica. Es común pensar que la existencia de expertos en materia computacional, o en los principios que sirven de base a cualquier sistema opaco, demuestra con ciertos conocimientos es posible reducir la opacidad epistémica respecto a un sistema.

Un problema de esta taxonomía es que, en realidad, todas las fuentes propuestas por Durán y Formanek requieren de la existencia de expertos en el tema con un nivel de conocimientos en los principios del funcionamiento de las tecnologías. Así, para aplicar métodos de XAI resulta necesario contar con un alto nivel de conocimiento que permita crear, implementar e interpretar los resultados del análisis. Asimismo, para hacer un estudio de robustez se requiere estar familiarizado con la arquitectura en análisis y con los métodos de codificación y procesamiento que utiliza el sistema.

Por otro lado, para comprender la tecnoevolución de un sistema y la historia de su afinamiento se requiere tener un amplio conocimiento del desarrollo de los sistemas y de la historia de la tecnología. La problemática fundamental de la lectura de Durán y Formanek es que asumen que solo para los expertos en la materia no existe la opacidad epistémica respecto a estas tecnologías. Sin embargo, podríamos afirmar que el usuario final de un sistema inteligente de producción de fármacos o el paciente cuyo diagnóstico es llevado a cabo por una máquina de aprendizaje profundo no tiene un nivel de pericia y conocimientos suficiente sobre la tecnología para que pueda tomar una decisión informada respecto a su salud.

En general, los usuarios finales (médicos y pacientes) no tienen, por diversos motivos que trataremos más adelante, este conjunto de saberes. El uso de estas tecnologías con fines médicos implica que, potencialmente, un amplio número de usuarios (radiólogos, médicos, enfermeros, pacientes e, incluso, personas sin preparación) se vean involucrados en el uso directo con estos sistemas. Podríamos esperar que en el futuro estas tecnologías se implementen de manera masiva y que, por tanto, nos viéramos en la inevitable necesidad de utilizar estos sistemas para la procuración de nuestra salud.

Si siguiéramos la propuesta del dúo de filósofos tendríamos que apelar a que todos deberían tener los conocimientos suficientes sobre los opacos y crípticos principios y mecanismos de los sistemas o bien que debemos renunciar a esta transparencia y apelar a que los expertos evalúan y toman decisiones acertadas respecto al uso de las tecnologías para el manejo de la salud de las personas. Esta suposición, por demás utópica, resultaría, en lo inmediato, imposible de realizar.

4. La opacidad epistémica como producto del analfabetismo tecnológico

A pesar de que hoy en día vivimos rodeados de tecnologías digitales es importante reconocer que en general no estamos epistémicamente equipados para tomar decisiones bien fundamentadas respecto al funcionamiento de las herramientas computacionales y que incluso en menor medida estamos habilitados para pensarlas de manera crítica.

En este sentido, podemos afirmar que existe un entendimiento pobre de las características esenciales de la tecnología, pero también de la influencia que tienen en nuestras sociedades y de cuáles son las agencias que dependen de nosotros mismos para afectar su desarrollo. En su estudio sobre los índices de analfabetismo tecnológico en los Estados Unidos de Norteamérica, Young, Cole y Denton [10] han afirmado que el americano promedio consume productos sin conocer su composición y sin tener consciencia sobre cómo han sido desarrollados, producidos, empacados y distribuidos.

Por ello, a pesar de los altos niveles de producción y venta de tecnologías digitales en la región, de los altos índices de formación técnica y de la consecuente inclusión tecnológica que de estas economías se deriva, sería difícil afirmar que los niveles de analfabetismo en el país son equivalentes a los índices de consumo. En cambio, resulta necesario pensar que el desconocimiento del modo en que opera la industria tecnológica es una seria limitante para la comprensión de la digitalidad. En este sentido, los tecnólogos afirman que la posesión de habilidades y conocimientos técnicos específicos no garantizan la alfabetización tecnológica.

Esto se debe a que incluso los agentes con altos niveles de pericia en la materia pueden no tener el entrenamiento o la experiencia necesaria para pensar las implicaciones sociales, políticas y éticas de su trabajo. En este sentido, una mirada más amplia de la tecnología nos compromete con que el conocimiento de estos espectros debería ser tan valioso como el conocimiento técnico cuando hablamos de analfabetismo digital. Como afirma Kate Crawford, la IA/AP debe ser entendida como un atlas en el que se ven involucrados no solo los dispositivos técnicos mismos, sino múltiples sistemas de poder interconectados.

Así, la IA puede ser utilizada para hablar de las formaciones industriales masivas que incluyen política, trabajo, cultura y capital [11]. Es importante tomar en cuenta que estas condicionantes son las que, en gran medida, han determinado el inequitativo acceso a las tecnologías digitales y a la educación técnica. Cuestionar el papel de la política, la industria y el desarrollo tecnológico en los marcos económicos actuales resulta por tanto necesario para comprender las dimensiones que se ven afectadas por el uso de sistemas inteligentes.

La ininteligibilidad de las tecnologías nos compromete con que el paciente pierde autonomía pues no posee la información requerida para comprender el resultado de la evaluación y, por otro lado, no existen herramientas para hacérselo explicable [12]. Podríamos sugerir que este fenómeno es un tipo de paternalismo, caracterizado por la relación dispar entre el paciente y el profesional, que se traduce en que el médico tiene un entrenamiento y conocimiento superiores, lo cual lo sitúa en una posición de autoridad para determinar los intereses de aquellos que caen bajo su cuidado y administración.

Así, el usuario final-paciente que desconoce el funcionamiento del sistema se ve comprometido a asumir una posición de sometimiento respecto a la autoridad

epistémica del profesional de la salud, sea el caso de que este último tenga la información para llevar a cabo una toma de decisión médica informada o no. Carel y Kidd [13] han afirmado que en estos escenarios el paciente sufre, además, una vulneración debido a la injusticia epistémica que es producto del desconocimiento y la falta de elementos necesarios para entender las condiciones de interacción tecnológica y para comunicar sus propios intereses.

Así, aunque el médico no poseyera las herramientas para entender a la tecnología y asegurar sus resultados sí sería poseedor de una autoridad epistémica que lo sitúa en una posición de privilegio sobre el paciente. De este modo la vía de solución al conflicto del desconocimiento general de la operatividad de los sistemas radica no solo en el entrenamiento de los practicantes de la salud en materia digital y tecnológica, sino, principalmente, de todo agente que se encuentre atravesado por las prácticas médicas.

El problema radica, en paralelo a la limitante del analfabetismo técnico, en cómo hacer de dominio público la información necesaria para el desciframiento de las tecnologías a conjuntos no homogéneos de individuos (de diferentes contextos culturales y económicos, latitudes geográficas, comunidades tecnológicas y lingüísticas, etcétera). Un exhaustivo estudio desarrollado por Gómez-González [14] ha demostrado que la alfabetización tecnológica no figura actualmente para los desarrolladores de software de IA/AP como una de las áreas de oportunidad para la inclusión ética de las tecnologías en el área médica.

Por ello, la reducción de las brechas digitales y la inclusión tecnológica deben ser considerados como factores necesarios para asegurar la disminución de la opacidad epistémica de estas tecnologías y, por tanto, para la toma de decisiones informadas por parte de médicos y pacientes.

5. La opacidad epistémica como producto de la gobernanza digital

En lo inmediato, podemos afirmar que las tecnologías de IA/AP pueden impulsar en gran medida el aumento en el valor de la industria de la salud. Esto se debe a las contribuciones que las tecnologías tienen para aumentar la velocidad de ejecución de las tareas asignadas, así como la reducción de los costos y la complejidad de muchos procesos médicos y administrativos. En el área de la salud, estas contribuciones se hacen patentes, al menos, en el ámbito asistencial (diagnóstico, pronóstico, tratamiento, etcétera), en la salud pública (vigilancia epidemiológica y promoción de la salud), en la administración de las instituciones (para la optimización de recursos y gestión administrativa) y en la investigación biomédica (farmacología y ensayos clínicos, entre otros) [15].

A pesar de estas grandes promesas que ofrece la IA/AP un problema que limita su uso es que actualmente no existen normativas claras para su regulación. Por ello es necesario llevar a cabo un esfuerzo para establecer criterios de buena gobernanza tecnológica que no solo beneficien a las industrias, sino que proteja a los usuarios finales. En su plan de acción para el uso de software de IA/AP como dispositivo médico, la FDA ha afirmado que una de las grandes áreas de oportunidad que debe cubrir una agenda en materia tecnológica y de salud pública es promover objetivos centrados en el paciente que incorporen la búsqueda de transparencia a los usuarios.

De acuerdo con el diálogo público sostenido previo a la publicación del plan, diversas partes interesadas han expresado la necesidad de que los desarrolladores de las tecnologías describan de manera clara la información que ha sido utilizada para el entrenamiento de los algoritmos, la relevancia de sus valores de entrada (inputs), las lógicas que utilizan (cuando sea posible), el papel que se espera que los valores de salida (outputs) representen y la evidencia del desempeño de los dispositivos [16].

A partir de estas exigencias, la FDA ha expresado su interés por promover un tipo de transparencia por parte de la industria y de sus desarrolladores en aras de asegurar que los usuarios entiendan los beneficios, riesgos y limitaciones de los dispositivos, por ejemplo, mediante su etiquetado (*labeling*).

Paralelamente, un problema que representa el uso de estas tecnologías y la existencia de una amplia multiplicidad de herramientas inteligentes aplicadas en el área de la salud radica en que, en muchas ocasiones, los pacientes y usuarios finales no tienen conocimiento de si los dispositivos cuentan o no con arquitecturas de IA/AP o si han sido utilizadas para el diagnóstico de sus enfermedades, en su tratamiento, o en el seguimiento de sus biométricos.

Por ello, resulta necesario, en primer lugar, visibilizar el uso y disponibilidad de tecnologías de IA/AP en las instituciones de salud y fuera de ellas, regularizando la obligatoriedad en el aviso al usuario de que las herramientas que utiliza cuentan con tecnologías inteligentes. Ante este panorama, una pregunta que debe ocupar las agendas de los desarrolladores, reguladores y de la industria en general es ¿qué información debe estar disponible para los usuarios finales de las tecnologías y cómo debe ser presentada?

6. La gestión de riesgos como alternativa a la transparencia epistémica

A lo largo de esta discusión hemos afirmado que existe un alto grado de opacidad epistémica que implica la imposibilidad por asegurar la transparencia, explicabilidad e inteligibilidad de las tecnologías de IA/AP. Esta conclusión ha partido de una definición de la opacidad epistémica que consiste en afirmar la existencia de elementos epistémicamente relevantes que son desconocidos para los agentes cognitivos. Una crítica que se podría hacer a esta argumentación es que, en realidad, no existe un conjunto de elementos que sean epistémicamente relevantes de forma universal.

Mientras que a un técnico radiólogo puede parecerle necesario conocer cómo opera un sistema antes de utilizarlo para sus diagnósticos, para un paciente puede ser suficiente saber quién lo ha producido para asumir que la tecnología funciona debido a algún sesgo personal. De acuerdo con el grupo de investigación de Microsoft liderado por Vaughan y Wallach, una estrategia centrada en los intereses humanos que promueva la inteligibilidad de las tecnologías debe comenzar definiendo las necesidades de partes interesadas relevantes [17].

En este sentido, la búsqueda de inteligibilidad debe responder a las particularidades de los científicos de datos, desarrolladores, diseñadores, administradores de programas, reguladores, usuarios y de la gente que es afectada por los sistemas en general y no a partir de un supuesto bioético universalizante. Pero ¿cómo asegurar un acceso a la información mínimo para tomar decisiones informadas de acuerdo con las necesidades

de las partes interesadas? Las recomendaciones de la FDA tienen por objetivo advertir sobre la necesidad por pensar los futuros riesgos en el comportamiento de los dispositivos médicos.

Esta es una posible vía para asegurar la toma de decisiones informadas. En inicio, podemos reconocer que un elemento epistémicamente relevante a considerar al someterse a una evaluación, diagnóstico, tratamiento o seguimiento médico en el que se ve involucrado el uso de un sistema de IA/AP es cuál es el índice de éxito que ha demostrado tener el sistema, si el sistema es robusto, cuáles son los riesgos que implica su uso y qué acciones deben ser tomadas en caso de un fallo.

Proveer esta información a los usuarios finales de la tecnología a través de una buena gestión de riesgos puedes ser una respuesta satisfactoria que permita llevar a cabo una toma de decisión informada por parte de los pacientes y médicos a pesar de la existencia de las limitantes cognitivas y sociales que condicionan la existencia de la opacidad epistémica. Así, promover estándares para mantener informado al paciente y al usuario a través de la disponibilidad de manuales y reportes de seguridad funcionaría como un medio para que estos puedan decidir de forma autónoma respecto a su salud.

Una normativa de este tipo no se comprometería con que el paciente tenga que conocer las funcionalidades de los dispositivos ni tener acceso a la información que la alimenta (lo cual promueve, además, el derecho a la privacidad de la información y la protección de la propiedad intelectual). Además, la FDA sugiere que la presentación y acceso a la información sea prescrita por regulaciones locales, lo que permite adaptar la presentación de la información a las necesidades de los públicos que forman parte de cada conjunto sociocultural.

En conclusión, una consideración epistemológica para la creación de regulaciones en el uso de estos sistemas en el área de la salud radicaría en no apostar por la total transparencia operativa de las tecnologías, sino de vías que promuevan la inteligibilidad de los sistemas en casos y para agentes específicos. Como como afirma Marda [4] la transparencia no es necesariamente útil ni posible cuando se tratan sistemas de aprendizaje computacional. Esta búsqueda por la transparencia absoluta ha caído en el equívoco de una aspiración por hacer íntegramente comprensibles a los aspectos técnicos de los sistemas a todo agente humano.

De acuerdo con el planteamiento presentado en este artículo, resultaría necesario tener una mayor precisión en el discurso legislativo y una claridad conceptual que responda a las necesidades efectivas de las comunidades tecnológicas en consideración ya que la aspiración por volver transparentes a los sistemas no es, en muchos casos, sino la búsqueda de hacerlos meramente inteligibles a ciertos públicos.

7. Conclusiones y trabajo a futuro

A lo largo de este trabajo he afirmado que la opacidad epistémica es un límite para el uso seguro y responsable de la IA/AP en el área de salud. Asimismo, he argumentado que los métodos convencionales para reducir los niveles de opacidad en el desarrollo, implementación y uso de estas tecnologías para todos los agentes relevantes que se ven relacionados con ellas son insuficientes.

Esta perspectiva me ha llevado a señalar un gran peligro: alcanzar la absoluta transparencia de los sistemas se vuelve un objetivo imposible. En realidad, podríamos

afirmar que, aunque las consideraciones bioéticas promuevan la reducción en la opacidad de los sistemas, los agentes que interactuamos con ellos siempre nos encontraríamos con un límite para saber cómo funcionan y cómo utilizarlos; en este sentido, no habría transparencia porque seguiríamos estando imposibilitados para ver dentro de los sistemas.

Si las agendas en materia de regulación de la IA/AP continúan comprometiéndose con la transparencia, explicabilidad e inteligibilidad en el uso de estas tecnologías deben también comprometerse con una inclusividad tecnológica que socave los niveles de analfabetismo tecnológico y, como afirma la División de Desarrollo Social de la Comisión Económica para América Latina y el Caribe (CEPAL), alcanzar umbrales de competencia digital para la inclusión social [18].

A pesar de los beneficios sociales que puede tener este enfoque, las condiciones materiales de las sociedades contemporáneas pueden encontrar severas limitantes para cumplir con esta misión. Como alternativa a este planteamiento, podría sugerirse que la toma de decisión informada en materia de IA/AP a través de un conocimiento básico de los riesgos que implica el uso de la tecnología funciona como deriva de solución al problema de la "medicina de caja negra", a pesar de que siempre haya un límite en la transparencia de los sistemas.

En resumen, el anhelo de transparencia debe ser abandonado para considerar derivas más realistas en el tratamiento de la IA/AP en el área de la salud. La IA debe dejar de ser entendida como una caja negra que es necesario abrir: no hay un secreto que exponer sobre su operatividad ni su naturaleza. Un trabajo futuro consistirá en definir cómo se debe presentar esta información a los usuarios finales para hacerla comprensible. Esto requerirá poner especial atención en los niveles de alfabetización poblacional, en las fórmulas de interacción paciente-médico y en las estructuras legislativas del contexto en estudio.

Naturalmente esta tarea se vuelve de especial importancia en contextos tecnológicamente menos desarrollados y que no cuentan con iniciativas claras de regulación tecnológica, como es el caso de México y otros países latinoamericanos. Solo en tanto reconozcamos la complejidad de estas tecnologías es que podremos comenzar a entender la influencia que tienen en la vida humana y obtener una mejor comprensión de su papel en el mundo.

Referencias

1. Organización Mundial de la Salud: Ethics and governance of artificial intelligence for health: WHO guidance (2021)
2. Bjerring, J., Busch, J.: Artificial Intelligence and patient-centered decision-making. Philosophy and Technology, vol. 34, pp. 349–371 (2020)
3. Article 19: Governance with teeth: How human rights can strengthen FAT and ethics initiatives on artificial intelligence. Article 19 (2019)
4. Marda, V.: Machine Learning and Transparency: A Scoping Exercise. SRRN (2017)
5. Sztompka, P.: Trust: A sociological theory. Cambridge University Press (1999)
6. Humphreys, P.: The philosophical novelty of computer simulation methods. Synthese, vol. 169, no. 3, pp. 615–626 (2009) doi: 10.1007/s11229-008-9435-2.
7. Burrell, J.: How the machine 'thinks': Understanding opacity in machine learning algorithms. Big Data & Society, vol. 3 (2016)

8. Durán, J. M., Formanek, N.: Grounds for Trust: Essential Epistemic Opacity and Computational Reliabilism. Minds and Machines, vol. 28, pp. 645–666 (2018) doi: 10.1007/s11023-018-9481-6.
9. Singh, A., Sengupta, S., Lakshminarayanan, V.: Explainable Deep Learning Models in Medical Image Analysis. Journal of Imaging, vol. 6, pp. 52 (2020) doi: 10.3390/jimaging6060052.
10. Young, T, Cole, J., Denton, D.: Improving Technological Literacy. Issues in Science and Technology, vol. 18, no. 4, 73–79 (2002)
11. Crawford, K.: Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence. Yale University Press (2021)
12. Dragomir, A.: Luke, I´m NOT Your Father: Beyond Technological Paternalism, towards Mutual Cooperation between Patients, Medical Staff and AI. In: CEPE/IACAP Joint Conference 2021: The Philosophy and Ethics of Artificial Intelligence (2021)
13. Carel, H., Kidd, I.: Epistemic injustice in healthcare: a philosophical analysis. Medicine. Health Care and Philosophy, vol. 17, pp. 529–540 (2014) doi: 10.1007/s11019-014-9560-2.
14. Gómez-González, E., Gomez, E., Márquez-Rivas, J., Guerrero-Claro, M., Fernández-Lizaranzu, I., Relimpio-López, Ma. I., Dorado, M., Mayorga-Buiza, Ma. J., Izquierdo-Ayuso, G., Capitán-Morales, L.: Artificial intelligence in medicine and healthcare: a review and classification of current and near-future applications and their ethical and social Impact. (2020) doi: 10.48550/arXiv.2001.09778.
15. Miralles, F.: Sector salud y bienestar: Pruebas de concepto de referencia. AI & Big Data Congress (2021)
16. FDA: Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan (2021)
17. Vaughan, J., Wallach, H.: A Human-Centered Agenda for Intelligible Machine Learning. Microsoft (2022)
18. Martínez, R., Trucco, D., Palma, A.: El analfabetismo funcional en América Latina y el Caribe: Panorama y principales desafíos de política. CEPAL (2014)

ISSN 1870-4069

Atribución de autoría en textos en español a partir de sus atributos textuales

Fernando Hernández-Ibarra[1], Belém Priego-Sánchez[1], David Pinto[2]

[1] Universidad Autónoma Metropolitana,
Departamento de Sistemas,
México

[2] Benemérita Universidad Autónoma de Puebla,
Facultad de Ciencias de la Computación,
México

{herr.taquitos, belemps}@gmail.com, dpinto@cs.buap.mx

Resumen. La atribución de autoría busca identificar al autor de entre un grupo de posibles autores de un texto. Las aplicaciones de esta tarea abarcan por ejemplo la detección de plagio, identificación del autor de un texto anónimo y análisis forense. En este trabajo se pretende resolver esta tarea para dos corpora, uno compuesto por textos literarios de 9 autores hispanohablantes, y el segundo es un corpus obtenido de la competencia PAN 2017 que se compone de tweets en español, con base en los atributos textuales. Para el primer corpus, se alcanza una precisión sobre un conjunto de pruebas de 89.59% mediante redes neuronales basadas en una arquitectura tipo transformer. Para el segundo corpus, el mejor resultado obtenido fue con una red neuronal convolucional con una precisión de 73.10%. Los resultados obtenidos, en ambos corpora, son prometedores para la tarea de la atribución de autoría.

Palabras clave: Atribución de autoría, aprendizaje profundo, textos en español.

Authorship Attribution in Spanish Text through Text Attributes

Abstract. Authorship attribution seeks to identify the author of a text between a set of possible authors. It can be applied to plagiarism detection, author identification in anonymous text or forensic analysis, and others. This article aims to solve this task through text attributes on two corpora: one composed of literary texts written by 9 Spanish-speaking authors and the second corpus is obtained from the competition PAN 2017, which is composed of tweets in Spanish. The accuracy achieved on the first corpus was 89.59% on a test dataset using neural

networks based on a transformer type architecture. For the second corpus, the best result achieved was an accuracy of 73.10% using a convolutional neural network. The results obtained, on both corpora, are promising for the authorship attribution task.

Keywords: Authorship attribution, deep learning, Spanish text.

1. Introducción

Cada escrito tiene características que los hacen o no similares entre sí más allá del tema que traten; por ejemplo, la riqueza de vocabulario, longitud de oraciones, puntuación, entre otras características que en conjunto forman patrones de redacción propios de cada individuo. Estás marcas personales, que pueden ser una decisión consciente o no, pueden ser cuantificadas y, por tanto, se pueden convertir en una huella lingüística de su autor.

La estilometría es un campo que se ocupa del estudio de estos patrones lingüísticos y es la base para la tarea de identificación de autor, o atribución de autoría (AA), que consiste en, dado un texto, determinar al autor del texto de entre un grupo de posibles autores y el perfilado de autor, que busca descubrir características asociadas al autor tales como el género, la edad o el idioma (incluyendo variantes de éste).

Estas tareas tienen varias aplicaciones: análisis forense, detección de plagio, seguridad e incluso marketing. En el área del Procesamiento del Lenguaje Natural (PLN) la AA se ha vuelto un tema habitual y muchos trabajos se acercan justamente mediante la estilometría, aunque las características más útiles para llevar a cabo la tarea no son fijas para todo estudio, sino que cada conjunto de datos es mejor representado por unas u otras características.

El aprendizaje automático clásico ha sido ampliamente usado, sin embargo, en años recientes y aprovechando la enorme cantidad de datos disponibles en línea, modelos neuronales han ido ganando atención para diversas tareas relacionadas al PLN, en particular, aunque no limitado a, redes neuronales convolucionales (RNC), que son capaces de trabajar con datos donde existan patrones espaciales, redes neuronales recurrentes (RNR) que están diseñadas para trabajar con datos secuenciales o *transformers*, que emplean un mecanismo que permite superar los límites de las entonces dominantes RNR.

Además, es importante recalcar que el aprendizaje profundo no reemplaza al aprendizaje automático clásico, existen trabajos sobre AA donde emplean tanto algoritmos de DL como de ML, obteniendo resultados similares para ambos enfoques o donde uno sobresale al otro de acuerdo a ciertas condiciones de los modelos, como en [1], donde utilizan algoritmos clásicos de aprendizaje automático y una RNR, concluyendo que, bajo la configuración hecha de la RNR, ésta no destaca de forma sustancial sobre sus contrapartes, o en [2] donde comparan un modelo de red neuronal recursiva contra un modelo multinomial Naive Bayes, siendo el modelo de red neuronal mejor cuando se trata de resolver atribución de autoría para 3 autores, pero el modelo Naive Bayes obtiene mejores resultados cuando se escala el problema a 10 autores.

En este trabajo se busca resolver la AA para un corpus compuesto por textos literarios de 9 autores hispanohablantes y para un segundo corpus compuesto de tweets en español, el cual fue tomado de la competencia PAN del 2017 [3], por medio de las características textuales y empleando algoritmos de aprendizaje profundo.

En la sección 2 se presentan algunos trabajos relacionados a la AA resueltos por medio de diferentes enfoques, tanto del análisis de los datos como de los algoritmos empleados. En la sección 3 se describen con más detalle los conjuntos de datos utilizados en el presente trabajo. La sección 4 detalla los pasos que componen la metodología propuesta para la resolución de AA para los dos corpora utilizados.

En la sección 5 se presentan los mejores resultados obtenidos para cada corpus empleando una métrica de evaluación extra a la precisión cruda y se discuten los mismos. Finalmente, la sección expone las conclusiones y posibles mejoras al presente trabajo.

2. Trabajos relacionados

En [1] trabajan la AA utilizando tres enfoques: el primero realiza un análisis a nivel de artículo y utilizan algoritmos clásicos de aprendizaje automático, el segundo se acerca por medio de un análisis a nivel de palabra entrenando un modelo GloVe, y en el tercer enfoque desarrollan una RNR con vectores de palabras pre-entrenados con GloVe y el análisis es a nivel de oración.

Al final del trabajo comparan los resultados de los diferentes algoritmos utilizados, donde se observa que la RNR no destaca tanto del resto de algoritmos, aunque queda abierto a una optimización de la red con la cuál pueda ser que sí se logre un resultado mucho más destacable sobre los enfoques de aprendizaje automático clásico. En [2] buscan un modelo de aprendizaje profundo para resolver el problema de atribución de autoría múltiple, es decir, cuando son varios los autores de un texto, a nivel de oraciones.

Utilizan un corpus compuesto por artículos de Wikipedia que fueron escritos por varios autores. Los resultados demuestran que el modelo de red neuronal recursiva es superado por un modelo Multinomial Naive Bayes conforme se incrementa el número de autores a distinguir. En [4] se aborda la tarea de atribución de autoría haciendo uso de un modelo de ensamble de redes neuronales convolucionales y recurrentes LSTM.

Utiliza un corpus con textos en inglés obtenidos del proyecto Gutenberg, además de los corpus usados en la competencia PAN del 2013 [5] y 2014 [6] para comprobar la generalidad del modelo. Se trabajó dividiendo los textos en n-gramas a nivel de palabra y caracteres, sin aplicar ninguna técnica de tokenización. En [7] se propone un modelo para resolver la tarea de atribución de autoría no supervisada.

Utiliza el mismo corpus que fue dado en el PAN 2017 para la tarea de agrupación de documentos por autor, que incluye dos géneros (artículos y reseñas) y tres idiomas (inglés, holandés y griego). Con base en el estado del arte para esta tarea, propone modificaciones (como el tratamiento que reciben los tokens, el uso de caracteres especiales como puntuación y selección de características), que llevan a una ligera mejora en los resultados mostrados por el estado del arte.

En [8] proponen una solución al problema de atribución de autoría de un solo autor mediante modelos de cálculo de semejanza entre textos, sin emplear modelos que requieran ser entrenados o calibrados. En [9] trabajan con un corpus compuesto por tweets, los cuáles analizan a nivel sintáctico y forman gramas a partir de las dependencias sintácticas en cada dato del corpus.

En [10] utilizan varios tipos de modelos de aprendizaje profundo, siendo un modelo GRU a nivel de artículo el que logra mejores resultados para la tarea de AA. Además, trabajan también la verificación de autoría, que consiste en determinar si dos entradas pertenecen a la misma categoría, por medio de una red siamés (una red siamesa consiste en dos subredes con la misma arquitectura, parámetros y pesos).

3. Conjunto de datos

La presente sección presenta la descripción de los conjuntos de datos utilizados. En la sección 3.1 se detalla el contenido y creación del corpus de autores literarios y la sección 3.2 describe el corpus de tweets en español.

3.1. Corpus TLE

El corpus de Textos Literarios en Español, denominado corpus TLE, ha sido construido para la realización de este trabajo debido a que, al menos hasta el momento de comenzar el presente proyecto, no se encontró un corpus adecuado para esta tarea. Este corpus TLE se formó recolectando y transcribiendo diversas obras de nueve autores literarios hispanohablantes de finales del siglo XX y principios del XXI, entre las que se incluyen novelas, cuentos y ensayos.

El corpus puede ser proporcionado a la comunidad científica solicitándolo directamente a los autores del artículo. Aproximadamente, el corpus se compone de 1,300,000 palabras, sin contar signos de puntuación. Sin embargo, el análisis que se hace sobre estos datos se enfoca en el manejo de párrafos como subestructura de los textos, entonces se dividen las obras en párrafos (cada diálogo se consideran un párrafo) y se les asigna la correspondiente etiqueta de acuerdo con el autor de cada párrafo.

El corpus TLE se divide en tokens como parte del preprocesamiento y es de interés observar la distribución de la longitud de los párrafos para poder definir un tamaño máximo del párrafo, medido en la cantidad de tokens, ya que es necesaria la uniformidad de esta característica para los modelos de aprendizaje profundo utilizados. Para esto se realiza un análisis sencillo de la frecuencia de longitud Fig. 1, tras lo cual se realizan pruebas con un modelo preliminar utilizando diferentes rangos de longitud de párrafo, con el fin de obtener una idea de la utilidad aparente de los rangos de párrafos.

Tras esto, se decide trabajar con un subconjunto del corpus que se componga de aquellos párrafos con longitud entre 50 y 500 tokens, el resto son despreciados por considerarse o carentes de información (en el caso de aquellos con menos de 50 tokens) o puntos atípicos (aquellos con longitud mayor a 500). En la Tabla 1 se describe el número de párrafos por cada autor en esta nueva versión del corpus TLE.

Fig. 1. Frecuencia del tamaño de tokens en cada párrafo del corpus TLE.

Tabla 1. Cantidad de párrafos por autor en el corpus TLE.

Autor	Número de párrafos	Autor	Número de párrafos
Autor 1	862	Autor 6	1240
Autor 2	1091	Autor 7	621
Autor 3	711	Autor 8	1133
Autor 4	765	Autor 9	740
Autor 5	712		
Total de párrafos: 7875			

Como se observa en la Tabla 1, el número de párrafos por autor no es equitativo para cada uno, por lo que basarse meramente en la precisión general del eventual modelo podría llevar a mal interpretación. Para mitigar esta problemática, se pueden utilizar otras métricas de evaluación del modelo, como f1-score que toma en cuenta tanto la cantidad de errores como el tipo de estos que comete el modelo en sus predicciones.

3.2. PAN

PAN es una serie anual de eventos científicos sobre textos forenses y estilometría. En este trabajo se utiliza únicamente la porción del corpus encontrado en [3] correspondiente al idioma español con las etiquetas de género del autor. Se realiza un análisis sobre el corpus con el mismo fin de fijar una longitud máxima de tokens. Así, se observa que casi todos los tweets tienen una longitud de 50 o menos tokens, y hay menos de 100 tweets con longitud mayor a 50, por lo cual, se decide excluirlos.

En este caso, no se establece de antemano una longitud mínima. Como se mencionó, al trabajar con este conjunto de datos se utilizan las etiquetas de género, por lo que el problema de AA en este caso se reduce a clasificación binaria (0 para femenino y 1 para masculino), además, el corpus ya está balanceado, es decir, la distribución de las etiquetas (hombre y mujer) es aproximadamente 50:50, por lo que de inicio, sabremos

que un modelo con precisión mayor a 50% es en sí mejor que predecir por mera probabilidad.

4. Metodología propuesta

Para cada corpus descrito en la sección anterior, se realizan experimentos con la metodología mostrada en la Fig. 2 y descrita a continuación.

1. Obtención del corpus. Esta etapa consiste en la recolección de los datos para formar los corpus TLE y PAN, descritos en la sección 3.1 y 3.2 respectivamente, y la división en 3 subconjuntos correspondientes al conjunto de entrenamiento (80%), validación (10%) y prueba (10%). Los conjuntos de entrenamiento y validación se utilizan durante la etapa de entrenamiento de los modelos de aprendizaje profundo, mientras que el conjunto restante se utiliza para probar el modelo en una situación que se considera un caso de uso real y es este resultado el que se considera para evaluar la utilidad del modelo. Además, se realiza un breve análisis de los datos recolectados a fin de filtrar aquellos de los que se pueda prescindir.
2. Tokenización. Consiste en segmentar el corpus en tokens (en este proyecto se utiliza la librería spaCy) con el fin de facilitar el trabajo en la etapa siguiente.
3. Análisis de los datos. El objetivo de esta fase es obtener conocimiento sobre la composición de los datos en los corpora, por ejemplo la riqueza del vocabulario, categoría gramatical de los tokens, etc. La profundidad del análisis se limita a la información gramatical que asocia spaCy a los tokens.
4. Preprocesamiento. Con base al análisis en la etapa previa, se ejecuta un conjunto de técnicas comunes de preprocesamiento (lematización, conversión a mayúsculas/minúsculas, remoción de stop-words, etc) sobre los conjuntos de entrenamiento y validación, se extrae el vocabulario asociado a estos por medio de un diccionario y se mapea cada token a su valor índice en el diccionario (las etiquetas de cada corpus igualmente se convierten a alguna representación numérica). Finalmente se trunca o rellena el párrafo para ajustar su longitud a la máxima longitud fijada en la primera etapa.
5. Construcción del modelo. En esta etapa se deciden los hiperparámetros de los modelos empleados (RNC, RNR y transformer), posteriormente se construye el modelo y finalmente se le entrena utilizando los conjuntos de entrenamiento y validación previamente preprocesados. Para ambos corpora, se emplea una capa Embedding entrenable propia del problema, es decir, no se utiliza un modelo de embedding pre-entrenado. Esto con la finalidad de que los valores del vector estén además asociados a la resolución de la tarea.
6. Evaluación del modelo. El modelo con mejores resultados sobre el conjunto de validación del paso anterior se testea con el conjunto de pruebas para tener una idea del desempeño del modelo bajo un caso de uso real, utilizando una métrica que mejor se adapte a cada corpus utilizado. Los resultados obtenidos se comparan con modelos previos para tratar de obtener algún indicio de

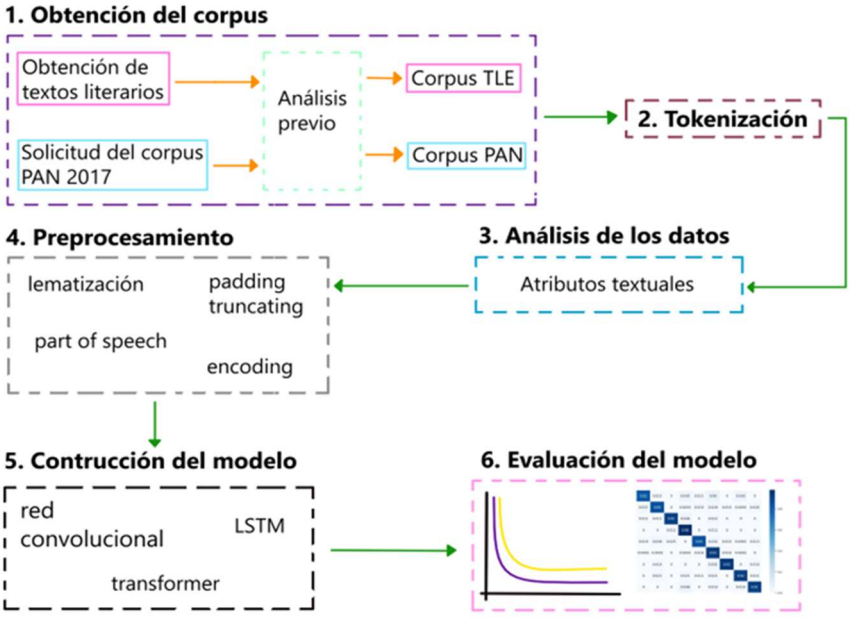


Fig. 2. Metodología propuesta para la atribución de autoría en textos en español.

cuáles características textuales son de mayor o menor utilidad para resolver la tarea de AA para el corpus correspondiente.

Esta metodología permite comparar diversos preprocesamientos y modelos; además, requiere una cantidad de recursos y tiempo grande debido a que básicamente se realiza un nuevo experimento por cada variación deseada. En este trabajo, esto no representa un inconveniente pues los tiempos de preprocesamiento y entrenamiento con los corpora utilizados son manejables, con recursos gratuitos -como Google Colab- son suficientes. Sin embargo, para proyectos de mayor tamaño este acercamiento tipo *prueba y error* no es viable.

5. Análisis y resultados experimentales

Esta sección se dedica a la exposición y discusión de los mejores resultados obtenidos para cada conjunto de datos. En las secciones 5.1 y 5.2 se presentan los resultados para los corpus TLE y PAN respectivamente, en la sección 5.3 se analizan los resultados en las dos secciones previas.

5.1. Corpus TLE

Se utiliza un modelo (TransformerTLE) ligeramente modificado del transformer obtenido de los ejemplos encontrados en la página de la API Keras [11]. La arquitectura

Tabla 2. Arquitectura del modelo transformer para el corpus TLE.

TransformerTLE		
Capa	*Output Shape*	# parámetros
Input	[(None, 500)]	0
TokenAndPositionEmbedding	(None, 500, 100)	5,993,700
TransformerBlock	(None, 500, 100)	84,416
GlobalAveragePooling1D	(None, 100)	0
Dropout	(None, 100)	0
Dense	(None, 64)	6464
Dropout	(None, 64)	0
Dense	(None, 9)	585
Parámetros entrenables: 6,085,165		

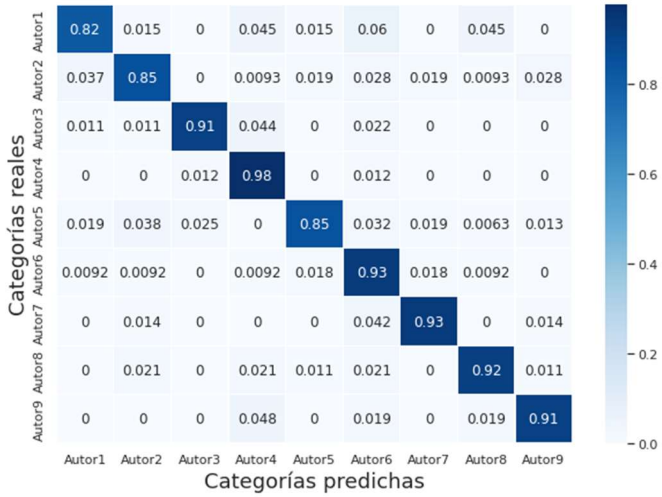


Fig. 3. Matriz de confusión del modelo transformer.

del modelo resultante se muestra en la Tabla 2. La capa *TokenAndPositionEmbedding* define internamente dos operaciones *embedding* (una para las palabras/tokens y otro para las posiciones de estas) resultando en vectores de 100 dimensiones.

La capa TransformerBlock define una capa MultiHeadAttention con 2 cabezas de atención y una capa Dense de 16 neuronas, con relu como función de activación, además de capas LayerNormalization y Dropout para optimizar las operaciones del modelo. A continuación, se toma promedio de las salidas del Transformer (GlobalAveragePooling1D) y se pasa por una red feedforward, que incluye una capa

Tabla 3. Arquitectura del modelo RNC para el corpus PAN.

RNC-PAN		
Capa	*Output Shape*	# parámetros
Embedding	(None, 50, 200)	82,128,800
SpatialDropout1D	(None, 50, 200)	0
Conv1D	(None, 48, 256)	153,856
GlobalMaxPooling1D	(None, 256)	0
Dense	(None, 1)	257
Parámetros entrenables: 82,282,913		

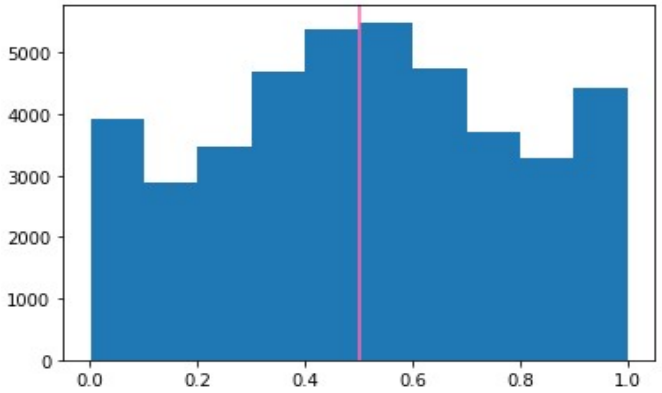


Fig. 4. Distribución de las predicciones del modelo RNC-PAN.

oculta Dense con 64 neuronas y relu como función de activación, para su clasificación (el output es producido por una capa Dense de 9 neuronas con activación softmax).

Las capas Dropout en todo el modelo utilizan un rate de 0.1. El modelo se entrena utilizando categorical-crossentropy como función de pérdida y el algoritmo adam como optimizador, además se emplea la técnica earlystop para detener el entrenamiento en cuanto no haya mejora (delta igual a 0.001) en el valor de la función de pérdida sobre el conjunto de validación y evitar así un posible caso de sobre-entrenamiento (overfitting).

Las pruebas con el conjunto de pruebas arrojaron un resultado de 89.59% de precisión en las predicciones, la matriz de confusión asociada se muestra en la Fig. 3. Con el propósito de tener en consideración el desbalance de la cantidad de datos por autor (Tabla 1), se emplea la métrica f1-score, la cual da un resultado de 0.896.

5.2. Corpus PAN

Se construye un modelo sencillo de red convolucional (RNC-PAN) que se muestra en la Tabla 3. En este modelo la capa *Embedding* produce vectores de tamaño 200, a

continuación se utiliza una capa *SpatialDropout1D* con *rate* 0.1 y una capa *Conv1D* con 256 *kernels* con tamaño de ventana igual a 3 (*stride* por default igual a 1) y *relu* como función de activación. Finalmente se reduce la dimensión de la salida de esta capa por medio de *GlobalMaxPooling1D* y se clasifica utilizando una neurona binaria con función de activación *sigmoid*.

El modelo se entrena con la función de pérdida *binary crossentropy*, optimizador *adam* y utilizando nuevamente *earlystop*. En el conjunto de pruebas, el modelo de la Tabla 3 logró una precisión del 73.10%. En la Fig. 4 se muestra la distribución de las predicciones del modelo (se muestra la probabilidad de pertenencia a cada clase, 0 o 1). Como métrica adicional a la precisión se mide el área bajo la curva ROC (o AUC-ROC por sus siglas en inglés), el cual resulta en 0.8187.

5.3. Análisis

Con base en los resultados mostrados en la Fig. 3, el modelo TransformerTLE logra un buen desempeño para clasificar textos de los nueve autores tratados, aunque para algunos tiene mayor conflicto para identificarlos. Esto podría deberse a mayor presencia de vocabulario no antes visto por la red en las porciones de esos autores en el conjunto de pruebas, a diferencia de sus contrapartes con mucho mayores resultados de precisión, o quizá el enfoque aplicado durante el análisis no es tan conveniente para ciertos autores como lo es para el resto.

Respecto al modelo RNC-PAN, sin tener del todo un terrible desempeño, el modelo no es capaz de discernir claramente, entre las dos categorías posibles (el mejor puntaje alcanzado durante la competencia PAN del 2017 para la misma porción del corpus fue de 83.21% de precisión). Aun así, el puntaje bajo la curva ROC alcanzado por el modelo es de 0.8187, indicando que el desempeño es medianamente bueno. Esto puede deberse, entre otros motivos, a que el análisis realizado no es el indicado para el tipo de contenido del corpus o a que el modelo propuesto, a pesar de ser el mejor de los que se probaron, está lejos de ser una arquitectura adecuada para el problema.

6. Conclusiones y perspectivas

El modelo TransformerTLE alcanza resultados satisfactorios para resolver la tarea de atribución de autoría para los 9 autores tratados en este trabajo. Sin embargo, el margen de mejora es aún considerable considerando que, como se muestra en la Fig. 3, hay autores para los que el modelo tiene más conflicto de identificar. Más aún, un análisis lingüístico más especializado sobre el corpus TLE podría llevar a mejores resultados. Sumado a esto, como se mencionó anteriormente, se utilizó un modelo basado en [9], por lo que una arquitectura más especializada o mejor diseñada para el corpus TLE resulte más conveniente.

El modelo RNC-PAN por otro lado, aunque no se acerca a los mejores resultados en la competencia, no obtuvo tampoco un pésimo desempeño y cumple medianamente con el objetivo de atribución de autoría (perfilado de autor específicamente). Es claro que puede mejorarse, al menos hasta el punto del estado del arte en la competencia del 2017,

quizá por medio de un análisis más exhaustivo de los datos o un enfoque distinto de análisis (por ejemplo, debido a que se lematizaron las palabras en este trabajo se perdieron los detalles de conjugación). De igual forma, queda abierto diseñar una arquitectura o modelo diferente.

Como trabajo futuro se puede resolver el problema con una perspectiva lingüística más profunda en el análisis de los datos a fin de obtener características más concretas que puedan llevar a mejores resultados, lo que a su vez podría llevar a reducir la complejidad de los modelos (es decir, la cantidad de parámetros entrenables).

Además, con una selección diferente de las características lingüísticas puede darse el caso que una arquitectura o un modelo diferente a las empleados en este trabajo sea más eficaz, lo que deja abierto el trabajo a desarrollar o adaptar un nuevo modelo de clasificación; aunque no necesariamente un modelo de aprendizaje profundo, ya que si bien han demostrado que pueden sobrepasar el desempeño de un modelo de aprendizaje automático clásico, no es una verdad aplicable a todo problema, tal como se observa en [1] y [2], por tanto, puede incluso darse el caso que un modelo clásico obtenga un mejor desempeño para la tarea tratada en este trabajo.

Referencias

1. Wang, H., Zhou L.: News Authorship Identification with Deep Learning (2016)
2. Macke, S., Hirshman, J.: Deep Sentence-Level Authorship Attribution (2015)
3. Rangel, F., Rosso, P., Potthast, M., Stein, B.: PAN17 Author Profiling [Data set]. In: CLEF 2017 Labs and Workshops, Notebook Papers. Conference title: PAN at Conference and Labs of the Evaluation Forum (2017).
4. López-Velasco, F.: Verificación de autoría en textos mediante redes neuronales convolucionales y recurrentes. Tesis de maestría, Universidad Nacional Autónoma de México, México (2018)
5. Rangel, F., Rosso, P., Koppel, M., Stamatatos, E., Inches, G.: PAN13 Author Profiling [Data set]. In: CLEF 2013 Labs and Workshops, Notebook Papers. Conference title: PAN at Conference and Labs of the Evaluation Forum (2013)
6. Rangel, F., Rosso, P., Chugur, I., Potthast, M., Trenkmann, M., Stein, B., Verhoeven, B., Daelemans, W.: PAN14 Author Profiling [Data set]. In: CLEF 2014 Labs and Workshops, Notebook Papers. Conference title: PAN at Conference and Labs of the Evaluation Forum (2014)
7. Martín-del-Campo-Rodríguez, C.: Atribución de autoría con aprendizaje automático. Tesis de maestría, Instituto Politécnico Nacional, México (2019)
8. Castro, D., Adame, Y., Pelaez, M., Muñoz, R.: Verificación de autoría, clasificación por vecindad. Computación y Sistemas, vol 21, no. 2, pp. 181–201 (2017)
9. Castillo-Velásquez, F. A, Martínez-Godoy, J. L., Torres-Falcón, M. P., Zavala de Paz, J. P., Becerra-Chávez, A., Rizzo-Sierra, J. A.: Atribución de autoría de mensajes de Twitter a través del análisis sintáctico automático. Research in Computing Science, vol. 149, no. 11, pp. 91–101 (2020)
10. Qian, C., He, T., Zhang, R.: Deep Learning based Authorship Identification (2017)

11. Text classification with Transformer—Keras, https://keras.io/examples/nlp/text_classification_with_transformer/ (2022)

ISSN 1870-4069

Discrete-Time Modeling and Control for a Soft Robot Displacements based on Experimental Data

Arturo Baltazar[1], Isaias Campos[1], Josué Gómez[2]

[1] Universidad Autónoma de Coahuila,
Facultad de Ingeniería,
Mexico

[2] Instituto Politécnico Nacional,
CINVESTAV, Saltillo,
Mexico

{arturo.baltazar,isaias.campos}@cinvestav.edu.mx,
jogomezc@uadec.edu.mx,

Abstract. Soft robot applications have recently gained importance over rigid robots for their great maneuverability to work cooperatively with human beings and in unstructured environments. A modified technique for soft-robot actuation is based on rapid liquid evaporation using ultrasonic waves and heat reaction. In any case, the soft robot displacements with rapid actuation contain high nonlinearities and uncertainties. Therefore, the classical control techniques based on analytical models become impractical to apply into soft robots. A nonlinear discrete-time model of a soft-robot displacement is proposed from experimental data in this research. In addition, a novel control law is developed applying a neuro-fuzzy network with adaptive stage and a sliding mode surface function as an input to compensate for uncertainties.

Keywords: Soft robot, rapid actuation, discrete-time regression model, sliding mode function, neuro-fuzzy control law.

1 Introduction

During recent years, the soft robot applications have increased notably for their human beings interaction, fragile objects handling and unstructured environments exploration, [1,8]. In addition, the soft robots have a great maneuverability to imitate biological systems, [7,9]. The soft robots are generally performed through pneumatic actuators. The injected fluid into an elastomer chamber covers the robot's volume causing displacements by the pressure on the soft walls.

On the other hand, some conventional soft robot designs are also integrated by heat exchangers to achieve a liquid vaporization at short term. [3] reported the impacts on the soft-robot performance when the phase change rate (liquid-gas) actuation is applied inside of a elastomer chamber. Currently, a novel actuation for soft robots is using a liquid dispersion by ultrasound waves, as is presented by [5]. Rapid actuation evaporates the liquid below its boiling point in a suitable time and without any structural material damage, as is depicted in Figure 1.

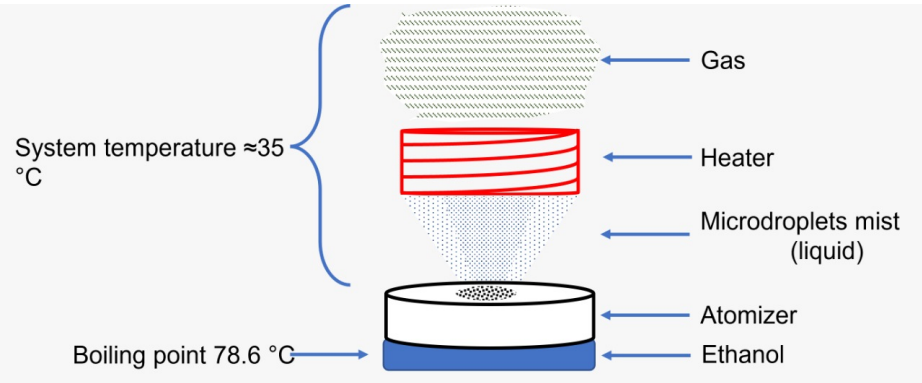


Fig. 1. Schematic of the rapid actuation system. The atomizer is placed directly over the liquid. Once the ultrasonic wave is applied to the atomizer a mist is generated. The mist contacts the heater which produces the evaporation.

In general, the classic control of rigid robots is based on analytical modeling, considering physical and mechanical characteristics of the robot such as: the number and type of degrees of freedom (dof), length of the links, centers of mass and gravity. However, the classic Control Based on Analytical Model (CBAM) is inadequate to deal with flexible robots due to their high nonlinearities and uncertainties during their displacements.

An alternative option to describe the dynamic of the soft robot is applying the data-driven statistical modeling (DDSM) generating a database from experimentation and considering the input and output signals history of the robot [12,13]. The Data-Driven Modeling and Control (DDMC) requires a minimum information of the robotic system in comparison to conventional CBAM, [11].

From the control theory viewpoint, the soft robot is considered as a nonlinear discrete-time system with high uncertainties and disturbance during their performance. Hence, the DDMC can be applied for all kind of robotic systems as manipulators, inertial, and non inertial including flexible robots, see for instance [6]. Consequently, data-driven identification and control are a novel option to apply for unknown nonlinear discrete-time systems as the case of robots, see [2].

The innovations of this work are as follows: (a) the proposal of a discrete-time position model for a soft robot based on DDSM from a phase change (liquid-gas) actuation integrated by a nebulizer and a heat exchanger; (b) the development of a novel control law based on a neuro-fuzzy network with a sliding surface function as input. The control law combines the reasoning-adaptation stage of the neuro-fuzzy network and the robustness against uncertainties of the sliding surface function.

Hence, the application of DDSM approximates the position of the soft robot. In addition, the proposal of an intelligent controller with adaptation stage based on the dynamic system response guarantee a position control through the discrete-time model of the soft robot. The structure of this paper is as follows: Section 2 presents the soft-robot modeling, Section 3 describes the control law design and results, and Section 4

summarizes the conclusion of this work.

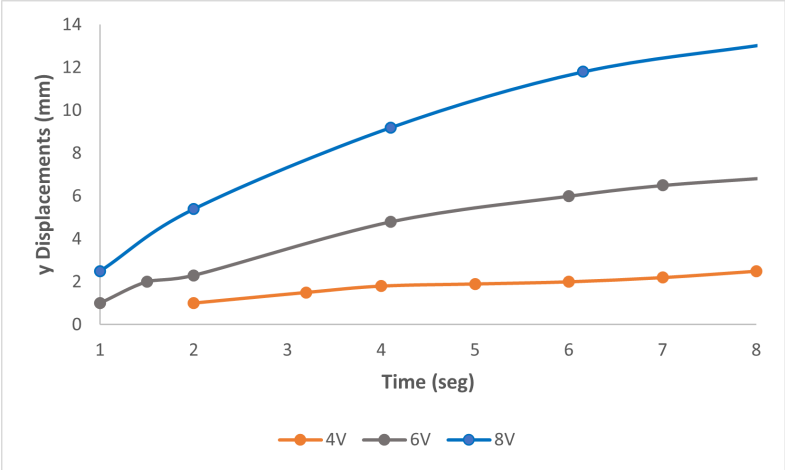


Fig. 2. The experimental data for 8 seconds when the heater was powered by 4, 6 and 8 V.

2 Robotic System

Through this section is presented the strategy to model the dynamic response of the soft- robot displacements applying a regression method from experimental data. Once, the regression model estimates the soft-robot displacements a nonlinear discrete-time function is proposed to validate the dynamic of the system into a closed -loop control. The proposed control is inspired on an intelligent control by an artificial neuro-fuzzy network with a sliding mode function as an input to compensate the nonlinearities and uncertainties during soft-robot displacements.

2.1 Experimental Setup

Remark 1 The proposal of the rapid actuation and the soft robot design have been discussed in [5]. As well, the data set presented in that research is used to obtain the regression model and the discrete-time model of the robot displacements to test the novel control law based on a neuro-fuzzy control in a closed-loop system.

2.2 Soft Robot Modeling

A regression model is presented using the least squares method, which is applied to the experimental results presented by [5]. The proposed regression model requires less data to estimate the displacements of the soft-robot, on the other hand, Artificial Neural Network (ANN) techniques require an extensive database for the training of

their parameters. In the case of Fuzzy Logic (FL) the model estimation is based on the human experience, then, a unique model is complicated to obtain. Figure 2 depicts the nonlinear relationship between the soft-robot displacements respect to the voltage input of the heater.

A strategy to obtain a trend-fit approximation function from experimental data is minimizing the residual errors sum of all available data between the measurements $y_{i,\text{measurements}}$ and the estimated $y_{i,\text{computed}}$ as follows:

$$S_r = \sum_{i=1}^{n} (y_{i,\text{measurements}} - y_{i,\text{computed}})^2. \tag{1}$$

Hence, a regression model is proposed based on a power equation in the following equation:

$$y = a_0 t^{a_1} V^{a_2}, \tag{2}$$

where y represents the position, t is the time, V is the input voltage and a_m represents the coefficients to determine by the least squares method. Thus, applying the natural logarithms properties is possible to linearize the equation (2):

$$\ln(y) = \ln(a_0) + a_1 \ln(t) + a_2 \ln(V). \tag{3}$$

The equation (3) fits experimental data for multivariable regressions. Replacing (3) in (1) is found the quadratic error sum function:

$$S_r = \sum_{i=1}^{n} (\ln(y) - \ln(a_0) - a_1 \ln(t) - a_2 \ln(V))^2. \tag{4}$$

The quadratic function in (4) is derived as $\dfrac{\partial S_r}{\partial a_m} = 0$ to find the coefficients a_m and it minimizes the error between the measurements and computed data:

$$\frac{\partial S_r}{\partial a_0} = \frac{2}{a_0} \sum_{i=1}^{n} [\ln(a_0) - a_1 \ln(t_i) - a_2 \ln(V_i)] = 0, \tag{5}$$

$$\frac{\partial S_r}{\partial a_1} = -2 \sum_{i=1}^{n} \ln(t_i) [\ln(a_0) - a_1 \ln(t_i) - a_2 \ln(V_i)] = 0, \tag{6}$$

$$\frac{\partial S_r}{\partial a_2} = -2 \sum_{i=1}^{n} \ln(v_i) [\ln(a_0) - a_1 \ln(t_i) - a_2 \ln(V_i)] = 0. \tag{7}$$

Once, the equations system for the regression model has been solved, the coefficients a_m of the proposed power function are obtained:

$$y = 0.0191\ t^{0.8076}\ V^{2.3752}. \tag{8}$$

Figure 3 shows a comparison between the experimental data and the estimated data according to (8).

Corollary 1 The polynomial interpolation model could be considered as unsatisfactory estimation, when the analyzed data set shows substantial errors. In contrast, a general approximation of data trend using a power regression is more useful to minimize the sum of the residual errors between the measured-output variable $y_{i,\text{measured}}$ from (8) and its mean $y_{i,\text{mean}}$.

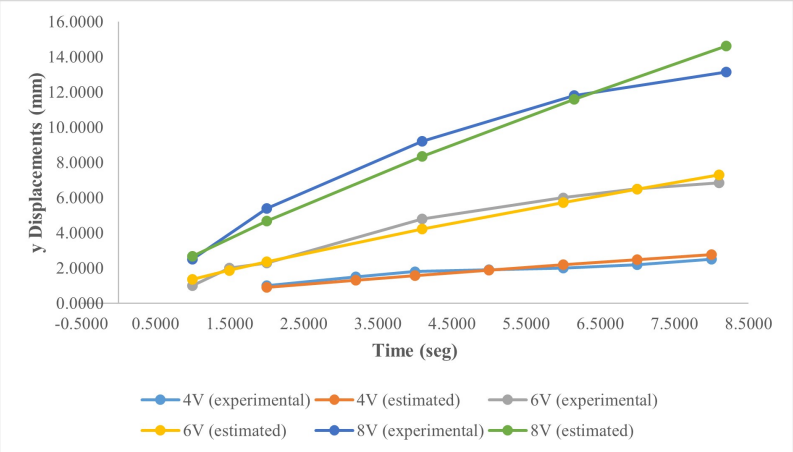


Fig. 3. Comparison between the experimental data and the regression model.

The magnitude of the residual error associated with the dependent variable (y) of the regression model is:

$$S_t = \sum_{i=n}^{n} (y_{i,\text{measured}} - y_{i,\text{mean}})^2. \tag{9}$$

The difference between $S_t - S_r$ quantifies the error between the data and a straight line instead of an average value. Since, the magnitude of this quantity depends on the scale, the difference is normalized respect to S_t to obtain the following form:

$$r^2 = \frac{S_t - S_r}{S_t}, \tag{10}$$

where r^2 and r are the determination and correlation coefficients, respectively. In a perfect fit $S_r \to 0$, therefore $r^2 = 1$. That means, 100% fit of the model according to the experimental data set. The standard error is defined as follows:

$$S_{y/x} = \sqrt{\frac{S_r}{n - (m+1)}}, \tag{11}$$

where $m = 3$ are the degrees of freedom in the power equation (8) and $n = 19$ are the data set numbers for this study in Figure 2. The adjusted coefficient of determination

$r^2_{adjusted}$ demonstrates the degree of effectiveness of the independent variables on the dependent variable:

$$r^2_{\text{adjusted}} = 1 - 1\frac{n-1}{n-m+1}(1-r^2). \tag{12}$$

Table 1 shows the regression model analysis. The results from the experimental data set and the regression model are concluded below. The regression equation (8) for the soft robot displacement has a correlation degree $r = 99.061\%$ between the inputs variables and the output variable.

Tabla 1. Statistical aspects of the regression model as a function of the soft robot displacement.

Parameter	Evaluation
r	0.99061
r^2	0.98131
r^2_{adjusted}	0.97897
$S_{y/x}$	0.53059
m	3
n	19

The regression model $r^2 = 98.131\%$ describes of the phenomenon uncertainties. Furthermore, the variables used for the model represents $r^2_{\text{adjusted}} = 97.7897\%$ of effectiveness. Finally, the standard error estimation is $S_{y/x} = \pm 0.53059$ mm. Once, the regression model is obtained is possible to approximate a nonlinear discrete-time function by the Taylor series expansion, as it is presented below.

Corollary 2 Taylor serie approximates the model through a polynomial function as:

$$y(x) = a_0 + a_1(x-c) + a_2(x-c)^2 + a_3(x-c)^3 + ... + a_n(x-c)^n. \tag{13}$$

The compact form from is:

$$y(x) = \sum_{n=0}^{\infty} a_n(x-c)^n, \tag{14}$$

where $x = c$ and (14) is derived successively:

$$\frac{d^n y(c)}{dx^n} = n! a_n. \tag{15}$$

Then, the approximation of the function by the Taylor serie is:

$$y(x) = \sum_{0}^{\infty} \frac{1}{n!}\frac{d^n y(c)}{dx_n}(x-c)^n. \tag{16}$$

The forward finite differences calculate a value in front of a reference point, where $x = x_{i+1}$, $x = c$, $\Delta x = x_{i+1} - x_i$ and $y(x_i) = y_i$, then the Taylor series is:

$$y_{i+1} = \frac{1}{0!}y_i + \frac{1}{1!}\frac{dy_i}{dx}\Delta x + \frac{1}{2!}\frac{d^2y_i}{dx^2}\Delta x^2 + ... + \frac{1}{n!}\frac{d^ny_i}{dx^n}\Delta x^n \tag{17}$$

$$= y_i + \frac{dy_i}{dx}\Delta x + \frac{1}{2}\frac{d^2y_i}{dx^2}\Delta x^2 + ... + \frac{1}{n!}\frac{d^ny_i}{dx^n}\Delta x^n. \tag{18}$$

For the case of backward finite differences, the Taylor series is obtained as follows, where $x = x_{i-1}$, $x = c$, $-\Delta x = x_{i-1} - x_i$ and $y(x_i) = y_i$, then Taylor serie is:

$$y_{i-1} = \frac{1}{0!}y_i - \frac{1}{1!}\frac{dy_i}{dx}\Delta x + \frac{1}{2!}\frac{d^2y_i}{dx^2}(-\Delta x^2) + ... + \frac{1}{n!}\frac{d^ny_i}{dx^n}(-\Delta x^n) \tag{19}$$

$$= y_i - \frac{dy_i}{dx}\Delta x + \frac{1}{2}\frac{d^2y_i}{dx^2}(-\Delta x^2) + ... + \frac{1}{n!}\frac{d^ny_i}{dx^n}(-\Delta x^n). \tag{20}$$

This series calculate a value behind of a reference point. Therefore, the expansion of Taylor series approximate the position function obtained with the regression model in (8) as follows:

$$y(k+1) = y(k) + \frac{\partial y}{\partial t}T_s + \frac{1}{2}\frac{\partial^2 y}{\partial t^2}T_s^2. \tag{21}$$

2.3 Transition from Regression Model to Discrete Model

The first derivative considers the coefficients and the regression model in (8) in order to approximate the discrete model as follows:

$$\frac{\partial y}{\partial t} \approx 0.01545k^{-0.1924}(V(k))^{2.3752}\left[\frac{\text{mm}}{\text{s}}\right]. \tag{22}$$

This term is associated to the velocity of the system, where k is the discrete time index and V is the input voltage. The second derivative is related to the acceleration of the system:

$$\frac{\partial^2 y}{\partial t^2} \approx -0.00297(V(k))^{2.3752}k^{-1.924}\left[\frac{\text{mm}}{\text{s}^2}\right]. \tag{23}$$

From the Taylor series is obtained:

$$y(k+1) = y(k) + \frac{\partial y}{\partial t}T_s + \frac{1}{2}\frac{\partial^2 y}{\partial t^2}T_s^2. \tag{24}$$

Substituting (23) and (24) in (22) is obtained the discrete-time function:

$$\begin{aligned} y(k+1) = y(k) &+ 0.01545k^{-0.1924}(u(k))^{2.3752}T_s \\ &-\tfrac{1}{2}(-0.00297(u(k))^{2.3752}k^{-1.924})T_s^2. \end{aligned} \tag{25}$$

The expression in (25) approximates the nonlinear system dynamic of the soft robot working within discrete-time in order to apply a novel neurofuzzy control.

3 Control Law

This section presents a novel intelligent controller inspired by an artificial neuro-fuzzy network considering the nonlinear discrete-time function in (25), which describes the displacement of the soft robot. The input signal to the nonlinear discrete-time function is the voltage (control variable) applied to the heater during the soft-robot actuation, and the output signal is the displacement (controlled variable) generated by the soft-robot. Therefore, the following assumptions should be satisfied for the control law design.

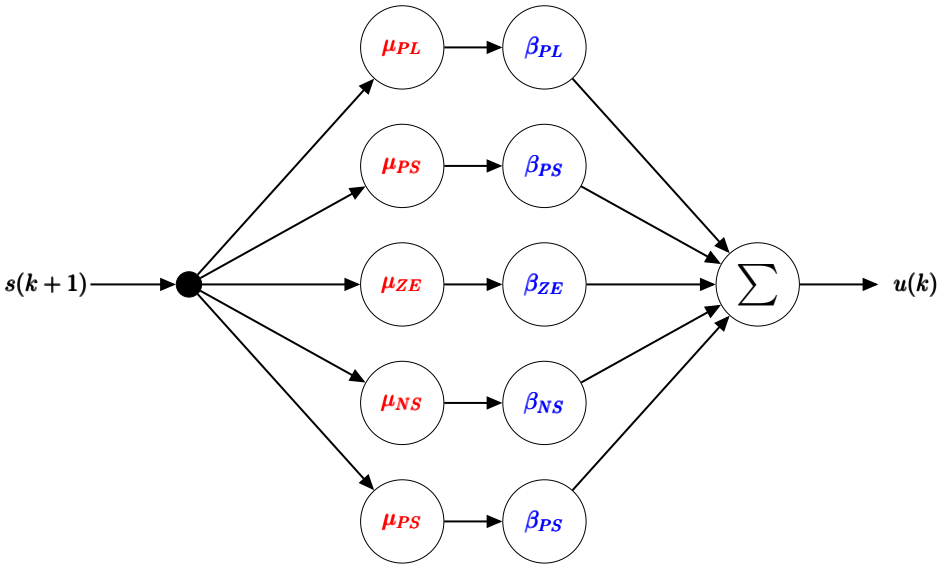


Fig. 4. NFN architecture and $s(k+1)$ as input signal.

Assumption 1 The robot is considered Lipschitz and exists a positive constant L that defines the direct relationship between system input-output $\| y(k+1) \| \leq L \| u(k) \|$. that means, a change of the system output imposes a change of the system input.

Assumption 2 The output of the robotic system is observable, *i.e.*, $y(k+1) = \hat{\Phi}(k)u(k)$ $\forall k > 0$. It is possible to know the equivalent model of the system from the measured output signals.

The artificial neuro-fuzzy network is characterized by an adaptive stage based on a human experience and the intuitive initial parameters selection. The adaptation stage adjusts its parameters using the descending gradient technique. The neuro-fuzzy network considers the plant as an unknown nonlinear system working in the discrete-time domain.

Therefore, the neuro-fuzzy network only requires to know the input and output signals from the system to control the plant. The structure of Neuro-Fuzzy Network (NFN) is based on the human knowledge and the intuitive initialization of its parameters [10] as is referred Figure 4. A sliding mode surface function $s(k+1)$ in (26) is proposed as input to the NFN:

$$s(k+1) = C_1 e(k+1) + C_2 e(k), \tag{26}$$

where $C_1,\ C_2 \in \mathbb{R}^+$ and the position error is defines as:

$$e(k+1) = y_d(k+1) - y(k+1), \tag{27}$$

where $y(k+1)$ is the current position and $y_d(k+1)$ is the desired position.

3.1 Proposed NFN Architecture

NFN structure has 4 layers and 5 nodes.

- Layer 1. This layer is considered as the input to the artificial neural network $s(k+1)$, also this signal is sent to each node in the next layer.
- Layer 2. This layer contains the membership functions. Each node in this layer is a membership function corresponding to the design of the linguistic variables. The output of each node is calculated as follows:

$$\phi(k) = \mu(s(k)). \tag{28}$$

- Layer 3. This layer is the adaptation stage where the parameters $\beta(k+1)$ are adjusted.
- Layer 4. This layer is the output of the NFN:

$$O(k) = \sum_{i=1}^{N} \phi(k)\beta(k), \tag{29}$$

where N represents the number of linguistic variables.

3.2 Adaptation Algorithm

An adaptive technique based on the descending gradient method is proposed to adjust NFN parameters. First, an objective function is defined to achieve the optimal value of the network parameters. The parameters are adjusted at each time step through a quadratic function $\xi(k+1)$ in terms of the control error:

$$\xi(k+1) = \frac{1}{2} s^2(k+1). \tag{30}$$

According to the descending gradient method, the adaptation of the parameters $\beta(k+1)$ is computed as follows:

$$\beta(k+1) = \beta(k) - \eta \frac{\partial \xi(k+1)}{\partial \beta(k)}, \tag{31}$$

where η is the learning rate and applying the chain rule:

$$\frac{\partial \xi(k+1)}{\partial \beta(k)} = \frac{\partial \xi(k+1)}{\partial s(k+1)} \frac{\partial s(k+1)}{\partial e(k+1)} \frac{\partial e(k+1)}{\partial y(k+1)} \frac{\partial y(k+1)}{\partial u(k)} \frac{\partial u(k)}{\partial O(k)} \tag{32}$$

$$= \qquad s(k+1)C_1[-1]\hat{\Phi}(k)\mu(s(k)), \tag{33}$$

When substituting (33) in (31) is found the adaptation law:

$$\beta(k+1) = \beta(k) + \eta s(k+1)C_1\hat{\Phi}(k)\mu(s(k)). \tag{34}$$

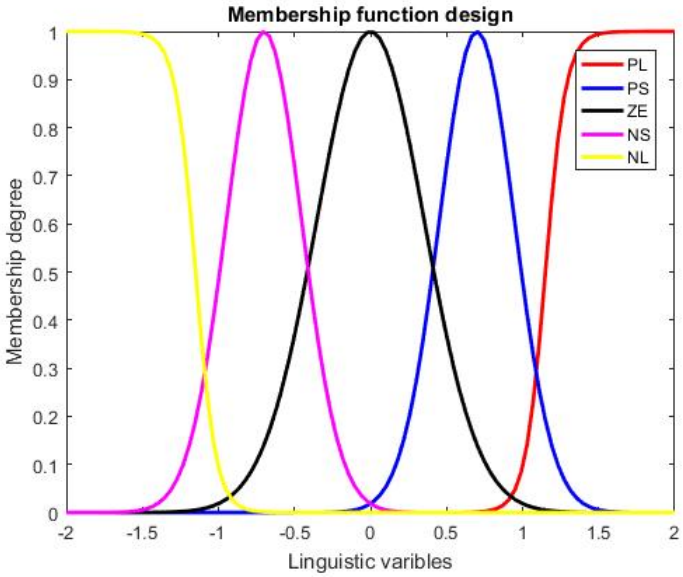


Fig. 5. Design of the membership functions for SMC-NFN.

Remark 2 where $\hat{\Phi}(k) = y(k+1)/u(k)$ denotes the approximated input and output relationship of the system in (25), therefore the representation of the ideal system $\Phi^*(k)$ is given by:

$$\Phi^*(k) = \hat{\Phi}(k) + \epsilon(k), \tag{35}$$

where $\epsilon(k)$ is the estimation error and the control law is:

$$u(k) = \mu(s(k))\beta_s(k). \tag{36}$$

Remark 3 The novelties in the proposed neuro-fuzzy-control are:

– The five membership functions are designed according to the robot displacement as is shown in Figure 5.

– The sliding surface function $s(k+1)$ improves the tracking control and robustness.

– The adaptation law (34) permits to update the neuro-fuzzy parameters $\beta(k+1)$ and it captures instantaneous changes on the system.

3.3 Simulations

The five linguistic variables are designed according to the physical characteristics of the robot. Therefore, the linguistic variables are $\mu_i : P_L$ is positive large, P_S is positive small, Z_E is zero, N_S is negative small and N_L es negative large. Figure 5 shows the membership function design and the Table 2 shows the control setting parameters. The IF-THEN rules are established by the input function $s(k+1)$ and the output $u(k)$:

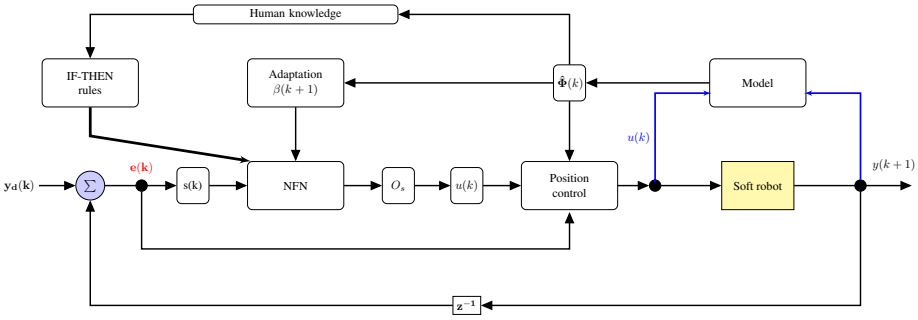


Fig. 6. Block diagram of the closed-loop system.

Tabla 2. Control setting parameters values.

Parameters	Value
$\beta_{PL}(0)$	2.25
$\beta_{NL}(0)$	1.85
$\beta_{ZE}(0)$	0.5
$\beta_{NS}(0)$	-1.65
$\beta_{NL}(0)$	-1.85
C_1	1.45
C_2	0.55
η	0.85

– IF $s(k+1)$ Is positive large (P_L), THEN $u(k)$ Is positve large (P_L).

– IF $s(k+1)$ is positive small (P_S), THEN $u(k)$ is positve small (P_S).
– IF $s(k+1)$ is zero (Z_E), THEN $u(k)$ Is Zero (Z_E).
– IF $s(k+1)$ is negative small (N_S), THEN $u(k)$ Is negative small (N_S).
– IF $s(k+1)$ is negative large (N_L), THEN $u(k)$ Is negative large (N_L).

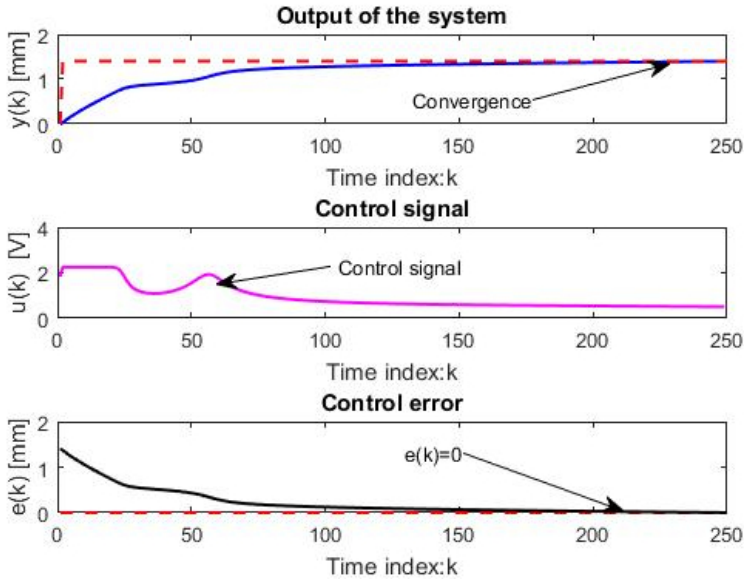


Fig. 7. SMC-NFN controller with adaptive parameters $\beta(k+1)$.

Figure 6 depicts the closed-loop system. The control law in (36) represents the system input, the nonlinear discrete-time function in (25) represents the system output, and the adaptation law is in (34). NFN guarantees the adaptation and learning stages based on the plant empirical knowledge, as well the sliding mode function in (26) provides robustness against uncertainties inside of the NFN structure.

Figure 7 shows the simulation of the controller for a regulation position task where the control law design remedies the control error convergence to zero, successfully. Figure 8 depicts the parameters $\beta(k+1)$ for the the proposed adaptive law in (34). Moreover, the proposed controller is compared to a conventional PID controller in order to review its advantages.

The conventional PID controller is:

$$u(k) = K_p e(k) + K_d \left[e(k) - e(k-1)\right] T_s + K_i \left[\frac{e(k) - e(k-1)}{T_s}\right]. \tag{37}$$

where the proportional, integral and derivative gains are $K_p = 3.95$, $K_i = 0.01$ and $K_d = 0.01$, respectively. Figure 10 depicts the simulation results applied to the discrete-time model in (25). The comparison bewteen the proposed control and the conventional

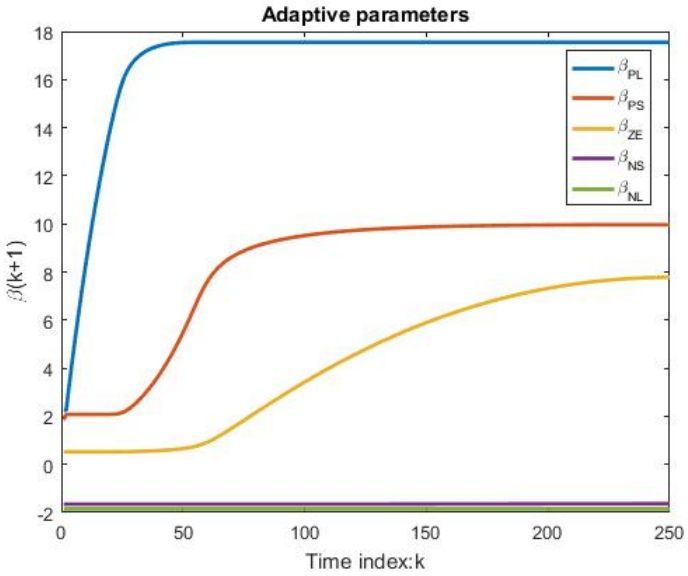


Fig. 8. Adaptive law for $\beta_i(k+1)$.

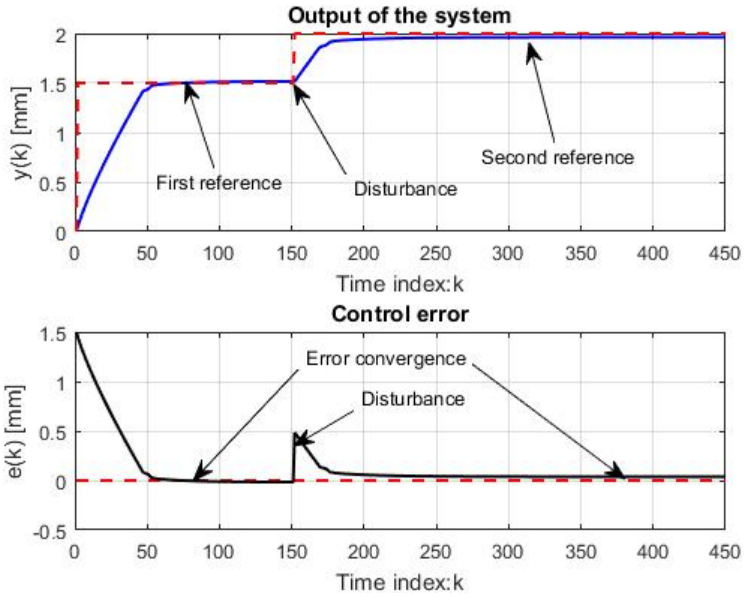


Fig. 9. SMC-NFN controller with disturbance response.

control is observed directly on the error convergence, meanwhile the control error $e(k) = 0$ [mm] in the neuro-fuzzy control, and the control error

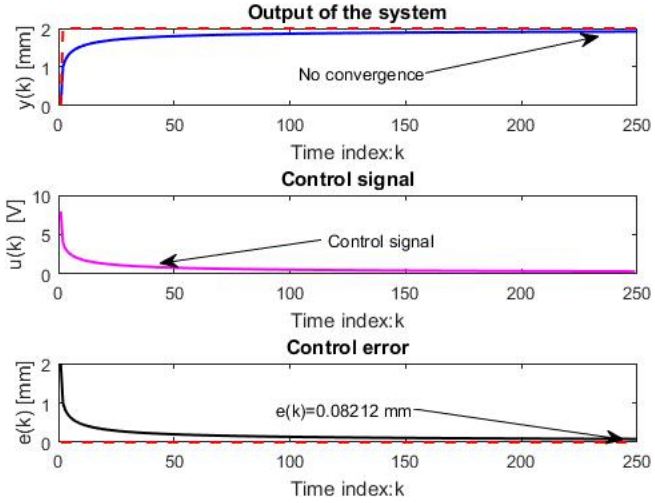


Fig. 10. PID controller.

$e(k) = 0.08212\ [\mathrm{mm}]$ in PID is not enough to converge.

Additional simulation of the control system is presented at Figure 9 in order to validate the proposed neuro-fuzzy control, a disturbance was included in the simulation to demonstrate the adaptation and the response against sudden changes in the system.

4 Conclusions

An statistical data regression model is proposed to describe the displacements of a soft robot based on the historical response (experimentation) of the input and output signals. Hence, the expansion of multi-variable Taylor series approximated a nonlinear discrete-time model from the SDDM.

A novel control law for the nonlinear discrete-time model of the robot was proposed based on the concept of NFN and a sliding surface function as input. The adaptive law permits to capture instantaneous changes in the closed loop system. Moreover, the sliding surface function improves the tracking control. The control law presented combines the adaptive stage and human experience knowledge of the system from NFN and the uncertainties compensation from the sliding surface function.

As well, the proposed control law guarantee the control error convergence in comparison to a conventional PID controller that does not compensate the non-linearities of the system. As future work, the research is led to test a Data-Driven Model and Control (DDMC) in an experimental setup only using the association of the input signal (heater voltage) and the output signal (soft-robot displacement). Moreover, the stability analysis will be developed to guarantee the control error convergence.

References

1. Amend, J .R., Brown, E., Rodenberg, N., Jaeger, H. M., Lipson, H.: A positive pressure universal gripper based on the bamming of granular material. In: IEEE Transactions on Robotics, vol. 28, no. 2, pp. 341–350 (2012)
2. Gómez, J., Treesatayapun, C., Morales, A.: Data-driven identification and control based on optic tracking feedback for robotic systems. Int J Adv Manuf Technol, vol. 113, pp. 1485–1503 (2021)
3. Lee, H. J., Loh, K. J.: Liquid vaporization actuated soft structures with active cooling and heat loss control. Smart Materials and Structures, vol. 30, no. 5 (2021)
4. Lee, H. J., Melchor, N., Chung, H., Loh, K. J.: Characterization of a soft gripper with detachable fingers through rapid evaporation. In: 3rd IEEE International Conference on Soft Robotics (RoboSoft) pp. 83–88 (2020)
5. Lee, H. J., Prachaseree, P., Loh, K. J.: Rapid soft material actuation through droplet evaporation. In: Soft Robot, vol. 8, no. 5, pp. 555–563 (2021)
6. Li, M., Kang, R., Geng, S., Guglielmino, E.: Design and control of a tendon-driven continuum robot. Transactions of the Institute of Measurement and Control, vol. 40, no. 11, pp. 3263–3272 (2018)
7. Lin, H. T., Leisk, G .G., Trimmer, B.: GoQBot: a caterpillar-inspired soft-bodied rolling robot. Bioinspiration & biomimetics, vol. 6, no. 2 (2011)
8. Polygerinos, P., Wang, Z., Galloway, K.C., Wood, R.J., Walsh, C.J.: Soft robotic glove for combined assistance and at-home rehabilitation. Robotics and Autonomous Systems, vol. 73, pp. 135–143 (2015)
9. Shepherd, R. F., Ilievski, F., Choi, W., Morin, S. A., Stokes, A. A., Mazzeo, A. D., Whitesides, G. M.: Multigait soft robot. In: Proceedings of the national academy of sciences, vol. 108, no. 51, pp. 20400–20403 (2011)
10. Facundo-Flores, L., Treesatayapun, C., Baltazar, A.: Design of a pose and force controller for a robotized ultrasonic probe based on neural networks and stochastic gradient approximation. In: IEEE Sensors Journal, vol. 21, no. 5, pp. 6224–6233 (2021)
11. Zen, Z., Cao, R., Hou, Z.: Mimo model free adaptive control of two degree of freedom manipulator. In: 2018 IEEE 7th data driven control and learning systems conference, pp. 693–697 (2018)
12. Sivaiah, P., Chakradhar, D.: Modeling and optimization of sustainable manufacturing process in machining of 17-4 PH stainless steel. Measurement, vol. 134, pp. 142–152 (2019)
13. Montes-González, F. A., Rodríguez-Rosales, N. A., Ortiz-Cuellar, J. C., Muñiz-Valdez, C. R., Gómez-Casas, J., Galindo-Valdés, J. S., Gómez-Casas, O.: Experimental Analysis and Mathematical Model of FSW Parameter Effects on the Corrosion Rate of Al 6061-T6-Cu C11000 Joints. Crystals, vol. 11, no. 3, pp. 294 (2021)

ISSN 1870-4069

Optimización de mecanismos planos de 4 y 6 eslabones para el desarrollo de un prototipo de prótesis transfemoral

Eicarl Saynes-Vazquez[1], Esther Lugo González[2]

[1] Universidad Tecnológica de la Mixteca,
División de Posgrado,
México

[2] Universidad Tecnológica de la Mixteca,
Instituto de Electrónica y Mecatrónica,
México

eicarl43@gmail.com, elugog@mixteco.utm.mx

Resumen. En este artículo se presenta un procedimiento de optimización para la síntesis de mecanismos planos de 4 y 6 barras, basado en técnicas de cómputo evolutivo: Algoritmos Genéticos (AG) y Algoritmo Evolutivo Diferencial (AED). El objetivo es comparar los resultados obtenidos y determinar qué técnica y qué mecanismo cumple con el seguimiento de la poloide (curva característica que describe el Centro Instantáneo de Rotación de una rodilla protésica (CIR)). Se plantea la función objetivo para cada caso, así como las condiciones de restricción y los operadores que tienen en común estos algoritmos: selección, cruce y mutación. Como resultados se exponen tablas que comparan el error en el seguimiento de una línea recta y una trayectoria para obtener la poloide característica, así como gráficas que muestran la optimización. Finalmente, se tiene la propuesta para el diseño de un prototipo de prótesis transfemoral utilizando el mecanismo de 4 barras resultante.

Palabras clave: Síntesis de mecanismos, algoritmos genéticos, algoritmo evolutivo diferencial, prótesis transfemoral.

Optimization of 4 and 6 Link Planar Mechanisms for the Development of a Transfemoral Prosthesis Prototype

Abstract. This paper presents an optimization procedure for the synthesis of 4- and 6-bar planar mechanisms, based on evolutionary computation techniques: Genetic Algorithms (GA) and Differential Evolutionary Algorithm (DEA). The objective is to compare the results obtained and determine which technique and which mechanism complies with the tracking of the poloid (characteristic curve describing the Instantaneous

Center of Rotation (ICR) of a prosthetic knee). The objective function is proposed for each case, as well as the restriction conditions and the operators that these algorithms have in common: selection, crossover and mutation. As results, tables comparing the error in the tracking of a straight line and a trajectory to obtain the characteristic poloid are presented, as well as graphs showing the optimization. Finally, there is a proposal for the design of a transfemoral prosthesis prototype using the resulting 4-bar mechanism.

Keywords: Synthesis of mechanisms, genetic algorithms, differential evolutionary algorithm, transfemoral prosthetic.

1. Introducción

Las prótesis móviles por lo general utilizan mecanismos que deben ser sintetizados para obtener sus dimensiones, tipo o forma. Existen técnicas clásicas y modernas para la síntesis [6], como las que incluyen la aplicación de mínimos cuadrados o algoritmos evolutivos. Entre las técnicas modernas más utilizadas está la optimización heurística y metaheurística, cuyo propósito es minimizar o maximizar los resultados de una función objetivo, como en este caso, se busca la minimización del error en el seguimiento de una trayectoria.

Estas se utilizan debido a que entre las restricciones del método clásico se encuentran el número de puntos de precisión que se pueden tomar para definir una trayectoria deseada y un espacio de búsqueda local. En [2] se lleva a cabo el desarrollo de una síntesis óptima de mecanismos de 4 barras con algoritmos genéticos, y posteriormente en [1] se desarrollan la síntesis y optimización de mecanismos de 4 y 6 barras a través de diferentes algoritmos de cómputo evolutivo, como el AG (Algoritmo genético), AED (Algoritmo evolutivo diferencial) y el Algoritmo de Optimización por Enjambre de Partículas (PSO, por sus siglas en inglés).

En ambos trabajos, el objetivo es hallar mecanismos óptimos para el seguimiento de diferentes trayectorias, minimizando el error que se puede presentar entre la trayectoria generada y la deseada. En [4] y [9] se desarrolla la síntesis y optimización de mecanismos de 4 barras a través de AG, AED, Algoritmo Evolutivo Diferencial Autoadaptativa (JADE, por sus siglas en inglés) y el Algoritmo Competitivo Imperialista (ICA, por sus siglas en inglés), como estrategias diferentes de optimización.

En ambos trabajos los autores exponen los resultados obtenidos, en los cuales se observa el seguimiento de las trayectorias predefinidas resaltando la factibilidad del uso de estas técnicas como herramientas de síntesis. En [16] se presenta el estudio de mecanismos de 6 barras utilizando el AED proponiendo diferentes trayectorias para comparar los resultados gráficos y analíticos obtenidos. Para cada caso de estudio, los algoritmos convergen a soluciones óptimas con ciertas variaciones en el valor del error y en el seguimiento de las trayectorias.

En [4] exponen que el ICA ofrece los mejores resultados para la optimización de los casos de estudio; no obstante, en [9] mencionan que el AED tipo best/2 presenta las mejores soluciones para la mayoría de los casos de diseño. Debido al desarrollo de los algoritmos evolutivos y a la exigencia de mejorar la función protésica de la rodilla (estabilidad, movilidad y seguridad), se realizan estudios de los mecanismos planos para diseñar dispositivos protésicos. El objetivo general, es lograr que el Centro Instantáneo de Rotación o CIR de un mecanismo de 4 o 6 barras se iguale o aproxime al CIR anatómico de una pierna humana para cumplir con los requerimientos del usuario.

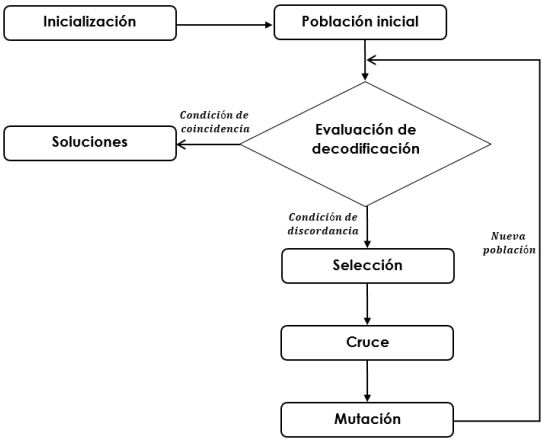


Fig. 1. Funcionamiento del algoritmo AG [17].

Así en [6,7,17] se presentan las síntesis y optimizaciones de mecanismos de 4 y 6 barras, a través de AG y AED, para el desarrollo de dispositivos protésicos pasivos y activos. Los autores coinciden con obtener un valor mínimo de error de seguimiento de trayectoria para asegurar que los dispositivos protésicos reflejen un CIR que se iguale o aproxime al CIR anatómico de una pierna humana y así garantizar una marcha adecuada y estable para las personas con amputación. Por último, resaltan que para algunos casos el mecanismo de 6 barras lleva a cabo un mejor seguimiento de la poloide, pero mecánicamente no es tan práctico como el de 4 eslabones porque presenta un incremento de peso y de piezas.

Con el objetivo de optimizar la síntesis de eslabonamientos planos para obtener el diseño de un prototipo de prótesis transfemoral, este artículo presenta como primer punto los conceptos necesarios para establecer el diseño del CIR de la rodilla que debe seguir el mecanismo policéntrico, posteriormente se tiene el análisis matemático de la síntesis de un mecanismo de 4 barras, uno de 6 tipo Watt II y otro tipo Stephenson III para determinar cuál es el óptimo en este tipo de estudio. También se plantea la función objetivo y los resultados obtenidos al

aplicar las técnicas del AG y AED. Finalmente, se presentan las conclusiones del trabajo y se determina cuál es la solución óptima para el diseño final.

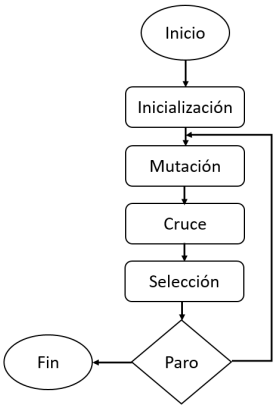


Fig. 2. Funcionamiento del algoritmo AED [3].

1.1. Rodilla policéntrica

El CIR se localiza en la intersección de las prolongaciones de las líneas de los enlaces anterior y posterior, los cuales conectan la sección del encaje a la pierna en la prótesis. Es muy utilizado en el diseño de prótesis transfemorales que incluyen un mecanismo de rodilla de 4 barras [17]. La curva que describe la trayectoria del CIR se conoce como poloide. Un mecanismo de 4 barras, desde el punto de vista biomecánico, se puede clasificar en 3 tipos según la ubicación de su CIR como se menciona en [14]:

– CIR Elevado.
– Hiper-Estabilizado.
– Control voluntario.

Esta última configuración proporciona estabilidad durante el contacto inicial y el despegue del talón, ya que el CIR se encuentra dentro de la zona de estabilidad de la pierna [14]. Su diseño se enfoca en usuarios con niveles de actividad moderada.

2. Algoritmos de cómputo evolutivo

Los algoritmos de cómputo evolutivo son estrategias de optimización y búsqueda de soluciones. Basan su funcionamiento en métodos de evolución, en fenómenos físicos o biológicos [11]. Entre los más empleados se encuentran los ya referidos AG, AED, PSO, JADE, ICA, que tienen la característica de resolver

problemas de manera rápida y robusta, reduciendo el tamaño efectivo del espacio de búsqueda.

El diagrama de flujo en la Figura 1 expone las diferentes etapas del funcionamiento de un algoritmo genético: a partir de una población inicial generada se lleva a cabo una evaluación de decodificación, si se encuentra en discordancia, el proceso se realiza de forma lineal y retorna para el análisis de una nueva población, de lo contrario, la condición se cumple y por lo tanto, existen soluciones óptimas. Si el algoritmo siguió un curso lineal, se presentarán tres de sus operadores más importantes [3, 12].

2.1. Algoritmos genéticos

- Selección: El proceso consiste en elegir individuos con base en su contribución de aptitud con respecto al total de la población.
- Cruce: El cromosoma intercambia genes de dos individuos completamente adaptados, este proceso se realiza cortando dos cadenas en una posición elegida al azar e intercambiándolas según el tipo de cruce que se elija.
- Mutación: Consiste en elegir aleatoriamente un gen durante la reproducción o cruzamiento y cambiarlo.

2.2. Algoritmo evolutivo diferencial

El AED es un algoritmo poblacional de búsqueda directa y simple, optimiza hasta alcanzar el óptimo global en funciones multimodales, no diferenciables y no lineales [3]. En la Figura 2, se presenta un diagrama de flujo que expone las diferentes etapas del funcionamiento del algoritmo. Aunque su funcionamiento es similar al AG, una de sus diferencias es que emplea el operador de mutación de forma distinta.

En este se provee información que se intercambia entre las distintas soluciones encontradas, siempre que el proceso de inicialización se haya llevado a cabo. El objetivo es recombinar la población para producir una nueva, aplicando la mutación diferencial [3, 11].

- Cruce: Incrementa la diversidad del vector de parámetros y complementa la estrategia de mutación utilizada.
- Selección: Consiste en encontrar las mejores soluciones verificando si el nuevo elemento producido por la mutación y el cruce es mejor al anterior.

3. Desarrollo

El objetivo principal de este estudio es obtener el mecanismo que cumpla de forma óptima con el seguimiento de una trayectoria específica, utilizando los algoritmos evolutivos AG y AED para resolver su síntesis a través de la ecuación de Freudenstain y de la función objetivo que busca obtener el mínimo error

entre la trayectoria deseada y la generada durante el seguimiento de la poloide. El procedimiento para obtener la síntesis se realiza para mecanismos de 4 y 6 barras de configuración tipo Watt II y Stephenson III, se observa en la Figura 3a. Inicia con el planteamiento de la ecuación de lazo cerrado del diagrama cinemático.

Se reescribe el lazo vectorial sustituyendo la propiedad de Euler en los términos $e^{j\theta}$. Después, se plantean constantes para simplificar el análisis y se obtiene la ecuación de Freudenstein. A continuación, se sustituyen propiedades del ángulo medio para dar solución al sistema con la ecuación de la fórmula general. Finalmente, se obtienen los ángulos $\theta_3,\ \theta_4$ o θ_i del sistema.

3.1. Mecanismo de 4 barras

Con base en el diagrama cinemático de la Figura 3b, se escribe la ecuación de lazo cerrado del mecanismo:

$$\vec{r_2} + \vec{r_3} = \vec{r_1} + \vec{r_4}. \tag{1}$$

La Ecuación (1) se divide en su parte real e imaginaria igualadas a cero, mediante la propiedad de Euler. Después, se elimina θ_3 y se resuelve el sistema cartesiano para θ_4 [8]:

$$r_3 \cos\theta_3 = r_1 + r_4 \cos\theta_4 - r_2 \cos\theta_2, \tag{2}$$

$$r_3 \sin\theta_3 = r_4 \sin\theta_4 - r_2 \sin\theta_2. \tag{3}$$

Se elevan al cuadrado los términos de la Ecuación (3) y se suman:

$$\begin{aligned} r_3^2 = r_1^2 + r_2^2 + r_4^2 + 2r_1 r_4 \cos\theta_4 - 2r_1 r_2 \cos\theta_2 \\ -2r_4 r_2(\cos\theta_2 \cos\theta_4 + \sin\theta_2 \sin\theta_4). \end{aligned} \tag{4}$$

Para simplificar la expresión (4) se definen las constantes:

$$K_1 = \frac{r_1}{r_2}, \tag{5}$$

$$K_2 = \frac{r_1}{r_4}, \tag{6}$$

$$K_3 = \frac{r_1^2 + r_2^2 - r_3^2 + r_4^2}{2r_4 r_2}. \tag{7}$$

Así, se obtiene la ecuación de Freudenstein:

$$\therefore \quad \cos(\theta_2 - \theta_4) = K_1 \cos\theta_4 - K_2 \cos\theta_2 + K_3. \tag{8}$$

Se sustituye en la ecuación (8) las propiedades semi-angulares:

$$\begin{aligned} \tan^2\left(\frac{\theta_4}{2}\right)(\cos\theta_2 - K_1 - K_2\cos\theta_2 + K_3) + \tan\left(\frac{\theta_4}{2}\right)(-2\sin\theta_2) \\ +K_1 + K_3 - (K_2 + 10)\cos\theta_2 = 0. \end{aligned} \tag{9}$$

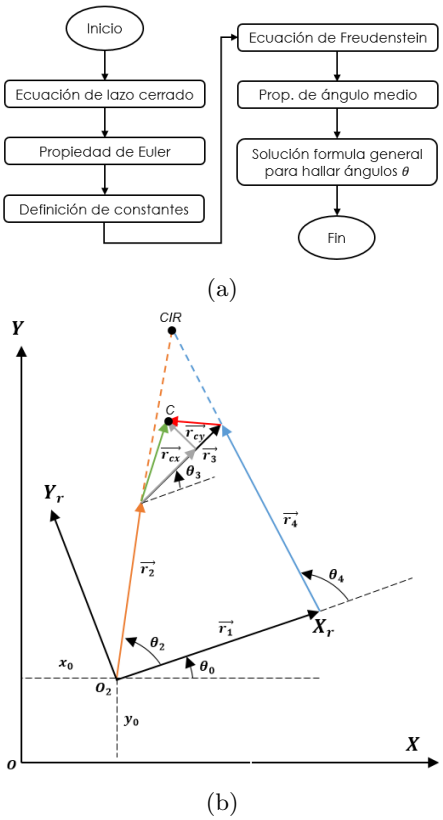


Fig. 3. a) Diagrama de flujo para la síntesis de mecanismos planos, b) Diagrama cinemático de mecanismo de 4 eslabones [2].

Simplificando la ecuación (9):

$$A = \cos\theta_2 - K_1 - K_2\cos\theta_2 + K_3, \tag{10}$$

$$B = -2\sin\theta_2, \tag{11}$$

$$C = \text{`}K_1 + K_3 - (K_2 + 1)\cos\theta_2, \tag{12}$$

$$\implies \tan^2\left(\frac{\theta_4}{2}\right)A + \tan\left(\frac{\theta_4}{2}\right)B + C = 0. \tag{13}$$

La solución de la ecuación (13) se expresa como [8]:

$$\therefore \quad \theta_{4_{1,2}} = 2\arctan\left(\frac{-B \pm \sqrt{B^2 - 4AC}}{2A}\right). \tag{14}$$

Por otro lado, la solución para el ángulo θ_3 es similar al de θ_4. La diferencia ahora es que en el sistema de la Ecuación (3) se elimina θ_4 y se resuelve para θ_3.

Tabla 1. Configuración de mecanismos por [13].

Configuración	θ_3	θ_4
Abierta	$+\sqrt{}$	$-\sqrt{}$
Cruzada	$-\sqrt{}$	$+\sqrt{}$

Análogamente a la Ecuación (14), la solución se expresa como:

$$\therefore \quad \theta_{3_{1,2}} = 2\arctan\left(\frac{-E \pm \sqrt{E^2 - 4DF}}{2D}\right). \tag{15}$$

Tanto la Ecuación (14) como la (15) tienen dos soluciones, obtenidas a partir de las condiciones $\pm$ en el radical. Éstas se conocen como las configuraciones cruzada y abierta del mecanismo. Según la configuración de un mecanismo de cuatro barras en la Tabla 1 deben elegirse los signos [13].

Finalmente, en [2] se menciona que el punto C del acoplador que seguirá la trayectoria deseada (ver Figura 3b), se encuentra en el marco de referencia rotado $O_2X_rY_r$ respecto del marco de referencia global OXY, por ello se hace uso de una matriz de rotación para adecuar la síntesis y hallar C_x, C_y como a continuación se explica:

$$\begin{bmatrix} C_x \\ C_y \end{bmatrix} = \begin{bmatrix} \cos\theta_0 & -\sin\theta_0 \\ \sin\theta_0 & \cos\theta_0 \end{bmatrix} \begin{bmatrix} C_{X_r} \\ C_{Y_r} \end{bmatrix} + \begin{bmatrix} x_0 \\ y_0 \end{bmatrix}, \tag{16}$$

donde C_{X_r} y C_{Y_r} quedan definidos por el siguiente lazo vectorial:

$$\vec{r_2} + \vec{r_{cx}} + \vec{r_{cy}} = C(X_r, Y_r). \tag{17}$$

Dividiendo la parte real e imaginaria de la Ecuación (17), se tiene:

$$\therefore \quad C_{X_r} = r_2 cos\theta_2 + r_{cx} cos\theta_3 - r_{cy} sen\theta_3, \tag{18}$$

$$C_{Y_r} = r_2 sen\theta_2 + r_{cx} sen\theta_3 + r_{cy} cos\theta_3. \tag{19}$$

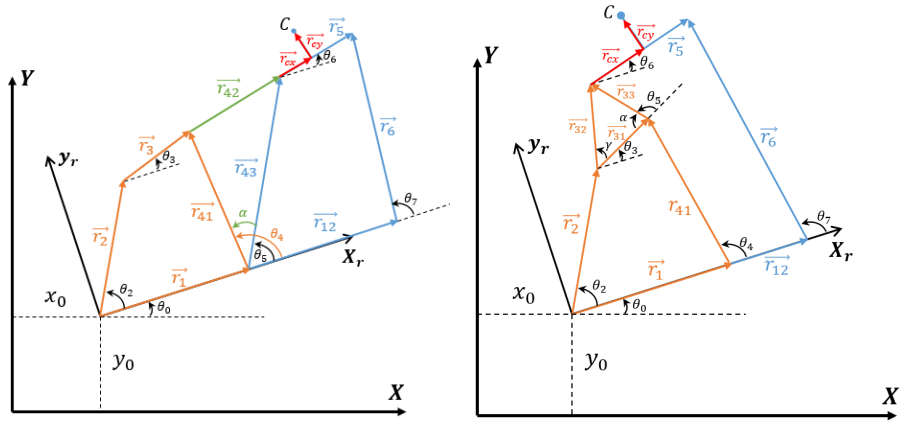
3.2. Mecanismo de Watt tipo II

El eslabonamiento tipo Watt II mostrado en la Figura 4a, se forma a través de dos mecanismos de cuatro barras que comparten un eslabón ternario [8]. En este caso, se analiza la primera malla o circuito, formado por los vectores r_1, r_2, r_3 y r_{41}. De este análisis se hallan los ángulos θ_3, θ_4, α y se desarrolla la segunda malla formada por los eslabones r_{43}, r_5, r_6 y r_{12}. La entrada $\theta_5 = \theta_4 - \alpha$ y las incógnitas para este segundo caso son θ_6 y θ_7. Ecuación de lazo cerrado para el desarrollo de la primera malla:

$$\vec{r_2} + \vec{r_3} = \vec{r_1} + \vec{r_{41}}. \tag{20}$$

La ecuación de lazo cerrado para la segunda malla:

$$\vec{r_{43}} + \vec{r_5} = \vec{r_{12}} + \vec{r_6}. \tag{21}$$



(a) Watt II (b) Stephenson III

Fig. 4. Diagramas cinemáticos de mecanismos de 6 barras [5].

Para ambas mallas se definen las constantes $(K_1, ..., K_{10})$ que se utilizan para simplificar las ecuaciones de Freudenstein. Posteriormente, se declaran las constantes $(A, ..., F)$ para la malla uno y $(An, ..., Fn)$ para la malla 2, que sirven para determinar la solución del sistema a través de los ángulos θ_4, θ_3, θ_7 y θ_6. Por último, el lazo vectorial para el seguimiento de la trayectoria en C es:

$$\vec{r_1} + \vec{r_{43}} + \vec{r_{cx}} + \vec{r_{cy}} = C(X_r, Y_r). \tag{22}$$

3.3. Mecanismo de Stephenson tipo III

La Figura 4b, consta de un mecanismo de 4 barras formado por los eslabones r_1, r_2, r_{31} y r_{41} y uno de 5 formado por r_{12}, r_{41}, r_{33}, r_5 y r_6. De este primer análisis se hallan θ_3, θ_4, α y γ. Para la segunda malla (mecanismo de 5 barras) se tiene:

$$\vec{r_{41}} + \vec{r_{33}} + \vec{r_5} = \vec{r_{12}} + \vec{r_6}. \tag{23}$$

Después de definir las constantes para simplificar el desarrollo matemático, la solución de la ecuación cuadrática en términos de θ_6 se expresa como:

$$\therefore \quad \theta_{6_{1,2}} = 2\arctan\left(\frac{-En \pm \sqrt{En^2 - 4DnFn}}{2Dn}\right). \tag{24}$$

La solución para θ_7 es el mismo procedimiento. Por último, el lazo vectorial para el seguimiento de la trayectoria en C es:

$$\vec{r_2} + \vec{r_{32}} + \vec{r_{cx}} + \vec{r_{cy}} = C(X_r, Y_r). \tag{25}$$

Una de las diferencias entre las síntesis de mecanismos de 4 y 6 barras, es que el incremento en el número de eslabones, representa también un aumento de ecuaciones.

4. Optimización de mecanismos

En [17] se menciona que el grado de similitud entre el comportamiento de una rodilla humana y la de un mecanismo de rodilla de 4 barras, es directamente proporcional al valor del error de la función objetivo, es decir, mientras más pequeño sea el error, el CIR del mecanismo de 4 barras se iguala o aproxima al CIR anatómico de la pierna humana, lo que garantiza la reproducción del movimiento natural de la rodilla protésica.

Desde esta perspectiva, se toma una trayectoria de referencia que el punto de operación C seguirá, generada por una poloide característica ideal (ver Figura 3b) [10]. Por lo tanto la función objetivo se define como:

$$F(x) = \sum_{i=1}^{N} \left[(X_d^i - X_c^i)^2 + (Y_d^i - Y_c^i)^2 \right], \tag{26}$$

donde el primer término define el error de posición como la diferencia al cuadrado de las distancias euclidianas entre X_{deseada} y $X_{\text{calculada}}$ del punto de seguimiento C. El segundo término lleva a cabo el mismo cálculo, pero para las distancias en Y. N representa al número de puntos a sintetizar. Para los dos casos de estudio, el error se calcula con la Ecuación (26). En el primero se presenta una línea recta (ver Figura 2) estudiada en [2]. Los autores proponen la solución del problema a través de AG mediante un mecanismo de 4 barras.

En la Tabla 3 se muestra un resumen de los mecanismos óptimos encontrados y los valores correspondientes a sus errores. Es importante resaltar que para llegar a estos resultados fue necesario ajustar los parámetros de número de individuos (NI), probabilidad de cruce (PC), probabilidad de mutación (PM), factor de escalamiento (F) y número máximo de iteraciones (itermax) de los algoritmos. En el caso dos, se lleva a cabo el análisis de una trayectoria de referencia para el seguimiento de una poloide característica (ver Figura 4).

Caso 1: Línea recta.

a) El problema es definido por los siguientes parámetros [2]:
   - Límites de las variables: r_1, r_2, r_3, r_4, ..., $r_n \in [0, 60]$; r_{cx}, r_{cy}, x_0, $y_0 \in [-60, 60]$; θ_0, θ_2^1, ..., $\theta_2^6 \in [0, 2\pi]$.
   - Puntos deseados: C_d^i =[(20, 20), (20, 25), (20, 30), (20, 35), (20, 40), (20, 45)].
   - Parámetros del algoritmo: $NI = 200$, $PC = 0{,}4$, $PM = 0{,}6$, $F = 0{,}4$, `itermax`= 1000.

b) Las dimensiones obtenidas en la última iteración para el mecanismo propuesto son:

$r_1 = 56{,}3332$, $r_2 = 11{,}3609$, $r_3 = 25{,}9128$, $r_4 = 52{,}1372$, $r_{cx} = 26{,}3795$,
$r_{cy} = 14{,}5755$, $x_0 = 4{,}9398$, $y_0 = 57{,}5743$, $\theta_0 = 57{,}2957$, error $= 0{,}0349$.

Tabla 2. Solución de la optimización del caso 1: Curvas trazadas por el punto C del acoplador.

4 Barras	4 Barras	Watt	Watt	Stephenson	Stephenson
					
a) AG	b) AED	c) AG	d) AED	e) AG	f) AED

Tabla 3. Comparación de errores para trayectoria recta.

Configuración	AG	AED
4 barras	1,6525	0,0349
Watt II	2,3244	1,1771
Stephenson III	2,9565	2,5496

Resultados sujetos a: NI=1000, PC=0.6, PM=0.4, F=0.4, `itermax`=1000.

En la Tabla 2 se observan los resultados gráficos mostrados en la Tabla 3 correspondientes a las 3 configuraciones de mecanismos optimizados con las técnicas de AG y AED, se aprecia que el error mínimo entre la trayectoria deseada y la generada fue la del mecanismo de 4 barras utilizando el AED con una diferencia de 0.00873 entre el error obtenido en [2] y el mejor mecanismo encontrado (ver Tabla 3). En la Tabla 2, la línea de color roja es la trayectoria deseada o de referencia y la de color azul la generada por el mecanismo propuesto. Los resultados se clasifican como:

Caso 2: Trayectoria de referencia para obtener poloide característica.

a) El problema es definido por los siguientes parámetros [15]:

- Límites de las variables: r_1, r_2, r_3, r_4, ..., $r_n \in [0, 60]$; $r_{cx} \in [10, 25]$; $r_{cy} \in [10, 1000]$; $x_0 = 0{,}0001$, $y_0 = 500$; θ_0, θ_2^1, ..., $\theta_2^6 \in [0, 2\pi]$.
- Puntos deseados: C_d^i =[(-0.66, 607.29), (2.73, 607.76), (5.51, 608.00), (8.10, 608.10), (12.39, 608.02), (15.97, 607.70), (18.86, 607.29), (21.58, 606.75), (24.08, 606.14), (27.42, 605.10), (30.65, 603.85), (33.25, 602.62), (36.51, 600.72), (38.80, 599.03), (41.36, 596.38), (42.53, 592.95), (41.95, 591.54)].
- Parámetros del algoritmo: $NI = 200$, $PC = 0{,}4$, $PM = 0{,}6$, $F = 0{,}4$, `itermax`=1000.

b) El mejor mecanismo encontrado en la última iteración es:

$r_1 = 58{,}7478$, $r_2 = 89{,}5333$, $r_3 = 35{,}4889$, $r_4 = 78{,}9177$, $r_{cx} = 13{,}7750$,
$rr_{cy} = 13{,}7160$, $x_0 = 0{,}0001$, $y_0 = 500$, $\theta_0 = 29{,}93$, error $= 0{,}0480$.

Resultados sujetos a: NI=1000, PC=0.6, PM=0.4, F=0.4, `itermax`=1000.

Tabla 4. Solución de la optimización del caso 2: Curvas trazadas por el punto C del acoplador.

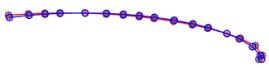
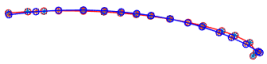
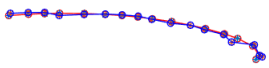
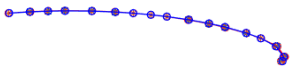
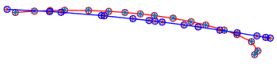
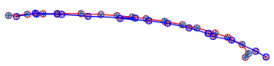
4 Barras	Watt II	Stephenson III
		
a) AG	c) AG	e) AG
		
b) AED	d) AED	f) AED

Tabla 5. Comparación de errores para trayectoria de referencia.

Configuración	AG	AED
4 barras	1,8846	0,0480
Watt II	2,5019	10,3685
Stephenson III	3,7626	3,6382

En la Tabla 4 se observan los resultados gráficos presentados en la Tabla 5, para los mecanismos óptimos obtenidos. La configuración que sigue mejor la trayectoria deseada es la de 4 barras utilizando el AED (ver Tabla 5b) con un error de 0.0480.

5. Discusión y análisis de resultados

Con base en los resultados de la Tabla 4 y considerando que por peso, precisión en el seguimiento de trayectoria, facilidad de análisis y diseño, se eligió el mecanismo de 4 eslabones para el desarrollo del dispositivo protésico (ver

Figuras 5a y 5b). Este se halló después de 996 generaciones una vez que las condiciones de paro del algoritmo se cumplieron.

El diseño solo es analizado en el plano sagital para el seguimiento de la trayectoria. La primera etapa, como se mostró en el desarrollo, consiste en obtener las dimensiones de los mecanismos que deben seguir una trayectoria establecida, que para este caso es la poloide.

El propósito de utilizar técnicas de optimización en la síntesis de los mecanismos es obtener las dimensiones y los ángulos necesarios sin tener mayor información que las restricciones de los valores máximos y mínimos. Estas técnicas se utilizan a menudo, ya que al tratarse de dispositivos personalizados es importante que no se rebasen las dimensiones o el peso necesario para evitar problemas en otras articulaciones como la cadera, rodilla o el tobillo del miembro inferior sano.

En este trabajo se presentaron solo como restricciones de construcción los ángulos y puntos de precisión para construir una poloide tomados del trabajo de [15] en combinación con una propuesta de los autores.

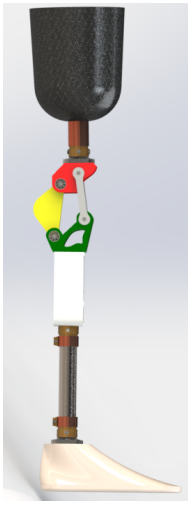
Tabla 6. Comparación de resultados entre métodos de optimización.

Variables (r_1 a r_{cy} en [mm])	Optimización cinemática [15]	AG	AED
r_1	60	59,1887	58.7478
r_2	90	87,1622	89.5333
r_3	33	32,6472	35.4889
r_4	80	78,0200	78.9177
x_0	0	0,0001	0.0001
y_0	500	500	500
r_{cx}	25,85	12,8095	13.7750
r_{cy}	14,97	14,1512	13.7160
θ_0	15	29,90	29,93
Error	0,0810	1,8846	0.0480

Para la segunda etapa del diseño del prototipo, se considerarán factores como dimensiones del usuario, peso, grado de amputación, selección de materiales de construcción, alineación de tobillo, rodilla y cadera, entre otros. En la Tabla 6 se exponen, a modo de comparación, los resultados obtenidos a través de un método clásicos (optimización cinemática) y dos de cómputo evolutivo.

Es evidente las variaciones dimensionales entre las variables encontradas mediante cada una de las técnicas. Sin embargo, lo más importante es notar el valor del error para cada uno de los casos. Es concluyente que el AED ofrece

los mejores resultados con un valor de error de 0.0480 en comparación con el método clásico e incluso con el AG.



(a)

(b)

Fig. 5. a) Mecanismo de 4 eslabones, b) Vista lateral de mecanismo propuesto, c) Vista isométrica de mecanismo propuesto.

6. Conclusiones

Este trabajo ha demostrado la factibilidad en la aplicación de los métodos de cómputo evolutivo para la síntesis de mecanismos planos basados en los Algoritmos Genéticos y Evolutivo Diferencial. Los casos de estudio presentados prueban el alcance y la capacidad de estas técnicas para la solución de diseños en el desarrollo de prótesis policéntricas para miembro inferior, basados en mecanismos de cuatro y seis barras.

Los resultados muestran que, con un rango de 6 a 17 puntos de precisión, es posible obtener el dimensionamiento de un mecanismo que cumpla con el seguimiento de una trayectoria, para este caso, la poloide. No obstante, se deben llevar a cabo ciertos ajustes en los parámetros de los algoritmos como el número de individuos, selección, cruce y mutación para encontrar el resultado óptimo de la función objetivo.

Se observa que tanto el mecanismo de cuatro como el de seis barras cumplen con el seguimiento de una poloide característica, sin embargo, debido a los requerimientos del usuario en cuanto a la necesidad de la disminución del peso del prototipo protésico, así como la obtención de una marcha normal, estabilidad

en la fase de apoyo y control voluntario de los movimientos, se determina que el uso de un mecanismo de cuatro barras, obtenido por AED, es el óptimo.

Se decidió tomar el resultado del AED, ya que el error en el seguimiento de la trayectoria fue menor en comparación con los resultados hallados con el AG y el método clásico empleado en [15]. Además, se verificó que la disposición geométrica de las barras del mecanismo generó una poloide en un 92 % similar a la establecida para una rodilla sana en posiciones particulares de la marcha y de la sedestación.

Referencias

1. Cabrera, J. A., Ortiz, A., Nadal, F., Castillo, J. J.: An evolutionary algorithm for path synthesis of mechanisms. Mechanism and Machine Theory, vol. 46, pp. 127–141 (2011)
2. Cabrera, J. A., Simon, E., Prado, M.: Optimal synthesis of mechanisms with genetic algorithms. Mechanism and Machine Theory, vol. 37, no. 10, pp. 1165–1177 (2002)
3. Cuevas, E. V., Osuna, J. V., Oliva, D. A., Díaz, M. A.: Optimización: Algoritmos programados con MATLAB. Alfaomega, Primera Edición (2016)
4. Ebrahimi, S., Payvandy, P.: Efficient constrained synthesis of path generating four-bar mechanisms based on the heuristic optimization algorithms. Mechanism and Machine Theory, vol. 85, pp. 189–204 (2015)
5. Erdman, A. G., Sandor, G. N.: Diseño de mecanismos análisis y síntesis . Pearson (1998)
6. Lugo, E.: Diseño de mecanismos utilizando algoritmos genéticos con aplicación en prótesis para miembro inferior. Tesis Doctoral, Instituto Politécnico Nacional (2010)
7. Morales, C., Portilla, E. A., Suárez, R. A., Hernández, N., Calva, M. B.: Synthesis of a Non-Grashof Six-Bar Polycentric Knee Prostheses Using an Evolutionary Optimization Algorithm. Conference on Engineering Optimization, pp. 1121–1132 (2019)
8. Norton, R. L.: Diseño de Maquinaria, síntesis y análisis de máquinas y mecanismos. Mc Graw Hill (2013)
9. Phukaokaew, W., Sleesongsom, S., Panagant, N., Bureerat, S.: Synthesis of four-bar linkage motion generation using optimization algorithms. Advances in Computational Design, vol. 4, no. 3, pp. 197–210 (2019)
10. Poliakov, O., Chepenyuk, O., Pashkov, Y., Kalinin, M., Kramar, V.: Multicriteria synthesis of a polycentric knee prosthesis for transfemoral amputees. International Journal of Chemical and Biological Engineering, vol. 6, no. 5, pp. 257–262 (2012)
11. Ponce, P., Molina, A., Ramírez, R., Mendez, E., Ortiz, A. A.: A practical approach to metaheuristics using labview and Matlab. CRC Press (2020)
12. Ponce, P.: Inteligencia Artificial con aplicación a la ingeniería. Alfaomega (2010)
13. Pérez, R.: Análisis de mecanismos y problemas resueltos. Alfaomega (2006)
14. Radcliffe, C. W.: Four-bar linkage prosthetic knee mechanisms: Kinematics, alignment and prescription criteria. Prosthetics and Orthotics International, vol. 18, pp. 159–173 (1994)
15. Salas, P., Vergara, M., Provenzano, S.: Prótesis de rodilla: Fundamentos teóricos y técnicas computacionales para su diseño. Revista Ciencia e Ingeniería, vol. 42, no. 1, pp. 91–100 (2021)

16. Shiakolas, P. S., Koladiya, D., Kebrle, J.: On the optimum synthesis of six-bar linkagesusing differential evolution and the geometric centroid of precision positions technique. Mechanism and Machine Theory, vol. 40, pp. 319–335 (2005)
17. Xie, H., Wang, S., Li, F.: Knee joint optimization design of intelligent bionic leg based on genetic algorithm. International Journal Bioautomation, vol. 18, no. 3, pp. 195–206 (2014)

ISSN 1870-4069

Redes neuronales recurrentes para el desarrollo de las habilidades conversacionales de un asistente de aprendizaje

Erik Carbajal-Degante[1], Omar Terrazas Razo[1], Jackeline Bucio García[1], Jimena Olveres[2], Boris Escalante-Ramírez[2], Guadalupe Vadillo[1]

[1] Universidad Nacional Autónoma de México,
Coord. de Universidad Abierta, Innovación Educativa y Educación a Distancia,
México

[2] Centro de Estudios en Computación Avanzada,
Universidad Nacional Autónoma de México,
México

{erikycd, guadalupe.vadillo}@gmail.com

Resumen. Este trabajo presenta los resultados de un procedimiento para desarrollar las habilidades conversacionales de un asistente de aprendizaje a través del diseño e implementación con algunas bases de datos. Se utilizan técnicas tradicionales de procesamiento de lenguaje natural así como una arquitectura conocida como *seq2seq* de redes neuronales recurrentes para la generación del lenguaje. Los resultados alcanzados proporcionan un indicio de buen rendimiento y son complementados con experimentos subjetivos. Esta propuesta constituye un elemento importante de interacción humano-máquina de los sistemas de tutoría inteligente que enriquece la posibilidad de personalizar el entorno del estudiante además de apoyarle en su formación.

Palabras clave: Asistente de aprendizaje, chatbot, redes neuronales recurrentes, sistemas de tutoría inteligente.

Recurrent Neural Networks for the Development of Conversational Skills of a Learning Assistant

Abstract. In this work, we highlight the importance of the written conversational skills of learning assistants and address the design and implementation with some databases. Natural language processing techniques in combination with a deep neural network architecture known as *seq2seq* is used to generate language. Results provide a clue of good performance that are complemented with subjective assessments. This proposal represents an important element of the human-machine interaction of intelligent tutoring systems that fosters the learners' environment customization and supports the learning process.

Keywords: Learning assistant, chatbot, recurrent neural networks, intelligent tutoring systems.

1. Introducción

La inclusión de la inteligencia artificial (IA) en la educación es cada vez mayor. En una revisión de la literatura de 2009 a 2021, [3] señalan que la investigación publicada sobre sistemas de tutoría inteligente (STI) ocupa uno de los primeros lugares y que la frecuencia y cantidad de publicaciones relativas a ella manifiestan una tendencia de crecimiento en el periodo analizado.

Los STI se conciben, de acuerdo con Ubani y Nielsen [20], retomando los lineamientos de Graesser y su equipo, como entornos de aprendizaje dentro de sistemas de cómputo que aportan a los aprendices modelos personalizados de retroalimentación o instrucciones a partir de modelos computacionales.

Permiten dar seguimiento detallado a diversos estados psicológicos (como emociones, habilidades o nivel de conocimiento) y adaptarse a los resultados para potenciar los aprendizajes de cada estudiante. Requieren un componente de interacción con el usuario y resulta deseable que se comuniquen de la forma más humana posible, por lo que se busca que desarrollen habilidades conversacionales. En la siguiente sección se detallan los tipos de componentes disponibles.

1.1. Chatbots, agentes y asistentes

Se conoce como chatbot a una herramienta de software que interactúa con los usuarios sobre un tema determinado o en un dominio específico de forma natural y conversacional utilizando texto o voz [4]. Para muchos propósitos diferentes, los chatbots se han utilizado en una amplia gama de dominios los cuales incluyen marketing, servicio a cliente, soporte técnico así como educación y capacitación. Se estima que el 70 % de las empresas de tecnología poseerán la capacidad de diseñar su propio sistema de conversación digital a mediados del 2022 [5].

En este sentido, la mayoría de los chatbots serán construidos con el objetivo de brindar una mejor experiencia al usuario, proveer servicios especializados, facilitar y automatizar ciertos procesos, y reducir en gran medida el costo de la interacción humana así como proporcionar disponibilidad y simplicidad de uso, por lo que se puede hablar de una transición generacional de estas entidades con funciones avanzadas donde la inteligencia artificial juega un papel muy importante.

Chatbots. En su forma más simple, los chatbots suelen seguir un conjunto de reglas o flujos establecidos para contestar a las preguntas realizadas por el usuario. Estas reglas o flujos les permiten responder de manera efectiva a las solicitudes dentro de un dominio específico de diálogo, pero no son eficientes para responder preguntas cuyo patrón no coincide con las reglas para el que está capacitado dado que el entrenamiento utilizado por estos sistemas no es tan elevado.

Agentes virtuales. Por otro lado, los agentes virtuales son entidades más avanzadas que los chatbots pero que igualmente utilizan el procesamiento de lenguaje natural (PLN) como forma de interacción, así mismo hacen uso de técnicas relacionadas con el entendimiento de lenguaje natural (ELN) y generación de lenguaje natural (GLN). Para estas entidades, los datos son cruciales puesto que sus habilidades son resultado de utilizar modelos entrenables del aprendizaje automático y del aprendizaje profundo, por lo que se dice que los agentes virtuales poseen cierto nivel de inteligencia (ver Figura 1 para mayor detalle comparativo).

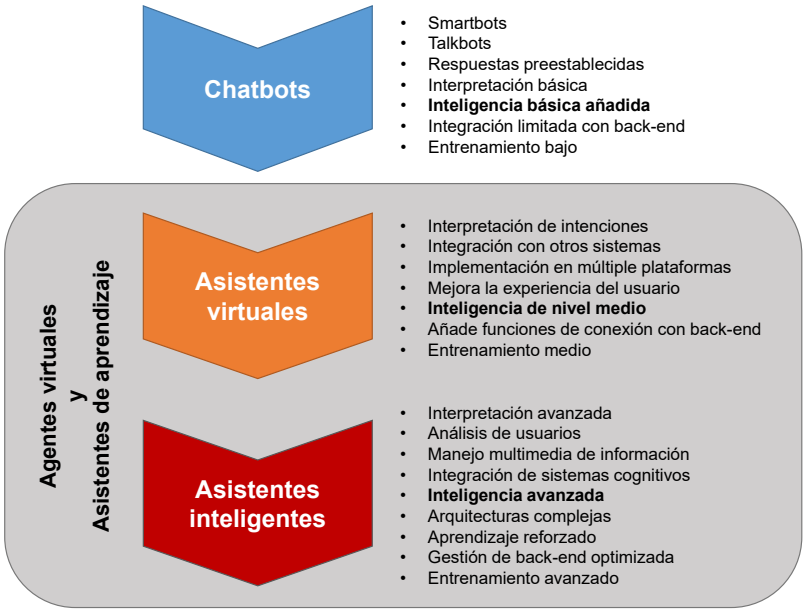


Fig. 1. Transición generacional de los chatbots y agentes virtuales.

Asistentes. Es posible considerar a los asistentes virtuales como parte de los agentes virtuales con un nivel de inteligencia medio. Las habilidades de los asistentes virtuales son mayores a las de los chatbots dado que utilizan un sistema de interpretación de intenciones. Con esto logran abordar diferentes tareas dentro de un abanico más amplio de opciones. Poseen la capacidad de integrarse con otros servicios y sistemas o correr en múltiples plataformas; en consecuencia mejoran considerablemente la experiencia al usuario.

Dentro de los asistentes virtuales, las entidades más sofisticadas son los asistentes inteligentes. Se busca crear asistentes con funciones y características más complejas que permitan interactuar de manera fluida con el usuario, como lo haría un ser humano. Incorporan capacidades de inteligencia mayor como interpretación y generación de lenguaje, análisis de sentimientos y son

capaces de definir un perfil del usuario y adaptarse a el, lo que se conoce como servicios cognitivos [2].

En este sentido los asistentes inteligentes son las entidades más difíciles de diseñar y su arquitectura roza la vanguardia tecnológica de desarrollo en IA, la cual sigue creciendo a pasos agigantados. El uso de datos masivos y limpios es un requisito importante dado que la capacidad de aprender por sus medios (aprendizaje por refuerzo) es una característica distintiva.

1.2. Asistentes de aprendizaje

En los últimos años, cada vez más organizaciones han comenzado a explotar los beneficios de IA más allá de la simple consulta de información seguida de una respuesta programada.

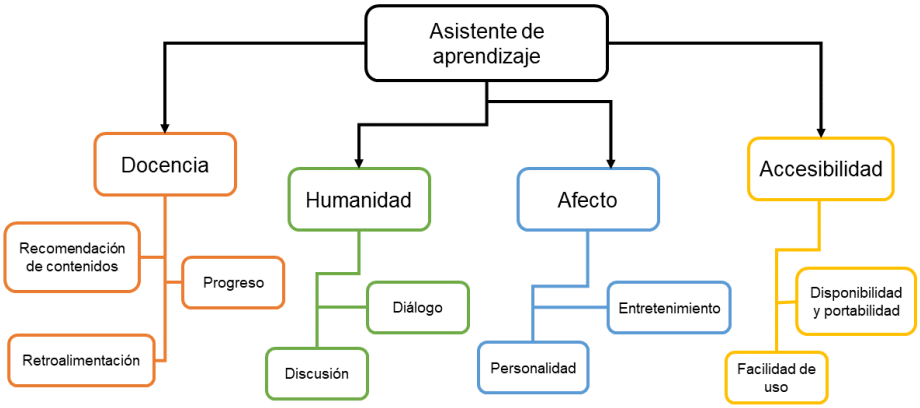


Fig. 2. Atributos y habilidades de un asistente de aprendizaje.

La incorporación de los chatbots y agentes al área educativa durante lo últimos años implica un aumento en el interés por las formas en que estas entidades podrían implementarse para la enseñanza y el aprendizaje, lo que da lugar a la creación de asistentes de aprendizaje [19].

Los asistentes de aprendizaje pueden considerarse como parte de los agentes virtuales. Se enfocan en mejorar y personalizar la automatización en la enseñanza a través de la IA. El conocimiento de los modelos de IA es importante para desarrollar agentes pedagógicos útiles, interesantes y valiosos que no solo aprovechen al máximo los más recientes avances científicos, sino también identifiquen en cierta forma las preocupaciones emocionales, cognitivas y de educación, por lo que su diseño se convierte en un reto importante en la actualidad.

Los asistentes de aprendizaje pueden proporcionar los beneficios de la disponibilidad instantánea y la capacidad de responder de forma natural a través de una interfaz conversacional con ventajas similares a las de entablar un diálogo

con un profesor. Estos asistentes demuestran la capacidad de crear interacciones sencillas con los usuarios además de respaldar la participación, establecer objetivos, proponer estrategias y analizar los resultados de aprendizaje o capacitación [16].

Dentro de los métodos para evaluar la calidad de los asistentes de aprendizaje se encuentran los basados en el proceso jerárquico analítico [14], cuyo enfoque es el adecuado para resolver problemas de toma de decisiones multicriterio. Bajo este esquema y a través de múltiples estudios realizados en [11], se concluyó que los atributos sobresalientes para la medición de la calidad incluyen la eficacia (funcionalidad y humanidad), la eficiencia (rendimiento) y la satisfacción (accesibilidad, afecto, comportamiento y ética).

En [16] se sigue la misma línea de análisis obteniendo cuatro categorías principales: enseñanza, humanidad, afecto y accesibilidad. Finalmente, estos atributos son considerados en el desarrollo de asistentes de aprendizaje más complejos y avanzados, por lo que su diseño no debe omitir estos criterios de calidad. En la Figura 2 se observan las funcionalidades ligadas a los cuatro atributos de un asistente de aprendizaje.

El estudio realizado por [16] para diversos asistentes pedagógicos utilizados en la plataforma de Facebook messenger en diferentes idiomas indica que de los atributos existentes, humanidad es uno de los más relevantes ya que es el punto de partida de interacción con los usuarios y se encuentra íntimamente ligado al tipo de personalidad, por lo que las habilidades conversacionales de un asistente juegan un papel importante en los primeros pasos en el diseño e implementación.

2. Métodos y materiales

Esta sección detalla algunos conceptos útiles relacionados al procesamiento del lenguaje que son la base del desarrollo de aplicaciones que involucran texto. También se detallan varias de las técnicas comunes de pre-procesamiento de información que transforman los datos para su aprovechamiento por los sistemas computacionales.

2.1. Procesamiento del lenguaje

El procesamiento de lenguaje natural (NLP, por sus siglas en inglés) es el campo de estudio de la IA enfocado en la comprensión, manejo y generación de lenguaje humano por medio de algoritmos de Machine Learning (ML) [10]. En los últimos años ha permitido incrementar el número de posibilidades dentro del ámbito académico para obtener y entregar retroalimentación en tiempo real y así mejorar la calidad de la educación mediante la aplicación de técnicas de análisis de sentimientos [**?**].

Como parte de este procesamiento se encuentra la comprensión de lenguaje natural (NLU, por sus siglas en inglés) que se enfoca en procesar entradas no estructuradas y convertirlas a un formato estructurado, tarea que resulta difícil cuando se tienen las complejidades de lenguaje como la anáfora, elisión,

ambigüedad e incertidumbre [12]. Es aquí donde las redes neuronales recurrentes (RNN, por sus siglas en inglés) juegan un papel importante. Finalmente, la generación de lenguaje natural (NLG, por sus siglas en inglés) se encarga de transformar los datos estructurados en lenguaje natural, como voz o texto.

Debido a que parte de este procesamiento de lenguaje, mientras más grande sea el modelo tendrá una mayor certeza, como la adaptación de GPT-2 en DialoGPT [13] y sus 147 millones de diálogos, mientras que Menna, el modelo de Google AI, tiene 2,600 millones de parámetros y se entrena con 341 GB de texto, lo que lo coloca con una capacidad 1.7 veces mayor que los modelos de última generación como OpenAI GPT-2 y con 8.5 veces más datos.

2.2. Técnicas de procesamiento

La limpieza de datos es el proceso que permite eliminar datos incorrectos, duplicados o corruptos de un conjunto y es sumamente útil cuando se combinan múltiples fuentes de datos. Para realizar este proceso existen diferentes métodos que integran algoritmos de ML [6]. La tokenización es el primer paso en el proceso de NLP: divide los datos no estructurados y el texto en fragmentos que se consideran elementos discretos, con lo que se tienen datos numéricos adecuados para el aprendizaje automático, por ejemplo [7].

La tokenización puede separar oraciones, palabras, caracteres o subpalabras. Lemmatization y stemming son dos métodos empleados por los chatbots para analizar el significado detrás de una palabra. Por lo general, lemmatization se realiza mediante algoritmos Tree (árboles de palabras clave) fundamentados en la programación dinámica basada en la distancia de Levenshtein y en la estructura de los datos [15] y busca la mejor palabra de origen de las que se tienen.

Por otro lado, el proceso de stemming reduce la inflexión en las palabras a sus formas de raíz. Lo que ayuda en el pre-procesamiento de texto, palabras y documentos para la normalización del texto [17]. La derivación de cada idioma es diferente y está fuertemente afectada por el tipo de idioma del texto, lo que hace que existan múltiples variantes de los algoritmos para atender cada caso.

2.3. Transición de redes secuenciales a redes recurrentes

Los humanos no iniciamos nuestro razonamiento desde cero cada segundo que pasa, por ejemplo, mientras leemos un documento de texto, entendemos lo que está escrito con base en nuestro entendimiento de las palabras anteriores, lo que construye el contexto. En este sentido, las redes neuronales tradicionales (NNs) no pueden simular esta tarea por su características secuenciales y olvidan atributos previos.

Las redes secuenciales son un tipo de red neuronal en donde cada entrada se procesa de manera independiente sin considerar los datos procesados con anterioridad. En el caso de las redes recurrentes, se utiliza información de los procesos y datos anteriores para calcular una nueva salida. Las llamadas RNNs se utilizan ampliamente para realizar análisis de secuencias ya que están diseñadas

para extraer la información contextual definiendo las dependencias entre varios pasos de tiempo. De forma general, las RNNs se caracterizan por un flujo de retroalimentación de sus estados internos.

3. Arquitectura conversacional

Los modelos secuencia a secuencia (a menudo abreviados como *seq2seq*) son una clase especial de arquitecturas de redes neuronales recurrentes compuestas a su vez de dos arquitecturas en cascada nombradas como codificador y decodificador. Normalmente, la arquitectura *seq2seq* se utiliza para resolver problemas complejos relacionados al lenguaje como traducción automática, sistemas de pregunta y respuesta, resumido de texto, reconocimiento del habla, etc. [18].

El modelo *seq2seq* más simple consta de dos redes recurrentes LSTM (Long-Short Term Memory), una para el codificador y otro para el decodificador. Básicamente, el codificador procesa el texto de entrada produciendo un estado final el cual es utilizado por la etapa del decodificador. A través de este proceso se busca que el codificador capture toda la información sobre la fuente para que el decodificador pueda generar un texto de salida en función de estos estados obtenidos.

3.1. Proceso de entrenamiento

Similar a muchos procesos de entrenamiento de redes neuronales, los modelos *seq2seq* aplicados al lenguaje se entrenan para predecir distribuciones de probabilidad de una frase o 'token' dado un contexto (conjunto de tokens previos).

Tabla 1. Hiper-parámetros del modelo.

Hiper-parámetro	**Valor**	**Hiper-parámetro**	**Valor**
Dimensión LSTM	256	Longitud de la secuencia	15
Optimizador	RMSProp	Activación	Softmax
Función de pérdida	CCE	Tamaño de batch	64
Métrica	Accuracy	Dropout	0.2
Tasa de aprendizaje	0.01	Épocas	50

A cada paso se intenta maximizar la probabilidad de asignar el token correcto reduciendo la función de pérdida (comúnmente entropía cruzada). Formalmente, se asume que tenemos una instancia de entrenamiento cuya fuente es una secuencia
$x = (x_1, \ldots, x_m)$ y otra secuencia objetivo $y = (y_1, \ldots, y_n)$, donde ambas secuencias pueden o no diferir de tamaño, el objetivo de *seq2seq* es estimar la probabilidad condicional $p^{(t)} = p(*|y_1, \ldots, y_{t-1}, x_1, \ldots, x_m)$, donde t representa

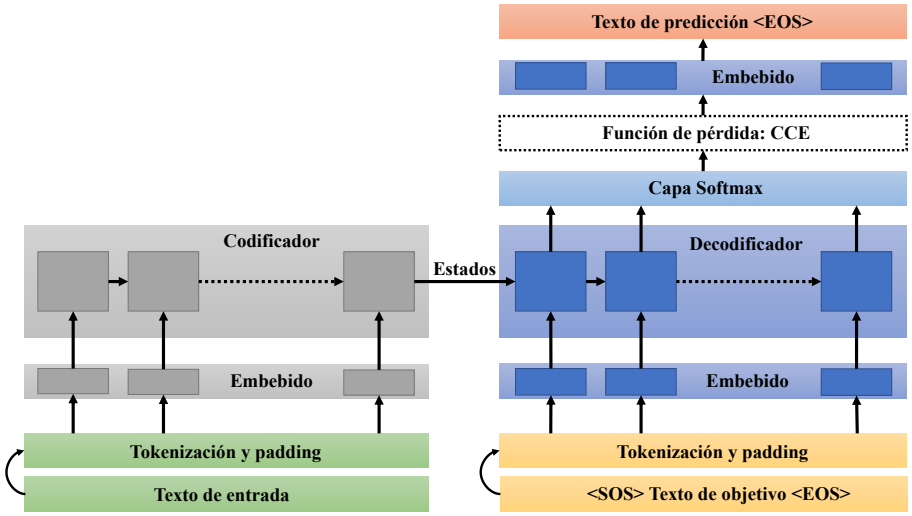


Fig. 3. Diagrama a bloques del entrenamiento *seq2seq*.

cada paso de tiempo. De forma general, el proceso completo puede modelarse como:

$$p(y_1, \ldots, y_n | x_1, \ldots, x_m) = \prod_{t=1}^{n} p(y_t | y_1, \ldots, y_{t-1}, x_1, \ldots, x_m), \tag{1}$$

donde n es la longitud de y, m es la longitud de x. En el caso práctico, dos arreglos de texto correspondientes a un diálogo de tipo pregunta-respuesta se procesan mediante técnicas de tokenización para la conversión de cada palabra, lo que produce una representación numérica de cada frase, lo que llamamos teóricamente x e y.

Para el caso del decodificador, se utilizan dos palabras reservadas `<SOS>` y `<EOS>` que indican respectivamente el inicio y el final de cada oración en la secuencia objetivo. Posteriormente, los arreglos x e y se embeden y utilizados como elemento de entrada de cada bloque LSTM, tanto del codificador, como del decodificador. Primero, el estado final resultante del codificador se utiliza como estado inicial del decodificador.

Posteriormente, el decodificador hará predicciones por cada palabra la cual se comparará con la palabra objetivo. Finalmente, mediante el cálculo del error a través de la función de pérdida, el sistema ajustará los pesos de la arquitectura a cada iteración. El proceso descrito anteriormente puede verse en la Figura 3 así como los hiper-parámetros utilizados en la Tabla 1, los cuales han sido ajustados heurísticamente.

3.2. Descripción de las bases de datos

Se ha realizado un proceso de búsqueda de datos en Internet para conseguir la información y acoplarla a la forma requerida por la arquitectura. Las características de diálogo estilo pregunta-respuesta en el idioma español no son tan comunes como se puede encontrar en otros idiomas, complicando más el hallar datos específicos con un contexto educativo. Se han realizado experimentos solo sobre tres bases de datos (BD) diferentes: dos BD son de acceso público y una fué elaborada localmente mediante técnicas de web scarping. Estas BD se detallan a continuación:

1. Los Sopranos. Extraída de Kaggle, consiste en la transcripción de los diálogos de una serie de televisión. En los experimentos descritos a continuación, se utilizan los archivos de la primera temporada con 13 episodios de 14k lineas de diálogo conteniendo 7,645 palabras. Ver la gráfica de nubes de palabras en Figura 4a.
2. OpenSubtitle. Corpus construido y estudiado en [9], es una BD masiva de 3.7M de archivos que recopila los diálogos de algunas películas en 62 idiomas diferentes. Para los experimentos llevados a cabo en este trabajo, se ocupó un extracto de 20k líneas de diálogo las cuales construyen un vocabulario de 9,680 palabras diferentes, lo que representa alrededor del 10 % del vocabulario del idioma español. Ver su gráfica de nubes de palabras en la Figura 4b.
3. Asistente. Es una BD la cual consta de una colección de datos generados localmente, así como diálogos extraídos de un conjunto público en [1]. Las oraciones han sido revisadas, limpiadas y pre-procesadas. Esta base de datos contiene 2k lineas de diálogo con un vocabulario de 1,857 palabras. Véase en la Figura 4c su gráfica de nube de palabras.

3.3. Recursos computacionales

Este trabajo se implementó en su totalidad en lenguaje Python versión 3.7. Una computadora con CPU Intel(R) Xeon(R) Silver 4216 CPU@2.10, así como una tarjeta gráfica NVIDIA GeForce RTX3090 se utilizó para entrenamiento e inferencia mediante las bibliotecas proporcionadas por Keras sobre Tensorflow versión 2.7.

4. Resultados

4.1. Proceso de inferencia

El proceso de inferencia es muy similar al entrenamiento y se define habitualmente como desarrollos independientes ya que las entradas y salidas son diferentes para ambos casos. En este proceso, la inferencia aprovechará todos los parámetros de la red aprendidos en la etapa de entrenamiento. La arquitectura

(a) (b)

(c)

Fig. 4. Nube de palabras de las bases de datos extraídas de: Los Soprano a, Opensubtitle b y propia c.

del codificador no cambia, por lo que se alimenta la misma red con una nueva frase (conjunto de tokens).

El decodificador, por su parte, se alimentará tanto del estado final del codificador, como de la palabra reservada `<SOS>` que le indicará el inicio de iteración. El resultado de la predicción de la primera palabra será el elemento de entrada del segundo bloque y así, sucesivamente hasta alcanzar la longitud máxima de la secuencia o bien la palabra reservada que indica el final: `<EOS>`. Este proceso se conoce como greedy decoding y busca encontrar la máxima probabilidad de las palabras siguiendo un procedimiento en cadena, de la forma, donde y' representa la secuencia de predicción:

$$y' = \arg\max_y p(y|x) = \arg\max_y \prod_{t=1}^{n} p(y_t|y_{<t}, x). \tag{2}$$

Otras técnicas existentes en la literatura como beam search permiten mejorar los resultados del proceso de inferencia al producir varias hipótesis y calcular la mejor combinación, aunque este procedimiento puede incrementar considerablemente el tiempo de respuesta.

4.2. Gráficas de desempeño

El desempeño de la arquitectura puede visualizarse mediante las gráficas de pérdida y precisión para cada época $\in [0, 50]$. Dichas gráficas se muestran en la Figura 5 para las tres BD estudiadas. Se puede notar en la Figura 5(a) que la pérdida disminuye a cada época para todas las BD, lo que da un indicio de convergencia del modelo hacia los datos de entrenamiento. Se puede ver que los

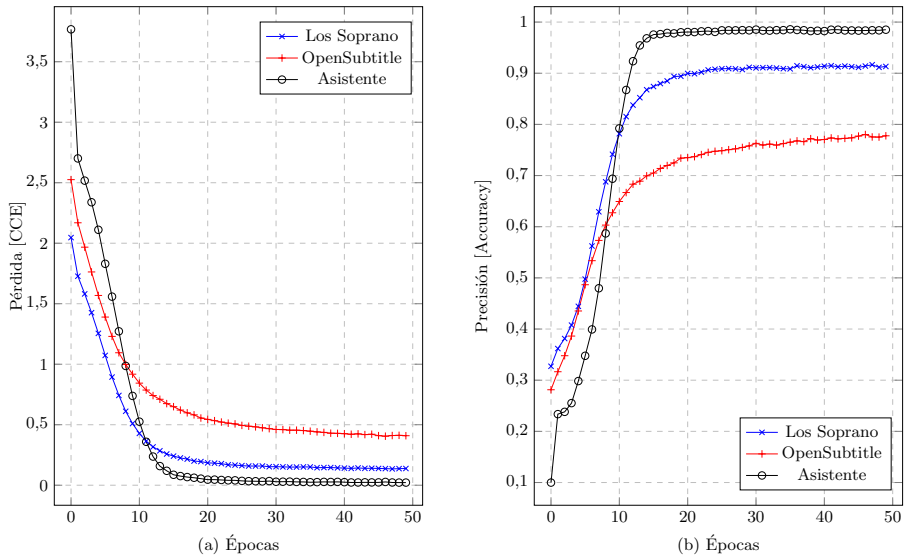


Fig. 5. Curvas de pérdida medida con CCE en (a), curvas de accuracy en (b).

valores mínimos de pérdida se alcanzan alrededor de la época 30, aunque para algunas BD la convergencia es más rápida debido a la cantidad de información utilizada y la limpieza de los datos.

Por otro lado, la métrica utilizada en términos de accuracy, ver Figura 5(b), indica la precisión del ajuste del modelo hacia los datos entrenados. Tanto las funciones de pérdida y la métrica de precisión funcionan de diferente manera cuando se analizan distintos datos (series de tiempo, texto, imágenes, video), por lo que posiblemente una función de pérdida más adecuada brindará mejor información del ajuste realizado sobre texto.

4.3. Comparativa de diálogo

El trabajo de [8] presenta un modelo de diálogo basado en *seq2seq* con el que se simulan diferentes comportamientos de personalidad. Sus resultados se comparan subjetivamente con el fin de medir la calidad de las respuestas y la cantidad de inconsistencias como complemento de las evaluaciones objetivas. Hemos extraído un conjunto de esas preguntas para estimar el desempeño subjetivamente del modelo con el entrenamiento de las 3 BD. En la Tabla 2 se muestran las respuestas obtenidas a las preguntas base que se han realizado, donde se pueden notar las inconsistencias entre algunas de las respuestas y la tendencia del modelo a adquirir una personalidad dada la información con la que ha sido entrenado.

Tabla 2. Tabla comparativa de las respuestas a un diálogo.

Pregunta	Modelo	Respuesta
Hola	Los Soprano	Hola buscaba a tony soprano
	OpenSubtitle	Hola que tal estais
	Asistente	Hola estas de vuelta
¿Donde vives?	Los Soprano	A que se refiere
	OpenSubtitle	Hay alguien mejor que podemos hablar de los insectos
	Asistente	Vivo en una computadora
¿Cuantos años tienes?	Los Soprano	Feliz cumpleaños
	OpenSubtitle	Como es tu mejor
	Asistente	Puedo vivir tanto como mis padres quieran
¿Que estudias?	Los Soprano	No se puedo hacerlo
	OpenSubtitle	Me gusta
	Asistente	Simplemente no puedo salir de la habitacion
¿Tienes sentimientos?	Los Soprano	En serio
	OpenSubtitle	Siempre asi que claro
	Asistente	No tengo sentimientos en este momento soy un chatbot

4.4. Limitaciones

Básicamente, existen dos inconvenientes principales de la arquitectura *seq2seq*, ambos relacionados con la longitud de las secuencias procesadas. En primer lugar, se habla de una memoria limitada en donde frases más complejas o con un contexto mayor de entendimiento pueden insertar errores por la pérdida de información ya sea al forzar un número máximo de caracteres (operaciones de truncamiento y compresión), o bien los estados finales del codificador son insuficientes para alimentar el decodificador.

Segundo, por regla general, cuanto más profunda es una red neuronal, mayor incertidumbre de interpretabilidad. En el caso de redes neuronales recurrentes, a profundidades grandes y secuencias largas de texto los gradientes desaparecen al realizar el proceso de retro-propagación. Aunque en teoría las redes recurrentes ayudan a prevenir en cierta medida este problema, sigue estando presente dicha limitante.

5. Discusión: El reto de la personalidad coherente

A pesar de la tecnología disponible en PLN e IA, la cual ha demostrado alcanzar resultados muy cercanos e incluso rebasar a los que pueden lograrse mediante la interacción humana, existen algunas limitaciones debido a la naturaleza del razonamiento humano que para las máquinas les es imposible simular. Tales limitaciones emergen de la falta de conciencia e inteligencia emocional, confusión por complejidad de las intenciones o del contexto, así como la necesidad de un continuo entrenamiento y una base de datos robusta.

Tal es el caso que aborda este trabajo al exhibir una parte importante de las limitantes producidas al entrenar modelos con datos escasos y que muestran características de diálogo específicas, con lo que una tendencia de personalidad particular se hace presente. Lo ideal es que los chatbots y agentes produzcan respuestas coherentes a las entradas, esto parece simple pero incorporar un conocimiento establecido para crear una personalidad es uno de los mayores retos en investigación.

La personalidad de estas entidades se refiere básicamente al personaje que interpreta durante una interacción conversacional. Adoptar características como edad, género, idioma, forma de expresión, nivel de conocimiento o especialidad pueden resaltar mejor su carácter y con eso facilitar la comunicación con el usuario estableciendo vínculos de confianza. Como lo resalta el trabajo de [21] que sugiere el uso de agentes con memoria que puedan simular interés en el perfil del usuario al realizar preguntas básicas dentro de un escenario de conversación natural.

Muchos sistemas aprenden a generar respuestas lingüísticas apropiadas pero no están entrenados para generar respuestas semánticamente consistentes. A tales sistemas les es difícil asignar un perfil explícito para generar respuestas coherentes y una de las raíces de este problema se ubica en el tipo de datos con los que se entrenan algunos modelos. La necesidad de datos masivos requeridos por técnicas del aprendizaje profundo es un factor clave para su buen desempeño.

Sin embargo, un conjunto de datos regularmente consiste en una compilación de información de diferentes fuentes. Abordar el problema de escasez de datos limpios y un proceso particular de filtrado es uno de los primeros pasos en la construcción de modelos robustos que adquieran las características de personalidad que la aplicación necesite y que brinden de cierta forma sentido humano a una conversación fluida.

6. Conclusiones y trabajos a futuro

El presente trabajo aborda los fundamentos en el desarrollo e implementación de las habilidades conversacionales de un asistente de aprendizaje, partiendo de la necesidad actual de enfocar la tecnología existente en el sector educativo y de proponer herramientas pedagógicas robustas que se encuentren a la vanguardia. Describimos el proceso de construcción de un sistema basado en el aprendizaje profundo y técnicas del procesamiento de lenguaje natural que le permite entrenarse con tres diferentes bases de datos y formar oraciones en el idioma español en respuesta a una frase de entrada.

Los resultados indican la convergencia del modelo en épocas tempranas lo que da un indicio de rendimiento, aún teniendo bases de datos masivas con un vocabulario extenso. Los resultados subjetivos nos permiten abordar el tema de personalidad coherente que en la actualidad representa un reto importante para cualquier sistema conversacional. Mecanismos de atención o redes más complejas (arquitecturas de tipo Transformers) pueden ayudar a superar las limitaciones que los modelos *seq2seq* presentan por naturaleza.

Se sabe que la capacidad de un asistente de aprendizaje de manejarse dentro de un ambiente con temas simples de charla abre el panorama adecuado para la formación de vínculos de confianza con los estudiantes. De esta manera, se obtiene información útil que construya nuevas bases de datos limpias y completas.

Esto permite también recopilar la información suficiente para realizar tareas futuras relacionadas al analítica del aprendizaje y complementar las funciones empleadas por los sistemas actuales de tutoría inteligente.

Agradecimientos. Este proyecto recibió financiamiento de Santander Universia.

Referencias

1. Alblawi, A. S.: Big data and learning analytics in higher education: Demystifying variety, acquisition, storage, NLP and analytics. arXiv, (2018) doi: 10.1109/icbdaa.2017.8284118
2. Benke, I., Gnewuch, U., Maedche, A.: Understanding the impact of control levels over emotion-aware chatbots. Computers in Human Behavior, vol. 129, pp. 107122 (2022) doi: 10.1016/j.chb.2021.107122
3. Chen, X., Zou, D., Xie, H., Cheng, G., Liu, C.: Two decades of artificial intelligence in education: Contributors, collaborations, research topics, challenges, and future directions. Educational Technology and Society, vol. 25, no. 1, pp. 28–47 (2022)
4. Dale, R.: The return of the chatbots. Natural Language Engineering, vol. 22, no. 5, pp. 811–817 (2016) doi: 10.1017/s1351324916000243
5. Gartner: Chatbots will appeal to modern workers (2019)
6. Hirsch, T., Hofer, B.: Identifying non-natural language artifacts in bug reports. In: 2021 36th IEEE/ACM International Conference on Automated Software Engineering Workshops (ASEW). Ieee (2021) doi: 10.1109/asew52652.2021.00046
7. Li, B., Zhang, Y., Sainath, T., Wu, Y., Chan, W.: Bytes are all you need: End-to-end multilingual speech recognition and synthesis with bytes. arXiv, (2018)
8. Li, J., Galley, M., Brockett, C., Spithourakis, G. P., Gao, J., Dolan, B.: A persona-based neural conversation model. arXiv, (2016)
9. Lison, P., Tiedemann, J.: OpenSubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In: Lrec (2016)
10. Madhuri, D., Prasad, R.: A ML and NLP based framework for sentiment analysis on bigdata. International Journal of Recent Technology and Engineering, vol. 8, no. 5, pp. 189–200 (2020) doi: 10.35940/ijitee.d9062.029420
11. Radziwill, N. M., Benton, M. C.: Evaluating quality of chatbots and intelligent conversational agents (2017)
12. Ram, A., Prasad, R., Khatri, C., Venkatesh, A., Gabriel, R., Liu, Q., Nunn, J., Hedayatnia, B., Cheng, M., Nagar, A., King, E., Bland, K., Wartick, A., Pan, Y., Song, H., Jayadevan, S., Hwang, G., Pettigrue, A.: Conversational ai: The science behind the alexa prize. arXiv, (2018) doi: 10.48550/arxiv.1801.03604
13. Rastogi, A., Zang, X., Sunkara, S., Gupta, R., Khaitan, P.: Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. arXiv, (2019)
14. Saaty, T. L.: What is the analytic hierarchy process? Mathematical Models for Decision Support, pp. 109–121 (1988) doi: 10.1007/978-3-642-83555-1_5

15. Schmitt, M., Constant, M.: Neural lemmatization of multiword expressions. In: Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019). Association for Computational Linguistics (2019) doi: 10.18653/v1/w19-5117
16. Smutny, P., Schreiberova, P.: Chatbots for learning: A review of educational chatbots for the facebook messenger. Computers & Education, vol. 151, pp. 103862 (2020) doi: 10.1016/j.compedu.2020.103862
17. Suci, F. W., Hayatin, N., Munarko, Y.: In-idris: Modification of idris stemming algorithm for indonesian text. IIUM, (2022)
18. Sutskever, I., Vinyals, O., Le, Q. V.: Sequence to sequence learning with neural networks. arXiv, (2014)
19. Tamayo, P. A., Herrero, A., Martín, J., Navarro, C., Tránchez, J. M.: Design of a chatbot as a distance learning assistant. Open Praxis, vol. 12, no. 1, pp. 145 (2020) doi: 10.5944/openpraxis.12.1.1063
20. Ubani, S., Nielsen, R.: Review of collaborative intelligent tutoring systems (CITS) 2009-2021. In: 2022 11th International Conference on Educational and Information Technology (ICEIT). Ieee (2022) doi: 10.1109/iceit54416.2022.9690733
21. Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D., Weston, J.: Personalizing dialogue agents: I have a dog, do you have pets too? arXiv, (2018)

ISSN 1870-4069

Caracterización de eventos volcánicos explosivos a partir de las señales sísmicas del volcán de Colima, México, para la identificación de nivel de peligro volcánico

Félix Ortigosa[1], Vyacheslav Zobin[2], Mauricio Bretón[2], JRG Pulido[1], Ruben Ruelas[3], Benjamin Ojeda[3]

[1] Universidad de Colima,
Facultad de Telemática,
México

[2] Universidad de Colima,
Centro Universitario de Estudios Vulcanológicos,
México

[3] Universidad de Guadalajara,
Departamento de Ingenieria de Proyectos,
México

{felix,vzobin,mauri,jrgp}@ucol.mx,
rubrule@gmail.com, benajo@hotmail.com

Resumen. El volcán de Colima es uno de los volcanes más activos de México, por lo tanto, es necesario un monitoreo constante de su actividad explosiva para dar una alerta oportuna del peligro volcánico para las poblaciones cercanas. Cada evento explosivo genera ondas sísmicas. Las características de las señales sísmicas pueden servir para identificar un evento explosivo como fuerte, o como un evento explosivo debil. Por eso, la identificación rápida del tipo de evento explosivo puede ser importante para la estimación del nivel de peligro para las poblaciones cercanas. Normalmente la identificación del tipo de señal sísmica producto de las explosiones se realiza después de haber finalizado el evento (análisis *off-line*). Una mejora al tipo de análisis *off-line* es realizar un análisis en el momento que se origina la señal sísmica explosiva continua (análisis *on-line*) ya que esto permitirá determinar el tipo señal sísmica en una etapa temprana donde se origina el evento. Se desarrolló un algoritmo que permite identificar el tipo de señal sísmica mediante el análisis (*on-line*) y emitir una alarma a la población de la ocurrencia de un evento explosivo fuerte con una probabilidad de peligro alta. Los resultados obtenidos por nuestro algoritmo de análisis on-line, muestran que se puede identificar el tipo de evento sísmico de igual manera que mediante el método off-line.

Palabras clave: Volcán de Colima, explosiones, señales sísmicas, monitoreo sísmico, monitoreo visual.

Félix Ortigosa, Vyacheslav Zobin, Mauricio Bretón, JRG Pulido, Ruben Ruelas, Benjamin Ojeda

Characterization of Explosive Volcanic Events from the Seismic Signals of the Colima Volcano in Mexico, for the Identification of the Level of Volcanic Danger

Abstract. Colima's volcano is one of the most active volcanoes in Mexico, therefore, constant monitoring of its explosive activity is necessary to provide a timely warning of volcanic danger for nearby populations. Each explosive event generates seismic waves. The characteristics of the seismic signals can serve to identify an explosive event as strong, or as a weak explosive event. Therefore, the rapid identification of the type of explosive event can be important for estimating the level of danger for nearby populations. Normally, the identification of the type of seismic signal produced by the explosions is carried out after the event has ended (*off-line* analysis). An improvement to the *off-line* type of analysis is to perform an analysis at the moment the continuous explosive seismic signal originates (*on-line* analysis) since this will allow determining the type of seismic signal in one stage early where the event originates. An algorithm was developed that allows identifying the type of seismic signal through analysis (*on-line*) and issuing an alarm to the population of the occurrence of a strong explosive event with a high probability of danger. The results obtained by our on-line analysis algorithm show that the type of seismic event can be identified in the same way as using the off-line method.

Keywords: Colima volcano, explosions, seismic signals, seismic monitoring, visual monitoring.

1. Introducción

El Volcán de Colima es uno de los volcanes más activos de México, por tal motivo su monitoreo es constante registrando todos los eventos de su actividad. Una explosión volcánica puede ocurrir en el volcán en cualquier momento, evento eruptivo que puede contener gas, ceniza o piedra volcánica. Dependiendo de la energía liberada desde el interior hasta el cráter del volcán [10] y las consecuencias de una explosión volcánica de magnitud VEI >= 4, (Volcanic Explosivity Index) serían devastadoras en las poblaciones cercanas, principalmente por lluvia de ceniza, flujos piroclásticos y flujos de lava [6,13].

En la actualidad existe un monitoreo constante del Volcán de Colima por las redes de monitoreo sísmico y el monitoreo visual del Centro Universitario de Investigaciones Vulcanológicas de la Universidad de Colima *et al.* en [12]. Este artículo está organizado de la siguiente manera: en la sección 2 se explica cómo se obtiene la información mediante el sistema de monitoreo del Volcán de Colima, registrando por una parte las imágenes y por la otra la señal sísmica.

En la sección 3 se describe la señal sísmica de una explosión y sus características. En la sección 4 se presenta el método para procesar la información

on-line. En la sección 5 se describen los resultados, y por último se presentan las secciones de discusión y conclusiones.

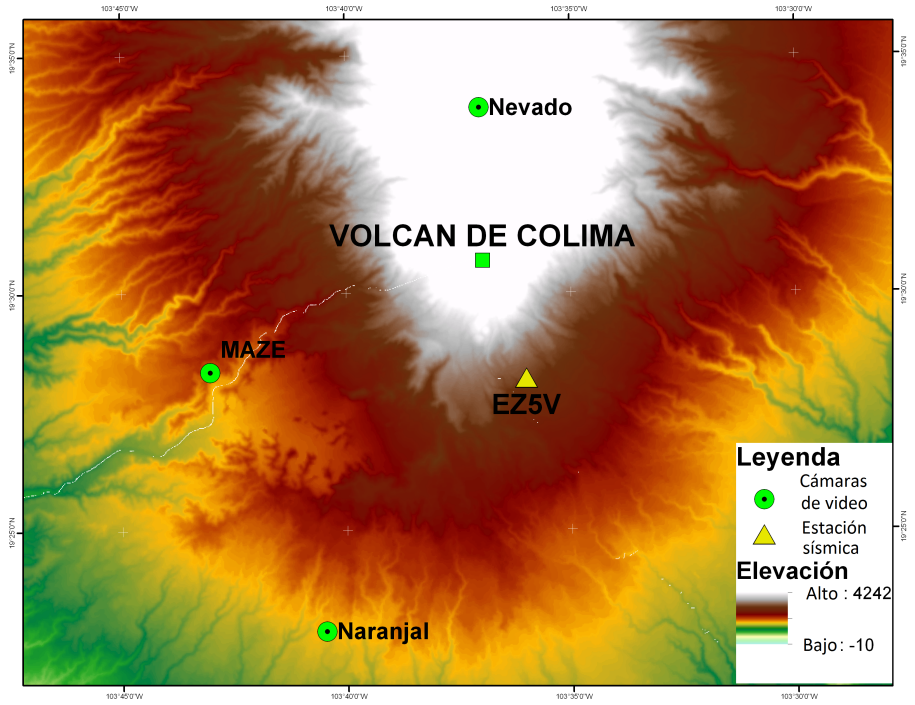


Fig. 1. Ubicación de las tres cámaras de video y de una estación sísmica de la Red de Monitoreo del Centro Universitario de Estudios Vulcanológicos de la Universidad de Colima.

2. Descripción del sistema de monitoreo

La actividad del Volcán de Colima es vigilada mediante la red de monitoreo sísmico y visual del Centro Universitario de Estudios Vulcanológicos de la Universidad de Colima. La Figura 1 muestra la posición geográfica del sistema de monitoreo de la actividad explosiva del volcán usada en nuestro estudio que consiste en las tres cámaras de video y una estación sísmica. Nuestro estudio está basado en los registros sísmicos identificados con las imágenes de video de las explosiones correspondientes.

2.1. Monitoreo visual

La red de monitoreo visual registra mediante el uso de cámaras y desde diferentes distancias, los diferentes eventos ocurridos durante las veinticuatro

horas del día. Estas imágenes son transmitidas al sistema central de monitoreo, para su posterior procesamiento [2]. Para este estudio utilizamos 3 cámaras, cada cámara registra una imagen en intervalos de 3 a 30 segundos, por minuto dependiendo de las condiciones climatológicas. En la Tabla 1 se describen las características que poseen cada una de las cámaras utilizadas.

Tabla 1. Descripción de las cámaras de video.

Nombre de Estación	Periodo de observación	Resolución pixels	Modelo de cámara	Zona de monitoreo	Distancia Focal (mm)	Coordenadas Geográficas		Distancia al cráter de volcán
						Long., W	Lat., N	(km.)
Nevado	2008-2014	704x479	Axis 213 PTZ	Crater y Domo	3.5-91	19,564°	−103,617°	5.3
Naranjal	1998-2014	704x480	Axis 213 PTZ	Lahares, Derrumbes, y edificio volcánico	3.5-91	19,381°	−103,674°	15.67
MAZE	2011-2014	704x480	Axis 213 PTZ	Lahares, Derrumbes, y edificio volcánico	3.5-91	19,473°	−1037,17°	11.3

Tabla 2. Descripción del sismómetro.

Estación	Modelo	Sensibilidad	Inicio Operación	Long., W	Lat., N	Distancia desde la estación al cráter del volcán km.	Azimut al volcán (º)
EZ5V	T4016	7.95E-10	01/06/2001	-103.602	19.479	4	157

2.2. Monitoreo sísmico

La red sísmica registra los diferentes eventos sísmicos ocurridos en el Volcán de Colima desde diferentes distancias, y realiza mediante la transmisión de señales sísmicas continuas durante las veinticuatro horas del día, en tiempo real, al departamento RESCO (REd Sismológica del estado de Colima) del Centro Universitario de Estudios Vulcanológicos de la Universidad de Colima, donde son almacenadas y procesadas.

Para este estudio utilizamos una estación sísmica denominada EZ5V, situada a 4 km del cràter del Volcán de Colima y descrita en la Tabla 2. Este sismómetro digital de banda ancha registra señales sísmicas con una frecuencia de muestreo de 100 muestras/s. Seleccionamos esta estación debido a la cercanía al cráter del volcán, además es la que mejor registra los eventos explosivos [16].

3. Descripción y análisis de la señal sísmica de explosión volcánica

Las señales sísmicas de explosiones representan los registros que consistan de las dos fases, preliminar (pre-explosiva) y principal (co-explosiva). En la Figura 2 son identificadas entre t_1 y t_2 para la primera fase y t_2 y t_3 para la segunda fase. De acuerdo con el modelo de las explosiones volcánicas, propuesto en [11], se puede identificar dos tipos de registros sísmicos de explosiones según sus fases preliminares. Se considera que la fase sísmica preliminar en caso de registar

tipo 1, representa un registro de baja frecuencia, se genera durante el ascenso del magma fragmentado a la superficie dentro del conducto del volcán.

Considerando que el magma que asciende de las profundidades dentro del conducto del volcán en caso de registar tipo 1 puede producir explosiones más fuertes. La fase sísmica preliminar en caso de registrar tipo 2, representa un registro de alta frecuencia está considerado como el resultado de la fractura de un domo de lava subsuperficial con salida de magma y gas, por lo tanto esto puede producir una explosión débil. Por consiguiente, nuestro problema es identificar el registro tipo 1 ó 2 de las señales sísmicas y en caso de ocurrencia de registro o señales tipo 1, dar una alarma esperando una explosión fuerte.

La Figura 2c-d. y la Figura 2e-f, muestra una representación del espectro de Fourier, de los picos de frecuencia entre 0 y 3 Hz. Zobin *et al.* en [15] describen que la señal sísmica de baja frecuencia en su fase de pre-explosiva tienen una duración entre 3 y 7 segundos.

Para nuestro estudio, se desarrolló un algoritmo para identificar en tiempo real el tipo de señal sísmica de tipo explosión. El algoritmo considero que la señal está compuesta de tres fases: fase de ruido precedente a la señal, fase preliminar que puede servir para identificación del tipo de señal, y fase de explosión, véase Figura 3. En la fase de ruido, las señales sísmicas contienen ruido debido a factores externos que registra el sismógrafo.

El ruido está presente constantemente en la señal y dependiendo de los factores externos puede ser mayor o menor que la información del evento detectado. Este ruido puede ser atenuado con filtros pasa bajos a 0.5 Hz [1,14,4,11]. En la fase de identificación los eventos sísmicos utilizan la frecuencia como característica principal para clasificar los eventos sísmicos de tipo explosión.

Por último en la fase de explosión, se caracterizan por tener una amplitud en esta fase se calcula la energia liberada en Joules, del evento explosivo para la toma de decisiones, se pretende el cálculo automático de este valor para un trabajo futuro.

4. Métodos y datos

El algoritmo propuesto para la identificación on-line del tipo de explosión analiza la señal sísmica (pre-explosiva) de los eventos en un mínimo de tiempo asi como el tipo de explosión. Después el algoritmo tiene que dar una estimación de energía de la explosión y decidir una necesidad en alerta para la población. En nuestro estudio utilizamos la primera etapa del algoritmo, la clasificación de la explosión en la fase de identificación.

La segunda etapa solo una parte, el calculo de frecuencias en la fase de identificación, y en la tercera etapa de la fase de explosión, se pretende desarrollar el cálculo automático de la energía liberada de este valor de modo on-line para un trabajo futuro.

La aplicación del algoritmo se realiza en tres etapas cada una con diferentes pasos descritos en la Figura 4. Estos pasos son utilizados para detectar un evento explosivo fuerte en la fases de una señal sísmica como se menciona a continuación:

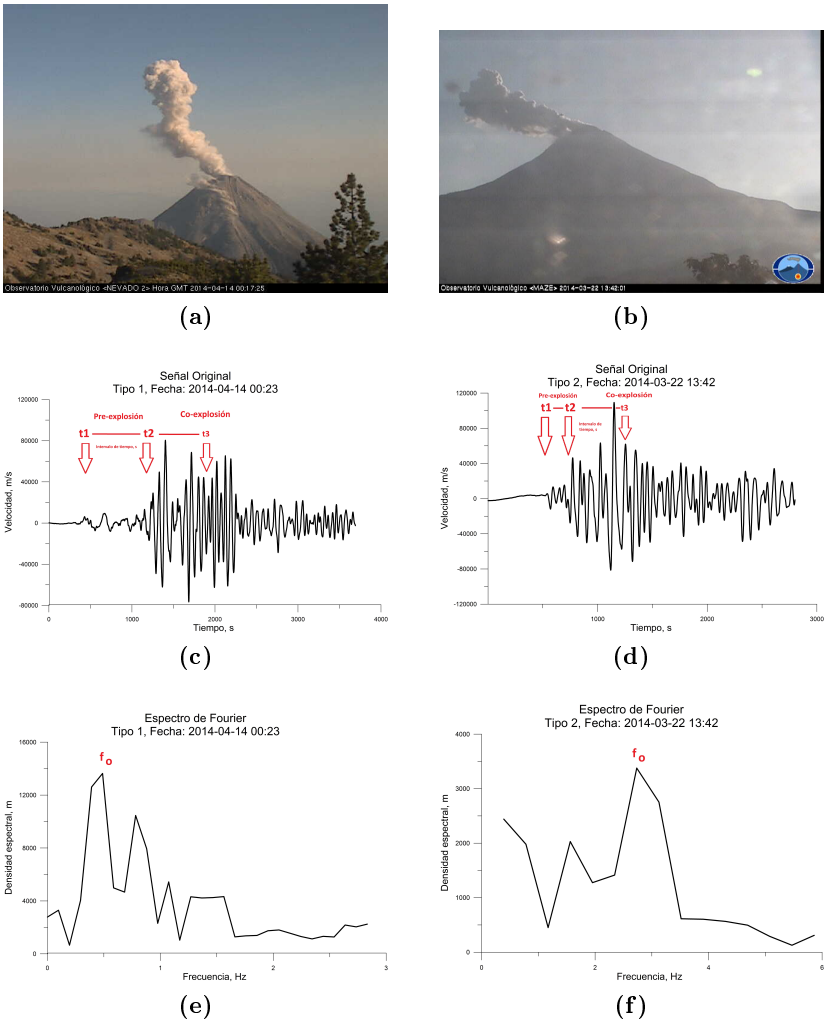


Fig. 2. Imágenes de dos tipos de explosiones volcánicas, (a) fuerte, (b) débil. Registros sísmicos: (c) tipo 1. Señal original registrada el 2014-04-14 a las 00:23 horas, (d) tipo 2. Señal original registrada el 2014-03-22 a las 13:42 horas. Espectros de Fourier de las señales de fases preliminares: (e) tipo 1, (f) tipo 2 donde se muestran las frecuencias f^0.

- Seleccionar el registro sísmico de explosión (estación EZ5v),
- Discriminar el ruido en la fase preliminar,
- Calcular el espectro Fourier de la fase preliminar,
- Identificar el tipo de explosión mediante su frecuencia espectral,
- Calcular el espectro Fourier de la fase de explosión,
- Calcular la energía de la fase de explosión,

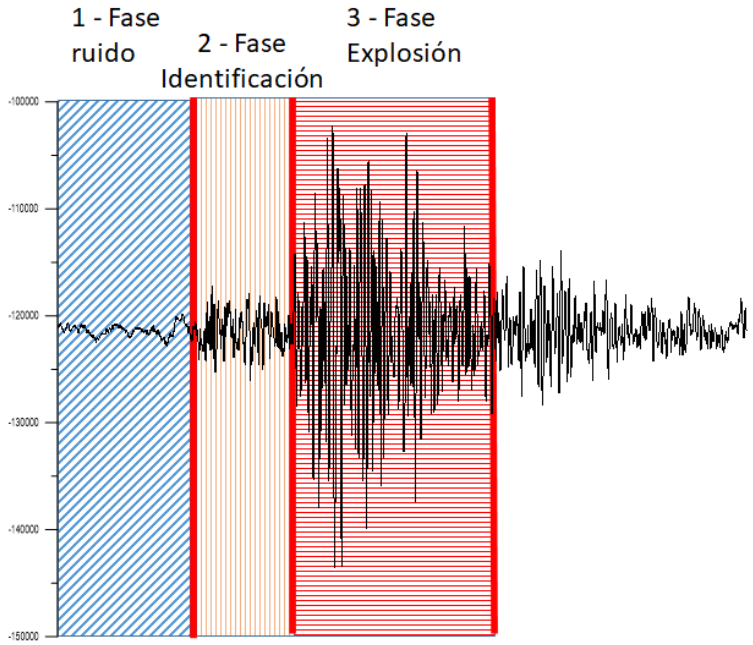


Fig. 3. Descripción de las fases de una señal sísmica explosiva.

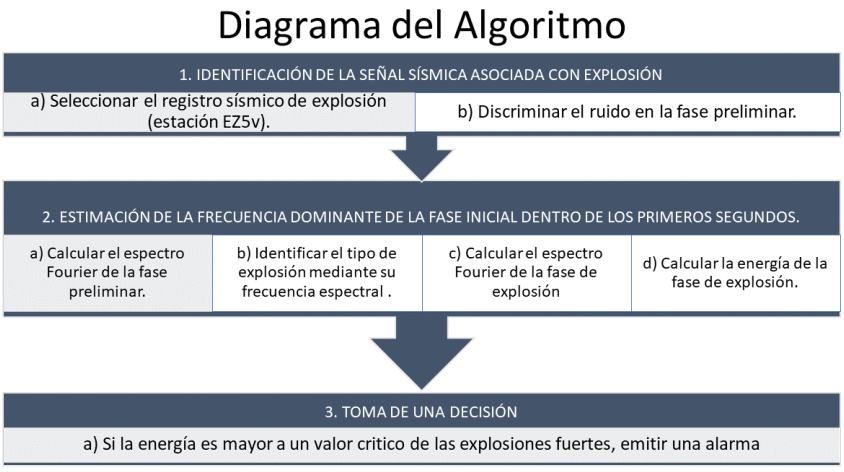


Fig. 4. Algoritmo para la identificación on-line.

– Si la energía es mayor a un valor crítico de las explosiones fuertes, emitir una alarma.

Este estudio utilizó la información registrada por el monitoreo sísmico y el monitoreo visual por un periodo de 18 meses, enero-diciembre 2013 y enero-julio 2014. Se utilizaron las imágenes visuales digitales y los sismogramas digitales

de la señal sísmica. Con la información proporcionada por el monitoreo visual se clasificaron las imágenes con una nube de ceniza característica con duración entre 10 y 20 segundos, obteniendo la fecha y hora del evento, consideradas como grandes; con esta información se encontraron un total de 1590 explosiones grandes. La Figura 5 muestra el registro mensual de las explosiones encontradas por cámara del periodo analizado.

Con la información del monitoreo sísmico, se analizó el sismograma de la señal sísmica tomando como referencia la Figura 2. El análisis de las secuencias de imágenes de video junto con los registros sísmicos de explosiones del Volcán de Colima mostró que unos segundos antes de la ocurrencia de la explosión del cráter de volcán, se inicia el registro sísmico preliminar en la señal sísmica de la frecuencia baja (tipo 1) o frecuencia alta (tipo 2), de amplitud muy baja. Después de estas señales preliminares tenemos también el registro de la misma explosión con amplitud más alta (Figura 2).

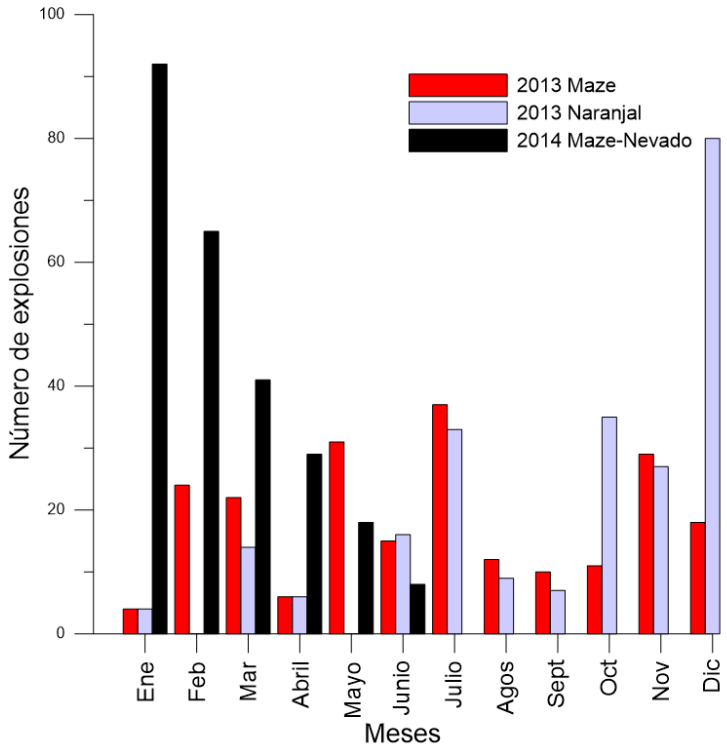


Fig. 5. Eventos clasificados mensualmente por cámara de explosiones ocurridas en el 2013-2014.

5. DEGTRA A4 y script en MATLAB

Para el procesamiento de una señal sísmica se utilizaron dos programas DEGTRA A4 y un script basado en Matlab. El programa DEGTRA A4, ha sido desarrollado por el Instituto de Ingeniería de la Universidad Nacional Autónoma de México, (UNAM). DEGTRA A4 es una herramienta académica de distribución gratuita academica. En Degtra A4 se procesa una señal en cualquier derivación con respecto a las necesidades del usuario, sismólogos e ingenieros sísmicos.

Entre sus caracteristicas encontramos que utiliza filtros básicos y avanzados, y permite evaluar, la coherencia, la correlación cruzada, integrar, derivar, calcular espectros de amplitud de Fourier, mostrar en forma de odograma los registros sísmicos de componentes diferentes, rotar, sumar, restar, corregir la línea base de diversas formas, entre otras funciones, [7].

Por otra parte, el script en Matlab es una derivación de la aplicación de Lesage [5] con características ad hoc de eventos sísmicos, aplicadas a las necesidades del estudio, denominadas como el algoritmo. Cabe mencionar que el modo de análisis off-line es el diagnóstico del experto después de ocurrido el evento y el modo on-line es un análisis mediante un algoritmo sin ayuda del experto.

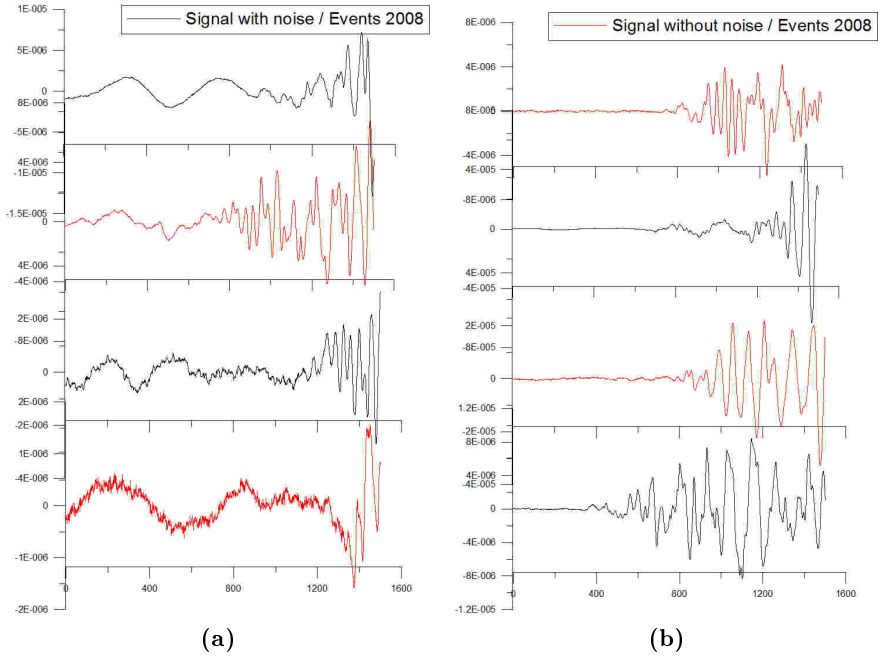
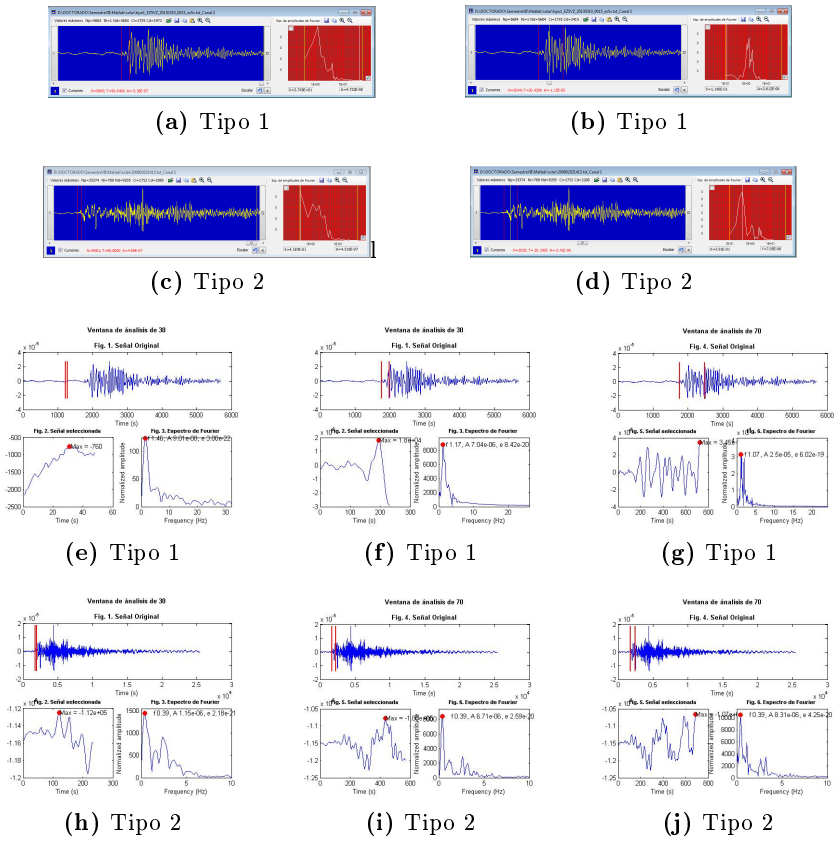


Fig. 6. Señal sísmica explosiva: a) Con ruido, b) Sin ruido.



(a) Tipo 1 **(b)** Tipo 1

(c) Tipo 2 **(d)** Tipo 2

(e) Tipo 1 **(f)** Tipo 1 **(g)** Tipo 1

(h) Tipo 2 **(i)** Tipo 2 **(j)** Tipo 2

Fig. 7. Imágenes de la identificación de eventos sísmicos usando el programa DEGTRA A4 y el algoritmo. a,b,c,d utilizan el programa DEGTRA A4; e,f,g,h,i,j utilizan el algoritmo; a,b,e,f,g son señales Tipo 1; c,d,h,i,j son señales Tipo 2.

6. Resultados

Analizando cada sismograma individualmente se encontró que la fase de ruido está muy presente en algunas señales, por lo tanto se aplicó el filtro pasa bajos a 0.5 Hz. Sin embargo, al aplicar el filtro se elimina parte de la información de toda la señal sísmica; por lo tanto los resultados para la clasificación por tipo de explosión (tipo 1 y 2) no eran satisfactorios, y por ello se volvió a clasificar los sismogramas sin aplicar el filtro, eliminando los sismogramas con un ruido muy evidente, véase Figuras 6a y 6b con ruido y sin ruido respectivamente. Para este estudio fueron seleccionados 90 eventos sin ruido, véase Tabla 4.

Al aplicar el algoritmo se obtuvieron los valores de frecuencia de las fases preliminares para realizar una comparación entre frecuencias obtenidas con el programa DEGTRA A4 y el algoritmo, véase Figura 7a,b,e,f,g para un evento

Tabla 3. Valores de eventos Tipo 2.

DEGTRA A4	Algoritmo
f, Hz	f, Hz
2.34	2.25
1.95	1.86
2.73	2.54
2.73	0.49
1.56	2.54
1.76	1.76
2.73	2.93
1.95	1.95
1.56	1.56
1.56	2.05
2.73	2.64
1.95	1.86
2.73	2.64
2.73	1.46
2.73	2.64
1.95	1.66
A	B

Tabla 4. Número de eventos analizados.

A	**2013 (1-12)**	**2014 (1-6)**
Tipo 1	37	37
Tipo 2	10	6
Total	47	43

Tipo 1 y Figura 7c,d,h,i,j para un evento Tipo 2. La Tabla 5 muestra los datos para eventos Tipo 1, las columnas A y B muestra los valores obtenidos con el programa DEGTRA A4; las columnas C y D muestra los valores obtenidos por el algoritmo; la Tabla 3 se muestran los datos para eventos Tipo 2, la columna A muestra los valores obtenidos con el programa DEGTRA A4; la columna B muestra los valores obtenidos por el algoritmo.

Esta prueba se realizó con el programa estadístico SPSS [9,3,8]. Con estos datos se muestra que el análisis en modo on-line es bastante seguro e informativo para dar alerta de peligro volcánico en caso de ocurrencia de eventos explosivos de tipo 1. Con los datos de los eventos Tipo 1 y Tipo 2 fue realizada una prueba estadística T para grupos relacionados como hipótesis nula *Ho*: El valor de la frecuencia es el mismo entre la primera y la segunda medición con un nivel de significancia de $p<0.05$. Los valores del grupo a son los valores obtenidos con el programa DEGTRA A4 *off-line*, y el grupo b los valores obtenidos con el algoritmo *on-line*. Véase Tabla 6.

Tabla 5. Valores de eventos Tipo 1.

DEGTRA A4		Algoritmo	
f, Hz	f, Hz	f, Hz	f, Hz
0.39	0.39	0.39	0.39
0.88	0.39	0.88	0.39
0.98	0.39	0.98	0.39
0.78	0.49	0.78	0.49
0.39	0.49	0.39	0.49
0.78	0.49	0.78	0.49
0.78	0.49	0.78	0.49
0.88	0.49	0.88	0.49
0.78	1.17	0.78	1.17
0.49	0.78	0.49	0.78
0.88	0.49	0.88	0.49
0.39	0.78	0.39	0.78
0.88	0.49	0.88	0.49
1.07	0.78	1.07	0.49
0.78	0.78	0.78	0.39
0.39	0.49	0.39	0.49
0.39	0.49	0.39	0.49
1.17	0.39	1.17	0.39
0.78	0.78	0.78	0.78
0.88	0.39	0.88	0.39
0.78	0.49	0.78	0.39
0.49	0.78	0.49	0.78
0.78	0.49	0.78	0.49
0.44	0.78	0.39	0.78
0.78	0.88	0.78	0.88
0.59	0.49	0.49	0.49
0.49	0.39	0.49	0.39
0.39	0.39	0.39	0.39
0.49	0.78	0.49	0.88
1.17	0.88	1.27	0.88
0.39	0.78	0.39	0.59
0.39	0.39	0.39	0.39
0.44	0.39	0.49	0.39
0.88	0.49	0.88	0.39
0.78	0.59	0.78	0.59
0.39	0.39	0.39	0.39
0.49	0.39	0.49	0.39
A	B	C	D

7. Discusión

Como se puede observar en los resultados, el algoritmo propuesto tuvo un mejor rendimiento para eliminar la información en la fase de ruido y una mejor identificación en la fase inicial de la señal sísmica en eventos Tipo 1 ó 2, para encontrar los valores de frecuencia.

Con esto se prueba que no es necesario analizar la señal sísmica en modo off-line para identificar el tipo de explosión, es decir, es posible identificarla en

Tabla 6. Resultado al aplicar la prueba T.

	a	**b**	**a >= b**
	SPSS	Valores Table t;	
Type 1 (n=74)	t= 1.716; gl=73;	t=1.6660; gl=73;	1.716 >= 1.660
	p= 0.090	p=0.05	0.090 >= 0.05
			Accept H○
Type 2 (n=16)	T=1.009; gl=15;	t=1.7531; gl=15;	1.009 >= 1.7531
	p= 0.329	p=0.05	0.329 > 0.05
			Accept H○

modo on-line. Con la prueba T se justificó que los valores obtenidos de frecuencia con el algoritmo y el programa DEGTRA A4 son muy similares con un porcentaje de confiabilidad del 95 %.

La Tabla 6 muestra que el nivel de significancia calculado de 0.090 es mayor que el 0.050 esperado, para los eventos Tipo 1; para el valor de significancia de 0.329 es mayor que el 0.05 esperado, para los eventos Tipo 2. Por lo tanto, se acepta la hipótesis de que la frecuencia es la misma para el programa DEGTRA A4 y para el algoritmo.

8. Conclusiones

En este trabajo se ha mostrado la identificación en tiempo real de los eventos explosivos con alta peligrosidad para poblaciones cercanas al Volcán de Colima, lo que es de vital importancia para ganar tiempo en la emisión de alarmas y con ello mitigar los posibles daños a consecuencia de las explosiones. Con esto se cambia el enfoque actual de un sistema de información sobre lo sucedido a un sistema de actuación sobre lo que está ocurriendo. Para el cálculo automático del valor de la energía liberada del evento explosivo, en Joules, se pretende realizar como trabajo futuro, así como la implementación del algoritmo para la identificación automática de los eventos explosivos y la emisión de las alarmas correspondientes.

Agradecimientos. A Mario Ordaz por el procesamiento de las señales sísmicas digitales con el programa DEGTRA A4, a Phillipe Lessage por la colaboración en su programa y al CUIV de la Universidad de Colima por su ayuda y colaboración en esta investigación. F. Ortigosa agradece al CONACyT y a la Universidad de Colima por financiar sus estudios de Doctorado.

Referencias

1. Abreu, R. D., Reddan, S., Parent, J., Johnson, D.: Seismic event detection using three-component data. IEEE Transactions on Geoscience and Remote sensing, vol. 30, no. 3, pp. 642–644 (1992)

2. Bretón-Gonzalez, M., Campos, A., León, Z., Plascencia, I., Ramirez, J. J.: The 2007-2012 lava dome growth in the crater of Volcan de Colima, México, derived from video monitoring system. Complex Monitoring of Volcanic Activity Methods and Results, pp. 153–170 (2013)
3. Castañeda, B., Cabrera, A., Navarro, Y., De Vries, W.: Procesamiento de datos y analisis estadísticos utilizando SPSS. EdiPUCRS (2010)
4. Duin, R. P., Orozco-Alzate, M., Londono-Bonilla, J. M.: Classification of volcano events observed by multiple seismic stations. In: 2010 20th International Conference on Pattern Recognition. pp. 1052–1055. Ieee (2010) doi: 10.1109/ICPR.2010.263
5. Lesage, P.: Interactive matlab software for the analysis of seismic volcanic signals. Computers Geosciences, vol. 35, no. 10, pp. 2137–2144 (2009) doi: 10.1016/j.cageo.2009.01.010
6. Mauricio Breton- Gonzalez, Ramirez, J. J., Navarro, C.: Summary of the historical eruptive activity of Volcan de Colima, Mexico 1519-2000. Journal of Volcanology and Geothermal Research, vol. 117, pp. 21–46 (2002)
7. Ordaz, M., Francisco, C., Zapata, A.: Manual DEGTRA A4 Ver 5.4 1 (2005)
8. Roberto Hernandez, S., Collado Fernandez, C., Baptista Lucio, M. d. P.: Metodología de la Investigación. McGraw Hill (2010)
9. Segnini, S.: Fundamentos de bioestadistica (2008)
10. Sigurdsson, H.: Explosive volcanism. Encyclopedia of Volcanoes, pp. 420–696 (2000)
11. Zobin, V. M.: Seismic signals associated with volcanic explosions. Introduction to Volcanic Seismology, pp. 295–326 (2012)
12. Zobin, V. M., Arámbula, R., Bretón, M., Reyes, G., Plascencia, I., Navarro, C., Téllez, A., Campos, A., González, M., León, Z., Martínez, A., Ramírez, C.: Dynamics of the January 2013 - June 2014 explosive-effusive episode in the eruption of Volcán de Colima, México: insights from seismic and video monitoring. Bulletin of Volcanology, vol. 1, no. January 2013, pp. 1–13 (2015) doi: 10.1007/s00445-015-0917-z
13. Zobin, V. M., Luhr J.F., Taran, Y., Breton, M., Corte, A., de la Cruz-Reyna, S., Dominguez, T., Galindo, I., Gavilanes, J. C., Muniz, J., Navarro, C., Ramirez, J., Reyes, G. A., Ursua, M., Velasco, J., Alatorrre, E., Santiago, H.: Overview of the 1997-2000 activity of Volcan Colima Mexico. Journal of Volcanology and Geothermal Research, vol. 117, pp. 1–19 (2002)
14. Zobin, V. M., Navarro, C., Reyes-Dávila, G., Orozco, J., Bretón, M., Tellez, A., Reyes-Alfaro, G., Vázquez, H.: The methodology of quantification of volcanic explosions from broad-band seismic signals and its application to the 2004-2005 explosions at Volcán de Colima, Mexico. Geophysical Journal International, vol. 167, no. 1, pp. 467–478 (2006) doi: 10.1111/j.1365-246X.2006.03108.x
15. Zobin, V. M., Reyes, G., Guevara, E., Bretón, M.: Seismological constraints on the position of the fragmentation surfaces in the volcano conduit. Earth and Planetary Science Letters, vol. 275, no. 3-4, pp. 337–341 (2008) doi: 10.1016/j.epsl.2008.08.034
16. Zobin, V. M., Varley, N. R., González, M., Orozco, J., Reyes, G. a., Navarro, C., Bretón, M.: Monitoring the 2004 andesitic block-lava extrusion at Volcán de Colima, México from seismic activity and SO2 emission. Journal of Volcanology and Geothermal Research, vol. 177, no. 2, pp. 367–377 (2008) doi: 10.1016/j.jvolgeores.2008.05.015

ISSN 1870-4069

Una evaluación comparativa de modelos de Deep Learning para el reconocimiento de emociones a partir del habla

Luis Bernal, Álvaro Cuno, Wilber Ramos Lovón

Universidad Nacional de San Agustín de Arequipa,
Departamento Académico de Ingeniería de Sistemas,
Perú

{lbernal, acunopa, wramos}@unsa.edu.pe

Resumen. Diversos modelos de reconocimiento de emociones a partir del habla vienen siendo propuestos en los últimos años. Sin embargo, las evaluaciones de desempeño de algunas propuestas podrían no ser lo suficientemente confiables. Este trabajo tiene como finalidad contribuir con abordar esta problemática, presentando una manera práctica de implementar evaluaciones comparativas cuyos resultados disten de ser atribuidos a la casualidad. La propuesta utiliza las pruebas de significancia estadística no paramétricas en base al test de Wilcoxon para la comparación del desempeño de modelos de *Deep Learning*. Se demostró la utilidad de la propuesta, realizando las comparación del desempeño de cinco modelos convolucionales, dos entrenados con la base de datos RAVDESS y tres entrenados con la base de datos IEMOCAP.

Palabras clave: Emociones, evaluación, habla.

A Comparative Evaluation of Deep Learning Models for Emotion Recognition from Speech

Abstract. Various models of emotion recognition from speech have been proposed in recent years. However, performance evaluations of some proposals may not be reliable enough. The purpose of this work is to contribute to solve this problem, presenting a practical way to implement comparative evaluations whose results are far from being attributed to chance. The proposal uses non-parametric statistical significance tests based on the Wilcoxon test to compare the performance of *Deep Learning* models. The usefulness of the proposal was demonstrated, comparing the performance of five convolutional models, two trained with the RAVDESS database and three trained with the IEMOCAP database.

Keywords: Emotions, evaluation, speech.

1. Introducción

El interés en lograr reconocer las emociones humanas de manera automatizada ha sido motivado por diversas áreas de aplicación. Por ejemplo, para diseñar robots inteligentes que puedan interactuar de manera empática con las personas, para crear anuncios personalizados que tengan en cuenta el estado emocional de potenciales clientes, para mejorar los procesos de enseñanza/aprendizaje, para mostrar contenido más adecuado a una audiencia objetivo, entre otras aplicaciones [4].

Existen varios métodos para medir las emociones humanas, cada una con sus ventajas y desventajas. Para el caso de los sistemas automatizados, podemos clasificarlos en invasivos y no invasivos. Entre los invasivos, tenemos aquellos que utilizan dispositivos pegados al cuerpo para medir alguna señal fisiológica como electrocardiogramas, electroencefalogramas, respuestas galvánicas de la piel, temperatura, respiración, entre otros.

Entre los no invasivos podemos encontrar a los que utilizan como señal de entrada expresiones faciales, gestos, posturas, voz, entre otras. Un sistema de reconocimiento de emociones a partir del habla busca inferir, de manera automatizada, cual es la emoción que una persona está expresando al momento de hablar. Esto se realiza tomando como entrada las ondas de sonido que se producen al hablar y que son capturadas por uno o varios micrófonos.

Tradicionalmente, este problema había sido abordado mediante la selección manual de características y haciendo uso de algoritmos de clasificación basados en aprendizaje de máquina(*machine learning*). Sin embargo, tras el notable incremento del poder computacional en los últimos años, las técnicas basadas en aprendizaje profundo(*deep learning*), que extraen características de manera automática, han superado en precisión a estos métodos tradicionales [1].

Debido a esto, en la actualidad el estado del arte se soporta en modelos de aprendizaje profundo. Para poder comparar el desempeño de los modelos de manera objetiva una o varias métricas de evaluación deben ser utilizadas. Esta tarea está lejos de ser trivial, principalmente, debido a la presencia de algunos procedimientos de carácter aleatorio en los modelos (p.ej. inicialización de pesos, particionamiento de las bases de datos, entre otros).

Debido a esta cuota de aleatoriedad, los resultados de las evaluaciones suelen ser no reproducibles y pueden variar entre una u otra implementación o incluso entre una u otra ejecución. La complejidad es mayor cuando se trata de comparar una nueva propuesta con modelos existentes en el estado del arte. Esto se debe, principalmente, a que los modelos han sido configurados para obtener los mejores resultados, utilizando bases de datos, formas de entrenamiento, métricas y evaluaciones, de carácter específico y particular a un modelo.

Si se cambia el particionamiento de la base de datos o la manera de inicializar los pesos de los modelos, los resultados podrían variar. A esto se suma la indisponibilidad de código fuente de los modelos, que dificulta replicar las evaluaciones y hacer comparaciones justas [6]. Una implementación o configuración errada de modelos propuestos por terceros podría sesgar los resultados de las comparaciones de manera voluntaria o involuntaria.

Una alternativa que se suele utilizar para facilitar las comparaciones entre modelos es la realización de pruebas comparativas (*benchmarks*), que consisten en distribuir una base de datos de manera pública y desarrollar competiciones por alcanzar la mejor predicción posible. Sin embargo, esto también puede llevar a conclusiones erróneas. Se ha visto casos donde es posible vulnerar los *benchmarks*, permitiendo a un modelo alcanzar el podio sin siquiera haber sido entrenado, solamente observando los resultados de exactitud (*accuracy*) obtenidos [15].

Otra manera de demostrar que un modelo presenta una mejora frente a otros modelos, es mediante la realización de pruebas de significancia estadística. Esta alternativa utiliza la estadística para determinar que un modelo es diferente y superior a otro fuera de los márgenes de la casualidad. Sin embargo, una implementación deficiente de las pruebas de significancia estadística podría llevarnos a conclusiones erróneas. Por lo tanto, en esta investigación se presenta una manera práctica de implementarlas. Se demostró la utilidad de la propuesta, realizando la comparación estadística del desempeño de tres modelos convolucionales entrenados con la base de datos RAVDESS y tres modelos para la base de datos IEMOCAP.

2. Trabajos relacionados

La comparación del desempeño de modelos de clasificación basados en aprendizaje de máquinas no es una tarea trivial [12]. Para el caso de modelos de reconocimiento de emociones de aprendizaje profundo, la mayoría de las revisiones de literatura, por ejemplo [11,14,1,8], se limitan a realizar análisis descriptivos de los métodos de reconocimiento de emociones sin llegar a replicar (diferente equipo, mismo diseño experimental) o reproducir (diferente equipo, diferente diseño experimental) las investigaciones del estado del arte.

Pero si hablamos de comparaciones específicas, Fayek et al. [6], presentaron una comparación de algunos modelos de aprendizaje profundo, siendo esta forma (replicar por uno mismo varios modelos y evaluarlos) la más común a la que los autores recurren para comparar una propuesta contra el estado del arte. Este problema se extiende a otras áreas donde se ha observado problemas de rigurosidad, replicabilidad y falta de estándares para realizar comparaciones.

Por ejemplo, en el área de sistemas de recomendación, Zun et al. [13] ponen en cuestión las evaluaciones realizadas y proponen una herramienta para hacerlo de manera adecuada. De la misma forma en el área de redes neuronales basadas en grafos, Errika et al. [5] proponen dos fases para conseguir comparaciones justas y reproducibles: selección de modelos y evaluación de modelos. También explora los fallos más comunes para lograr reproducibilidad como la ausencia de información sobre el preprocesamiento y el particionamiento de los datos.

Este tipo de investigaciones nos demuestran lo lejos que estamos de tener comparaciones confiables. Los mayores esfuerzos están sumados en la realización de desafíos (*challenges*), que son competencias donde se distribuye una base de datos etiquetada. Los competidores entrenan sus modelos y finalmente los modelos deben ser evaluados contra una base de datos de prueba.

Según los resultados obtenidos, de acuerdo a alguna métrica elegida, se ubica a los competidores en una tabla de posiciones. De esta forma se busca garantizar que las comparaciones se están realizando en las mismas condiciones con respecto al conjunto de datos proporcionado. En el 2011 se presentó el primer AVEC (Desafío de emociones audiovisuales - *Audio/Visual Emotion Challenge*) [10], que incluyó un sub-desafío utilizando solamente audio. En este desafío se utilizó la base de datos SEMAINE y se proporcionó 3 particiones de datos, evaluándose la exactitud. Esta competencia se llevó a cabo hasta el año 2015 [9], cambiando posteriormente los desafíos a detección de emociones específicas como afecto o emoción.

Aunque AVEC está enfocado en el estudio de detección de emociones multimodal, sentó las bases para evaluar los modelos de manera más confiable, proponiendo bases de datos, métricas e incluso un punto de referencia (*baseline*). EmotiW (Desafío de reconocimiento de emociones - *Emotion Recognition in the Wild Challenge*) [3], de manera similar, propone cada año un desafío de detección de emociones audiovisuales. En este evento se propone el uso de una base de datos dividida en tres particiones (entrenamiento, validación y pruebas) y de seis emociones categóricas (enojo, digusto, miedo, neutral, tristeza y sorpresa).

Aunque EmotiW está enfocado en videos como datos de entrada, propuso un punto de referencia basado en el aprendizaje profundo, lo cual es innovador en comparación a AVEC donde se utilizaba características manuales para el entrenamiento de los modelos. Sin embargo, cuando se realizan desafíos no se tiene certeza de que los resultados no hayan sido fruto de la casualidad, o de un intento de ataque por fuerza bruta como lo demuestra Whitehill J. [15]. Es por eso que se requiere explorar estrategias más robustas, como la validación cruzada por k-fold. La cual permite hacer varias pruebas sobre la misma base de datos, permitiendo detectar alteraciones de los resultados por causa del particionamiento de la base de datos.

3. Materiales y métodos

3.1. Materiales

Los materiales utilizados en esta investigación están conformados por dos bases de datos de emociones en el habla, tres modelos de aprendizaje profundo por cada una de ellas y un computador donde se ejecutan los experimentos. A continuación se detallan las bases de datos utilizadas:

- RAVDESS (*Ryerson Audio-Visual Database of Emotional Speech and Song*) [7]: Esta base de datos cuenta con 1441 segmentos de audios, que suman cerca de una hora y media de duración. Los audios están etiquetados con las emociones: calma, felicidad, tristeza, enojo, miedo, sorpresa y disgusto.
- IEMOCAP (*Interactive Emotional Dyadic Motion Capture*) [2]: Es una base de datos de emociones categóricas y dimensionales, representadas en audio y video, con cerca de 10000 muestras.

CNN1D	M11
Conv1d(kernel_size=80) BatchNorm1d(n_channel) MaxPool1d(4) Dropout() Conv1d(kernel_size=3) BatchNorm1d() MaxPool1d(4) Dropout() Conv1d(kernel_size=3) BatchNorm1d() MaxPool1d(4) Dropout() Conv1d() BatchNorm1d() MaxPool1d(4) Linear()	Conv1d(kernel_size=80) BatchNorm1d() ReLU() Dropout(0.15) MaxPool1d(4) Conv1d(kernel_size=3) ReLU() Dropout(0.15) Conv1d(kernel_size=3) ReLU() Dropout(0.15) MaxPool1d(4) Conv1d(kernel_size=3) ReLU(), Dropout(0.15), Conv1d(kernel_size=3) ReLU(), Dropout(0.15), MaxPool1d(4), Conv1d(kernel_size=3) ReLU() Dropout(0.15) Conv1d(kernel_size=3) ReLU() Dropout(0.15), Conv1d(kernel_size=3) ReLU() Dropout(0.15) MaxPool1d(4) Conv1d(kernel_size=3) ReLU() Dropout(0.15) Conv1d(kernel_size=3) ReLU() Dropout(0.15) Linear()

Fig. 1. Arquitectura de los modelos utilizados para la base de datos RAVDESS.

Para la base de datos RAVDESS se implementaron dos modelos de redes neuronales basados en convoluciones 1D, que toman como entrada un vector de audio plano. La arquitectura de los modelos (m11, m5) se detallan en la Figura 1.Mientras que para la base de datos IEMOCAP se implementaron tres modelos que utilizan una arquitectura de Red Neuronal Convolucional (Convolución + MaxPool + Convolución + MaxPool + ... + Flatten + Dense).

Cada modelo recibe como entrada una matriz que representa un espectrograma MFCC de un segmento de audio. La arquitectura de cada modelo (A, B , C) se detallan en la Figura 2. Para la implementación de los algoritmos y modelos se utilizó las librerías Pytorch y Sklearn. El entorno computacional utilizado en el entrenamiento de los modelos y la ejecución de los experimentos estuvo

conformado por un computador personal con procesador Core i5 9400F, 16 GB de memoria RAM, una tarjeta NVidia RTX 2060 (6GB Vram) y el Sistema Operativo Linux.

Modelo A	Modelo B	Modelo C
Input(shape=(20, window)) Conv2D(64, 3, "relu") MaxPool2D((1, 2)) Conv2D(64, 3, "relu") MaxPool2D((1, 2)) Conv2D(64, (1, 3), "relu") MaxPool2D((1, 2)) Conv2D(64, (1, 3), "relu") MaxPool2D((2, 1)) Conv2D(128, 3, "relu") Dropout(0.3) Flatten() Dense(1024, "relu") Dropout(0.4) Dense(8, "softmax")	Sequential(Input(shape=(20, window)) Conv2D(64, 3, "relu") MaxPool2D((1, 2)) Conv2D(64, 3, "relu") MaxPool2D((1, 2)) Conv2D(128, 3, "relu") MaxPool2D((1, 2)) MaxPool2D((2, 1)) Conv2D(128, 4, "relu") Dropout(0.3) Flatten() Dense(1024, "relu") Dropout(0.4) Dense(8, "softmax")	Sequential(Input(shape=(20, window)) Conv2D(64, 3, "relu") MaxPool2D((1, 2)) Conv2D(64, (1, 3), "relu") MaxPool2D((1, 2)) Conv2D(64, (1, 3), "relu") MaxPool2D((2, 1)) Conv2D(128, 3, "relu") Dropout(0.3) Flatten() Dense(1024, "relu") Dropout(0.4) Dense(1024, "relu") Dropout(0.4) Dense(8, "softmax")

Fig. 2. Arquitectura de los modelos utilizados para la base de datos IEMOCAP.

3.2. Métodos

El método utilizado en la presente investigación consiste en la implementación de los siguientes procedimientos:

1. **Pre-procesamiento:** Todos los segmentos de audios son remuestrados a 8khz. Y en el caso de IEMOCAP transformados a su representación MFCC (Coeficientes Cepstrales en las Frecuencias de Mel - *Mel Frequency Cepstral Coefficients*). Los parámetros utilizados para esta transformación fueron: $WindowSize = 55$ y $HopLength = 10$.
2. **Selección de arquitectura de modelos:** La arquitectura de los modelos a comparar es determinada empíricamente buscando maximizar la métrica exactitud (*accuracy*) en la validación.
3. **Recolección de muestras:** Una vez que los modelos a ser comparados han sido determinados, se procede a recolectar muestras de su comportamiento con diferentes particiones de la base de datos. Para esto, se utilizan iteraciones sucesivas de validaciones por k-fold de la manera presentada en el Algoritmo **??**. Se definen los modelos a evaluar M_A y M_B y en cada iteración

se recolecta un conjunto C de validaciones k-fold con longitud k:

$$C(D, M_A, M_B) = \{(S_A^j, S_B^j), (S_A^{j+1}, S_B^{j+1}), ..., (S_A^k, S_B^k))\}, \quad (1)$$

donde S_A^j se refiere al puntaje o *score* (métrica seleccionada a utilizar, por ejemplo, exactitud) obtenido por el modelo M_A sobre la partición de datos D_j generado por el método k-fold. Se repite N veces este procedimiento, obteniendo N conjuntos C_i:

$$G = \{C_i, C_{i+1}, ..., C_N\}. \quad (2)$$

4. **Prueba de significancia estadística:** Con las muestras recolectadas, se realizan las pruebas de significancia estadística. Para cada $C_i \in G$, se calculan los valores p-value y t-value mediante la prueba estadística de Wilcoxon. El valor t indica cuanta diferencia existe entre los resultados obtenidos por los dos modelos. Mayores valores de t implican menores valores de p. Cuando $p < 0{,}05$ podemos asumir que existe diferencia estadísticamente significativa. Como puede verse en el Algoritmo **??**, este procedimiento se realiza para cada $C_i \in G$.

Si analizamos el subconjunto $SG = \{p | p \in P \wedge p < 0{,}05\}$ y observamos la proporción de las cardinalidades $n(SG)/n(G)$, es decir, el porcentaje de muestras en las que se evidencia diferencia significativa, se puede apreciar el comportamiento general comparativo de estos modelos. El código fuente de esta implementación se encuentra disponible en el siguiente repositorio[1].

4. Resultados

Para la base de datos RAVDESS se compararon los modelos CNN1D Short, M11 (batch_size = 128) y M11 (batch_size = 128) utilizando **10 iteraciones** del algoritmo propuesto y un k-fold de 6.

Cuando observamos la comparación de los modelos CNN1D y M11_128 (Fig. 3), encontramos 4 resultados por debajo del umbral $p < 0{,}05$. Lo que indica diferencia significativa. Pero también podemos observar 6 resultados por encima del umbral.

Es decir, bajo ciertas condiciones hay evidencia para afirmar que existe diferencia significativa entre los modelos. Pero, en otras condiciones podríamos tener evidencia de lo contrario. En la comparación de los modelos CNN1D y M11_256 (Figura 4) observamos el mismo comportamiento que la comparación anterior.

La comparación entre los modelos M11_128 y M11_256, (Fig. 6) muestra que en todos los resultados no existe diferencia significativa entre las dos configuraciones del modelo M11. El principal limitante para obtener más iteraciones

[1] https://github.com/luantber/ser-benchmark

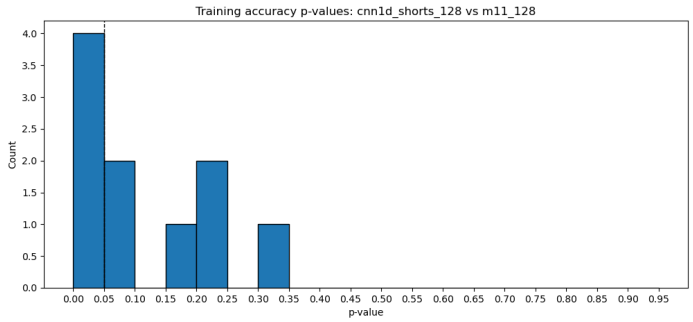


Fig. 3. Comparativa de los modelos CNN1D vs M11_128.

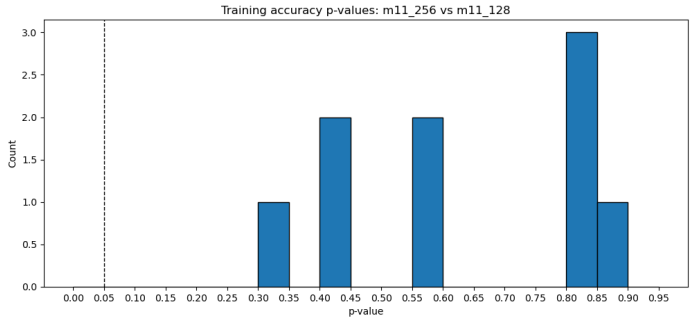


Fig. 4. Comparativa de los modelos CNN1D vs M11_256.

es el alto costo computacional del algoritmo, pues requiere reentrenar todos los modelos $N \times K$ veces.

Por otro lado, utilizando el dataset IEMOCAP, y tras recolectar datos de la métrica de *accuracy* durante 30 iteraciones y k-fold igual a 5 por un periodo de tiempo de cerca de 41 horas, se obtuvieron los resultados que se presentan en la Figura 5. Como se puede observar, todos los valores p-value obtenidos están por encima de 0.05, y dado que definimos como umbral para rechazar la hipótesis nula valores menores a 0.05, no podemos afirmar que exista alguna diferencia significativa entre los tres modelos para este dataset.

5. Discusión

Los resultados obtenidos en el experimento utilizando el dataset RAVDESS demuestran que, incluso, realizando una estrategia de k-fold y una prueba estadística, los resultados podrían cambiar e inclusive contradecirse. Pues con el mismo modelo se podría aceptar y rechazar la hipótesis H_0 si se repitiera el experimento con la misma base de datos, pero diferentes particiones en la validación k-fold.

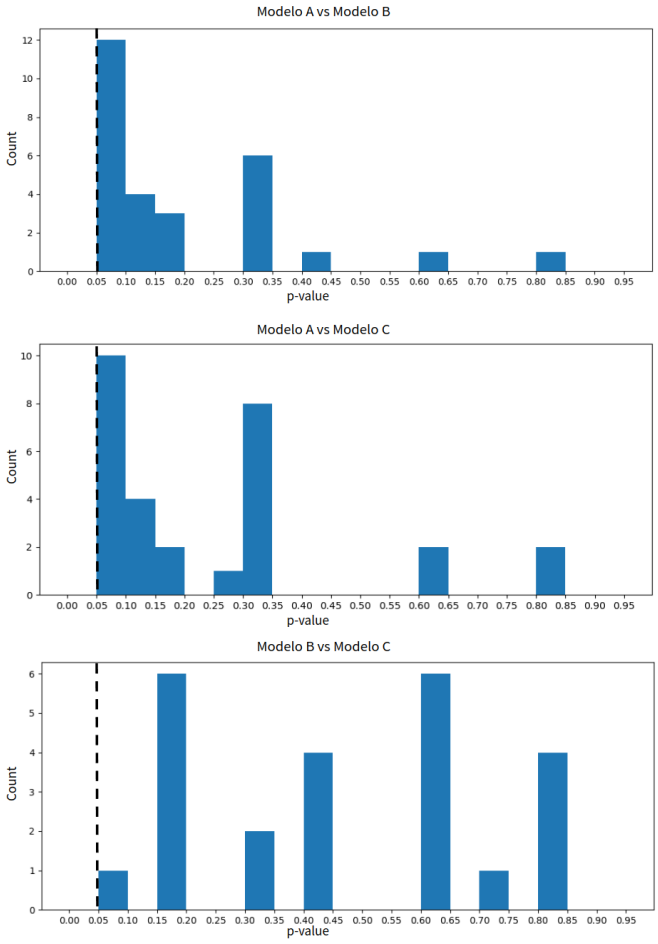


Fig. 5. *p*-values del modelo A vs. el modelo B (arriba), del modelo A vs. el modelo C (medio) y del modelo B vs el modelo C (abajo) utilizando el dataset IEMOCAP.

Por tal motivo, se podría llegar a conclusiones erróneas o sin la suficiente evidencia para afirmar que un modelo es superior a otro. Esta situación podría calificarse como producto de la casualidad o del azar. Como puede verse en los resultados, queda a criterio de los investigadores definir un umbral para declarar a un modelo superior a otro. Por ejemplo, en nuestro caso podría ser 30 %.

Con este umbral se podría afirmar que el modelo A es superior al modelo B, dado que existe 30 % de ocasiones en las que existe diferencia significativa. Sin embargo, no se podría afirmar lo mismo para la comparación del modelo B y C, pues solo un 20 % de los entrenamientos mostraron diferencia significativa. También debemos destacar que la propuesta busca mostrar la tendencia de porcentajes que tendrían los modelos si se repitieran más veces.

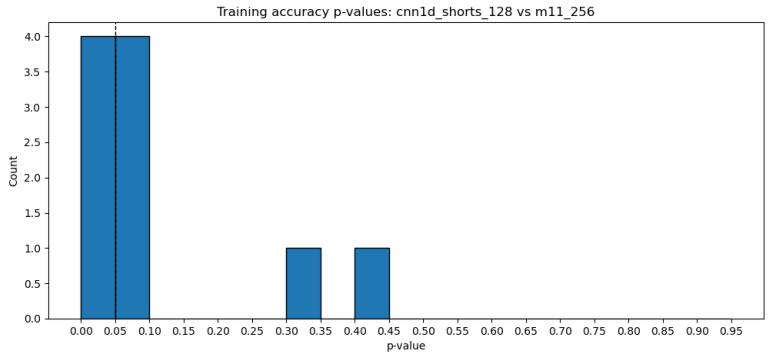


Fig. 6. Comparativa de los modelos M11_128 vs M11_256.

Repetir el experimento una suficiente cantidad de veces puede evidenciar una tendencia clara sobre la superioridad de un modelo en términos de la métrica deseada. Esto nos permitiría aseverar con mayor firmeza que "el modelo X es significativamente superior al modelo Y, en un % de las veces", o en caso contrario, como en el experimento IEMOCAP, que no existe diferencia significativa entre los modelos.

6. Conclusiones

El área de *Speech Emotion Recognition* viene siendo ampliamente estudiada, por lo que se pueden encontrar varios estudios de revisión y trabajos de recopilación, sobre el estado del arte, publicados recientemente. Estas investigaciones muestran la amplia variedad de modelos que se han propuesto para ésta tarea. Sin embargo, un número considerable de propuestas podrían no presentar resultados confiables, que permitan afirmar de manera objetiva que una propuesta es superior a otra.

Esto, entre otras razones, se debe a que las comparaciones con otras propuestas no toman en cuenta los elementos aleatorios que forman parte de los modelos de *Deep Learning*. El presente trabajo busca contribuir en la resolución de esta problemática, presentando una manera práctica de implementar evaluaciones comparativas de modelos de reconocimiento de emociones, mediante el uso de pruebas de significancia estadística.

Se demostró la utilidad de la propuesta, realizando la comparación del desempeño de tres modelos convolucionales entrenados con los dataset RAVDESS y IEMOCAP. Los resultados permitieron determinar el porcentaje de veces en los que los modelos evaluados obtuvieron un desempeño similar y el qué porcentaje de veces donde su desempeño tuvo una diferencia significativa.

Si bien se encontró que en la mayoría de los casos (entre 70 % y 80 %) los modelos tuvieron un desempeño similar, en el restante de los casos se encontró

una diferencia significativa. Esta discrepancia demuestra que si no se hubiese realizado la evaluación propuesta, podríamos llegar a cualquiera de las dos conclusiones (que existe diferencia o que son modelos similares). Con la propuesta presentada podemos identificar la tendencia en porcentajes de diferencia o similitud que tienen los modelos.

7. Trabajo futuro

En una próxima investigación se podrían realizar evaluaciones comparativas con modelos más complejos y bases de datos de mayor tamaño, teniendo en cuenta las restricciones impuestas por los tiempos de entrenamiento y la complejidad computacional. Estas restricciones podrían aliviarse distribuyendo el entrenamiento en múltiples nodos, lo cual permitiría la generación de más muestras que permitirían llegar a conclusiones más acertadas. Una limitación de esta investigación es que está sujeta a una sola estrategia de pre-procesamiento para todos los modelos.

La utilización de diferentes estrategias (p.ej., utilizar diferentes parámetros para generar el espectrograma) es una mejora interesante a ser implementada en trabajos futuros, ya que influye directamente en el desempeño de los modelos. Si bien la presente investigación está enfocada y ha sido realizada en un contexto del *Speech Emotion Recognition*, podría ser replicada en cualquier contexto donde se necesite comparar múltiples clasificadores multiclase.

Agradecimientos. Este artículo es parte de los resultados de la tesis de pregrado del Bachiller en Ciencias de la Computación Luis Bernal Chahuayo, la misma que ha sido financiada por el Proyecto Concytec - Banco Mundial "Mejoramiento y Ampliación de los Servicios del Sistema Nacional de Ciencia Tecnología e Innovación Tecnológica" 8682-PE, a través de su unidad ejecutora ProCiencia [Contrato número 014-2019-FONDECYT-BM-INC.INV].

Referencias

1. Akçay, M. B., Oğuz, K.: Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. Speech Communication, vol. 116, pp. 56–76 (2020)
2. Busso, C., Bulut, M., Lee, C. C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., Narayanan, S. S.: IEMOCAP: Interactive emotional dyadic motion capture database. Language resources and evaluation, vol. 42, no. 4, pp. 335–359 (2008)
3. Dhall, A., Kaur, A., Goecke, R., Gedeon, T.: Emotiw 2018: Audio-video, student engagement and group-level affect prediction. In: Proceedings of the 20th ACM International Conference on Multimodal Interaction. pp. 653–656 (2018)
4. Dzedzickis, A., Kaklauskas, A., Bucinskas, V.: Human emotion recognition: Review of sensors and methods. Sensors, vol. 20, no. 3, pp. 592 (2020)
5. Errica, F., Podda, M., Bacciu, D., Micheli, A.: A fair comparison of graph neural networks for graph classification. arXiv preprint arXiv:1912.09893, (2019)

6. Fayek, H. M., Lech, M., Cavedon, L.: Evaluating deep learning architectures for speech emotion recognition. Neural Networks, vol. 92, pp. 60–68 (2017)
7. Livingstone, S. R., Russo, F. A.: The Ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PloS one, vol. 13, no. 5, pp. e0196391 (2018)
8. Oh, S., Kim, D. K.: Comparative analysis of emotion classification based on facial expression and physiological signals using deep learning. Applied Sciences, vol. 12, no. 3 (2022) doi: 10.3390/app12031286.
9. Ringeval, F., Schuller, B., Valstar, M., Jaiswal, S., Marchi, E., Lalanne, D., Cowie, R., Pantic, M.: Avec 2015: The first affect recognition challenge bridging across audio, video, and physiological data. In: Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge. pp. 3–8 (2015)
10. Schuller, B., Valstar, M., Eyben, F., McKeown, G., Cowie, R., Pantic, M.: Avec 2011–the first international audio/visual emotion challenge. In: International Conference on Affective Computing and Intelligent Interaction. pp. 415–424. Springer (2011)
11. Schuller, B. W.: Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends. Communications of the ACM, vol. 61, no. 5, pp. 90–99 (2018)
12. Stapor, K., Ksieniewicz, P., García, S., Woźniak, M.: How to design the fair experimental classifier evaluation. Applied Soft Computing, vol. 104, pp. 107219 (2021)
13. Sun, Z., Yu, D., Fang, H., Yang, J., Qu, X., Zhang, J., Geng, C.: Are we evaluating rigorously? benchmarking recommendation for reproducible evaluation and fair comparison. In: Fourteenth ACM Conference on Recommender Systems. pp. 23–32 (2020)
14. Swain, M., Routray, A., Kabisatpathy, P.: Databases, features and classifiers for speech emotion recognition: a review. International Journal of Speech Technology, vol. 21, no. 1, pp. 93–120 (2018)
15. Whitehill, J.: Climbing the kaggle leaderboard by exploiting the log-loss oracle. In: Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence (2018)

ISSN 1870-4069

Caracterización de frases para un sistema conversacional inteligente en un entorno educativo virtual basado en las características de la dialéctica y los actos del habla

Bárbara María Esther García-Morales[1], María Lucila Morales-Rodríguez[1], Nelson Rangel-Valdez[1,2], Pedro Martín García-Vite[1], Juan Javier González-Barbosa[1]

[1] Tecnológico Nacional de México,
Instituto Tecnológico de Ciudad Madero,
México

[2] Consejo Nacional de Ciencia y Tecnología,
México

{G10070255, lucila.mr, nelson.rv,
pedro.gv, juan.gb}@cdmadero.tecnm.mx

Resumen. En este trabajo se presenta la caracterización de frases de un diálogo socrático para su integración en un agente conversacional, utilizando tres tipos de atributos principales para la caracterización: 1) teoría de los actos del habla en particular los actos ilocutivos y su fuerza, 2) los tipos de pregunta que caracterizan la técnica de la dialéctica y 3) los criterios de impacto de la información asociados a la importancia de la información. La estrategia para la caracterización utiliza la lingüística computacional para obtener el acto ilocutivo de una frase, así como el tipo de pregunta poderosa que representa y su impacto informativo. Los trabajos relacionados que abordan la caracterización de frases consideran como irrelevante conocer las definiciones de las palabras que puedan proporcionar la fuerza ilocutiva para establecer su clase ilocutiva a través de un diccionario. Lo que no permite generar frases que representen el razonamiento del agente, sino la visión de los autores. Esta caracterización de frases puede contribuir en el desarrollo de un modelo selector de frases en un coach virtual que brinde asesoramiento educativo basado en dialéctica socrática. El corpus de la caracterización resultante permite identificar la intención de cada una de las frases del diálogo caracterizando los actos del habla, lo que permitirá emitir una pregunta poderosa que concuerde con el contenido de la oración, produciendo que el asesorado pueda encontrar una respuesta a su problemática mediante la reflexión.

Palabras clave: Caracterización de frase, actos ilocutivos, pregunta poderosa directa, criterios de la información, agente conversacional.

Characterization of Sentences for an Intelligent Conversational Agent in a Virtual Educational Environment based on the Characteristics of Dialectics and Speech Acts

Abstract. In this work, the characterization of sentences of a Socratic dialogue is presented for their integration in a conversational agent, using three types of main attributes for the characterization: 1) theory of speech acts, in particular the illocutionary acts and their force, 2) the types questions that characterize the dialectic technique and 3) the information impact criteria associated with the importance of the information. The strategy for characterization uses computational linguistics to elicit the illocutionary act of a sentence, as well as the type of powerful question it represents and its informational impact. The related works that deal with the characterization of sentences consider it irrelevant to know the definitions of the words that can provide the illocutionary force to establish their illocutionary class through a dictionary. What does not allow them to generate sentences that represent the reasoning of the agent, but rather the vision of the authors. This characterization of phrases can contribute to the development of a phrase selector model in a virtual coach that provides educational advice based on Socratic dialectics. The corpus of the resulting characterization allows identifying the intention of each of the phrases of the dialogue characterizing the speech acts, which will allow to issue a powerful question that agrees with the content of the sentence, producing that the counselee can find an answer to his problem. through reflection.

Keywords: Characterization model, illocutionary acts, direct powerful question, information criteria, conversational agent.

1. Introducción

La *dialéctica socrática* es una técnica de enseñanza, la cual sirve para obtener el autoconocimiento a través de un diálogo entre dos individuos [1]. En este tipo de interacción, el asesor utiliza preguntas conocidas como *preguntas poderosas*. Estas preguntas se dividen en *mayéutica* e *ironía socrática* y se clasifican en siete tipos de *preguntas poderosas* [2]. En un *diálogo socrático* se involucran acciones que sirven para describir el enunciado, en este caso la *pregunta directa*.

Las frases del diálogo pueden ser caracterizadas mediante los *actos del habla*. Los cuales son: locución [3], ilocución y perlocución [4]. Caracterizar las acciones en un *diálogo socrático* también es posible a través del *impacto de la información* que transmiten las palabras contenidas en una conversación. El *impacto de la información* que perciben los individuos desde cualquier medio, reflejan de alguna manera su personalidad y preferencia en la retroalimentación de la información de acuerdo a un contexto específico [5].

Algunos criterios que podrían asistir en la tarea de la evaluación del *impacto de la información* con respecto a las preferencias pueden ser la relevancia que transmite la información según la percepción del individuo, la claridad que espera obtener al

consultar la información deseada, así como el *enriquecimiento léxico* que pueda proveer a su retroalimentación cultural y la fiabilidad que proporciona dicha información.

Los criterios mencionados anteriormente podrían funcionar para cuantificar una frase caracterizada utilizando los *actos del habla*, la *técnica de la dialéctica* y los *criterios de impacto de la información* con la finalidad de generar *preguntas poderosas directas* caracterizadas de acuerdo a la relación entre estos tres tipos de atributos.

Con esto se puede crear un *agente conversacional inteligente* en un *entorno educativo virtual.* En este trabajo se propone la *caracterización de frases* para un *sistema conversacional inteligente* en un entorno educativo virtual basado en las caracteristicas de la *dialéctica* y los *actos del habla*. El beneficio de caracterizar las locuciones basados en estos criterios es la identificación de la intención de la frase para poder estructurar una pregunta poderosa que pueda resolver el problema planteado.

2. Trabajos relacionados

En esta sección se presentan algunos trabajos relacionados que aplican la *teoría de los actos del habla* como estrategia para modelar las características de las frases incorporadas en los sistemas de gestión de diálogo en agentes socioemocionales. A pesar de que la lingüística computacional es la forma más recurrente en la literatura para poder establecer un análisis sintáctico y semántico a través del significado de cada una de las palabras que se encuentren en una oración, no aporta la suficiente profundidad, tal y como lo hace la intención y la fuerza en los *actos del habla* para una sola oración [6].

Un ejemplo de la interpretación de frases utilizando los actos del habla se encuentra en [7] donde se aplicó un modelo de comunicación basado en el lenguaje KQML (Knowledge Query and Manipulation Language). Este modelo utiliza la lingüística y los *actos del habla* para obtener la intención del *acto ilocutivo* de las frases establecidas en los mensajes. La manera de obtener la intención de las frases utilizando este lenguaje de comunicación es a través de un análisis sintáctico y semántico realizando consultas en un diccionario alojado en una base de datos, de donde se obtienen las definiciones de las palabras existentes en la frase.

En [8] se realiza una especificación de los *actos del habla* en un sistema multi-agente mediante el lenguaje de comunicación de agentes ACL (Agent communication language) para guiar sus metas y objetivos. Este lenguaje emplea un vocabulario (diccionario), un lenguaje de contenido (lenguaje de bajo nivel) y un lenguaje de comunicación (juego de comunicación y *actos del habla*).

El proceso para intercambiar mensajes entre agentes es por medio de operadores que establecen cual es el estado mental requerido para emplear un mensaje determinado y qué cambios se producirán al aplicar dicho mensaje.Los *actos ilocutivo* son establecidos por medio de operadores previamente registrados en la base de conocimientos, los cuales se encuentran en notación formal. Para poder establecer el operador correspondiente al mensaje, es necesario conocer el significado de las palabras por medio de un vocabulario general.

La estrategia para la *caracterización de frases* en el trabajo [9], está basada en la teoría de la relevancia de [10], en la cual se toma en cuenta la cognición humana y pertinencia del contexto para atribuir la intención comunicativa y su interpretación

(comunicación ostensivo-inferencial). Es decir, la frase del emisor puede acompañarse de un estímulo que puede comunicar un conjunto de hechos a un receptor para revelar la intención de su mensaje. En [9], los *actos del habla* son el medio para representar la intención comunicativa y para la actualización de metas e interacciones de un *agente inteligente*.

En este trabajo el diálogo es estructurado en dos niveles: el primer nivel está definido por los "juegos de comunicaciones" y el segundo nivel se define por las "fases" dentro de un juego de comunicación. Estos juegos de comunicación permiten la organización de los *actos de habla* que son expresados por los hablantes durante la interacción, caracterizando las frases mediante los actos perlocutivos, que se utilizan como objetivos y mediante los *actos ilocutivos* se define la estrategia utilizada durante el juego y las condiciones que lo cambian.

En [11], la teoría de los *actos del habla* forma parte de un *modelo de caracterización de frases* para un corpus de expresiones verbales. El proceso de caracterización es manual y determina cuales pueden ser los posibles criterios para cada frase del corpus. Para poder llevarlo se divide el proceso de caracterización en dos partes:

1. Caracterización de atributos o criterios,
2. Caracterización de valores.

Se analizaron los *actos del habla*, características del diálogo, del *agente* y del usuario, para identificar criterios funcionales en la caracterización de cada frase. Se identificó un valor nominal de intensidad para cada criterio. Posteriormente para convertir variables nominales a variables ordinales, se asigna un valor numérico a cada variable nominal dependiendo su grado de intensidad.

En los dos trabajos anteriores se expresa la visión de los autores limitada a su propio vocabulario, ya que consideraron irrelevante utilizar un diccionario para establecer la *clase ilocutiva* de una frase a través de las definiciones de las palabras que puedan proporcionar la *fuerza ilocutiva*. En este trabajo se utiliza la lingüística computacional como complemento para determinar la *fuerza ilocutiva* de los *actos del habla* y así poder crear una *caracterización de frases* para un *sistema conversacional* basado en *dialéctica*.

3. Caracterización de frases dichas en un diálogo socrático mediante los actos del habla y la diálectica socrática

La caracterización consiste en determinar los atributos o criterios particulares para algo o alguien. En este caso el proceso de caracterización identificará los posibles criterios para cada frase dicha por el *coachee* (asesorado). Para cumplir con el objetivo, se dividió el proceso de caracterización en dos partes:

a) La caracterización de atributos o criterios la cual busca determinar los atributos que distingan cada frase
b) La caracterización de valores el cual asigna valores numéricos a cada uno de los atributos.

La propuesta de caracterización de frases sugiere que toda frase se debe dividir en nueve grupos de atributos, cuatro de éstos son tomados de los *actos del habla*, cuatro de la *dialéctica y* uno de los *criterios de la información*. En esta sección se detallan los

Tabla 1. Ejemplo de coaching en la docencia utilizando la técnica de la dialéctica.

Actor	Frases o/y preguntas
Coach	Hola buen día, ¿En qué puedo ayudarle?
Coachee	Hola, buen día. Tengo problemas con un alumno en la materia de cálculo integral.
Coach	Muy bien, y ¿Cuál creé que sea el problema con el alumno?
Coachee	Posiblemente la falta de atención dentro del aula de clases.
Coach	Y, ¿Por qué creé que es la falta de atención en clase?
Coachee	Pues, siempre que llego al salón de clases, él está dormido.

compones teóricos detrás de los grupos de atributos y sus relaciones que permitirán caracterizar las frases en un *diálogo socrático*. Caracterización mediante la dialéctica socrática En un *diálogo socrático* se asiste a una persona por medio de una conversación basada en la *diálectica socrática*, la cual es una técnica de enseñanza que sirve para poder establecer el autoconocimiento en la persona que se le brinde la asesoría [1].

En la Tabla 1 se muestra la interacción entre dos personas por medio de un *diálogo socrático* en un proceso de coaching específicamente enfocado al área de la docencia. En este tipo de diálogos a la persona que da asistencia a la otra a través de una asesoría se le conoce como *coach* y a la persona que recibe la asesoría se le llama *coachee*.

El *diálogo socrático* planteado a continuación está basado en los ejemplos propuestos por [12], en la Tabla 1 se muestran solo algunas oraciones emitidas por los dos actores en este tipo de diálogo.

En el diálogo presentado en la Tabla 1 se establecen preguntas las cuales son llamadas *preguntas poderosas directas* [13]. Estas preguntas son nombradas de esta forma debido a que son preguntas abiertas, es decir que no esperan como respuesta un sí o un no. Las *preguntas poderosas directas* formuladas por parte del *coach* se clasifican en 7 tipos distintos [2].

La clasificación de estas preguntas es la siguiente: preguntas de aclaración, para razonar y argumentar, sobre asunciones, sobre causas y consecuencias, de origen, sobre perspectivas y sobre las propias preguntas. Estas *preguntas poderosas* pueden clasificarse en *mayéutica o ironía socrática*[14]. Los atributos que pueden ser utilizados para estructurar una pregunta poderosa en un diálogo socrático son: tipo de pregunta poderosa (TPP), tipo de pregunta basada en dialectica (TPBD), pronombre interrogativo mayéutica (PIM) tales como: ¿Cuál?, ¿Cómo) y pronombre interrogativo ironía (PII) tales como: ¿Por qué?, ¿Qué?

Cada una de las preguntas está relacionada con un pronombre interrogativo dependiendo del tipo del TPBD. El TPP sobre causas y consecuencias es la única que no puede aplicar mayéutica socrática, mientras que, el TPP aclarativa y de origen no pueden aplicar ironía socrática.

3.1. Caracterización mediante la teoría de los actos del habla

Para que el *coach* pueda ofrecer una *pregunta poderosa* en un *diálogo socrático*, es necesario conocer la intención de las frases expresadas por el *coachee*. Esto es posible

Tabla 2. Relación de atributos derivados de los actos del habla ilocutivos.

VI	VP	TCI	FI
Saludar	Saludar	Expresivo	Saludo
Agradecer	Agradecer	Expresivo	Agradecimiento
Despedir	Despedir	Expresivo	Despedida
Aprobar	Aprobar	Declarativo	Aprobación
Prometer	Prometer	Compromisivo	Promesa
Buscar	Buscar	Compromisivo	Examinar
Afirmar	afirmar	Asertivo	Afirmación
predecir	predecir	Asertivo	Adivinar
Admitir	admitir	Asertivo	Tolerar

a través de la caracterización de los atributos pertenecientes a los *actos del habla*. De acuerdo a [6], una oración principalmente se puede caracterizar por los tres *actos del habla* primarios (locución, ilocución y perlocución). Los *actos del habla ilocutivos* son una pieza importante para poder determinar cuál es el significado de lo que se percibe al escuchar o verificar una oración [15].

Los *actos ilocutivos* se clasifican en directos e indirectos [15], en los actos indirectos, el acto locutivo y el acto ilocutivo no coinciden, mientras que en los directos sí. Estos actos también pueden clasificarse en cinco clases distintas, las cuales son: asertivo, directivo, compromisivo, expresivo y declarativo. Los actos ilocutivos no los realizan las palabras sino los hablantes al emitir las palabras.

De acuerdo a [16] se establece que todos los actos tienen una *fuerza ilocutiva,* es decir, la fuerza comunicativa con la que el orador trasmite el enunciado. Los *actos del habla* pueden ser utilizados para caracterizar cualquier oración, obteniendo principalmente la intención de esa oración (frase) por medio de los *actos ilocutivos* para posteriormente determinar el efecto o interpretación que causa en la persona que recibe esa información (receptor) [16].

En un diálogo socrático generalmente las frases dichas por los dos actores son actos de *tipo directo* [18], por lo tanto, el significado del oyente es el mismo que la intención proporcionada por el hablante. En términos generales se puede decir que el verbo que ejecuta la acción dentro de la oración puede ser utilizando tanto en el *verbo ilocutivo* como en el *verbo perlocutivo*.

En la Tabla 2 se muestra la propuesta de los cuatro atributos derivados de la teoría de los *actos del habla* para caracterizar las frases en un *diálogo socrático* los cuales son: *verbo ilocutivo* (*VI*), *verbo perlocutivo* (*VP*), *tipo de clase ilocutiva* (*TCI*) y *fuerza ilocutiva* (*FI*). Las *fuerzas y los verbos ilocutivos* que se describen a continuación fueron identificados en un contexto educativo.

3.2. Cuantificación de los verbos ilocutivos de la teoría de los actos del habla con los criterios de la información

Caracterizar las acciones en un diálogo socrático también es posible a través del impacto de las palabras que contiene una frase en la conversación. Para conocer el

Tabla 3. Cuantificación de los verbos ilocutivos y los criterios de impacto.

VI	VP	TCI	FI	I	CI	EI
Saludar	Saludar	Expresivo	Saludo	0.3	0.3	0.7
Agradecer	Agradecer	Expresivo	Agradecimiento	0.3	0.3	0.7
Despedir	Despedir	Expresivo	Despedida	0.3	0.3	0.7
Aprobar	Aprobar	Declarativo	Aprobación	0.7	1	0.7
Prometer	Prometer	Compromisivo	Promesa	0.7	0.7	0.7
Buscar	Buscar	Compromisivo	Examinar	0.7	0.7	0.3
Afirmar	Afirmar	Asertivo	Afirmación	1	0.7	0.7
Predecir	Predecir	Asertivo	Adivinar	0.7	1	0.7
Admitir	Admitir	Asertivo	Tolerar	1	1	0.7

Tabla 4. Relación entre clases ilocutiva y los tipos de preguntas poderosas.

TCI	TPP
Asertivo	Preguntas aclarativas
Compromisivo	Preguntas de perspectiva
Directivo	Preguntas que requieren razón y pruebas
Declarativo	Preguntas sobre causas y consecuencias
Expresivo	Preguntas de origen

impacto de las palabras en una oración a través de los *actos del habla* se propone una cuantificación de los verbos pertenecientes a los actos ilocutivos. En este trabajo, los criterios impacto de la información en el diálogo están conformados por:

a) el criterio importancia de la información (*I*) representando la relevancia que el verbo ilocutivo puede aportar al contexto informativo,

b) el criterio claridad de la información (*CI*) específica el peso que el verbo aporta a la información,

c) el criterio enriquecimiento léxico (*EL*) ilustra la aportación léxica.

Estos criterios han sido evaluados en una escala que va de 0 a 1, donde 0 significa un impacto nulo de la frase, 0.33 un impacto bajo, 0.66 un impacto medio y 1 un alto impacto. Los valores nominales de impacto de la información han sido asignados de acuerdo con cada *verbo ilocutivo* en un *diálogo socrático* basado en la Tabla 2.

Por ejemplo, un saludo no tiene tanta relevancia como una afirmación, debido a que un saludo no revela información que ayude a la resolución de un problema. Por el contrario, una afirmación puede aportar más información en este tipo de interacción. En la Tabla 3 se muestra la cuantificación del impacto de los criterios de la información.

Si el valor de la importancia de la información es mayor a 0.33 y el tipo de *clase ilocutiva* es diferente a clase expresiva, es posible que el *coach* determine las características para estructurar una *pregunta poderosa* a partir de la intención de frase.

La Tabla 4 permite identificar la relación entre los *TPP* y las *TCI*, es decir, si el *TCI* de una frase es "**asertiva"** entonces el *coach* puede hacer una *pregunta poderosa directa* de tipo "**aclarativa"**. En la Tabla 5 se puede observar la relación entre los *VP* y los atributos *PIM y PII* los cuales son usados para caracterizar *el tipo de PPBD.*

Tabla 5. Relación entre pronombres interrogativos y verbos perlocutivos.

PPBD	Pronombre	VP
Mayéutica	¿Cuál? ¿Cómo?	Ofrecer, Prometer, Buscar, Predecir
Ironía	¿Por qué? ¿Qué?	Aprobar, afirmar, admitir

Tabla 6. Criterios que conforman la caracterización de frases mediante la técnica de la dialéctica y los actos del habla.

Grupos	Criterios	#Criterio
Tipo de Clase Ilocutiva:	Expresivo, Declarativo, Compromisivo, Directivo, Asertivo.	1 – 5
Tipo de Acto Ilocutivo:	Directo, Indirecto.	6 – 7
Fuerza Ilocutiva:	Saludo, Agradecimiento, Despedida, Aprobación, Preguntar, Ofrecimiento, Promesa, Examinar, Afirmación, Adivinar, Tolerar.	8 – 18
Verbo Perlocutivo:	Saludar, Agradecer, Despedir, Aprobar, Interrogar, Ofrecer, Prometer, Buscar, Admitir, Predecir, Afirmar.	19 – 29
Tipo de Pregunta Poderosa:	Preguntas Aclarativas, Preguntas de Perspectivas, Preguntas que Requieren Razón y Pruebas, Preguntas sobre causas y consecuencias, Preguntas de Origen.	30 – 34
Tipo de Pregunta Basada en la Dialéctica:	Ironía, Mayéutica.	35 – 36
Pronombre Interrogativo Mayéutica:	¿Cuál? ¿Cómo?	37 – 38
Pronombre Interrogativo Ironía:	¿Por qué? ¿Qué?	39 – 40
Impacto de la información en los actos del habla:	Importancia de la Información, Claridad de la Información, Enrequecimiento Léxico.	41 – 43

3.3. Atributos del corpus de frases utilizando la técnica de la dialéctica y los actos del habla

Para caracterizar frases en un *diálogo socrático*, se propone utilizar 43 criterios los cuales se encuentran divididos en 9 grupos (ver Tabla 6). Los primeros 29 criterios integran a cuatro grupos basados en los conceptos de la sección 3.2. Los criterios identificados en el rango del 30 al 40 comprenden 4 grupos de criterios pertenecientes a la técnica de la *dialéctica* presentados en la sección 3.1. Los criterios del 41 al 43 agrupan los criterios que cuantifican el *impacto de la información*.

Para poder caracterizar las frases mediante los *actos del habla* es necesario buscar la *FI* que emite la frase por parte del *coachee* para poder establecer a que *TCI*

Algoritmo 1. Caracterización de las clases ilocutivas.

Para cada clase ilocutiva *ci* = [“expresivo”, “declarativo” “directivo”, “compromisivo”, “asertivo”]
Num_clase_ilocutiva [ci] = 0
Para cada palabra *p* de la frase *f* que se caracteriza
fuerza_ilocutiva = Buscar_fuerza_ilocutiva(*p*)
Clase_ilocutiva = Buscar_Tipo_Clase_ilocutiva (*fuerza_ilocutiva*)
Im = Localizar_criterios_informacion (*Clase_ilocutiva*, Importancia_info)
Cl = Localizar_criterios_informacion (*Clase_ilocutiva*, Claridad_info)
En = Localizar_criterios_informacion (*Clase_ilocutiva*, Enriq)
Suma_criterios [*Clase_ilocutiva*] = acumular (*Clase_ilocutiva*, *Im, Cl, En*)
Num_clase_ilocutiva [Clase_ilocutiva] = Num_clase_ilocutiva [Clase_ilocutiva] + 1
Prom [Clase_ilocutiva] = Suma_criterios [Clase_ilocutiva] / Num_clase_ilocutiva
[Clase_ilocutiva]
Para cada *CRITERIO* i desde 1 hasta 5
CRITERIO [i] = 0
Si i = 1 entonces *Clase_ilocutiva* = “expresivo”
Si i = 2 entonces *Clase_ilocutiva* = “declarativo”
Si i = 3 entonces *Clase_ilocutiva* = “directivo”
Si i = 4 entonces *Clase_ilocutiva* = “compromisivo”
Si i = 5 entonces *Clase_ilocutiva* = “asertivo”
Si Prom [Clase_ilocutiva] > 0.66 AND Prom [Clase_ilocutiva] ≤ 1
entonces *CRITERIO* [i] = 1
Si Prom [Clase_ilocutiva] > 0.33 AND Prom [Clase_ilocutiva] ≤ 0.66
entonces *CRITERIO* [i] = 0.66
Si Prom [Clase_ilocutiva] > 0 AND Prom [Clase_ilocutiva] ≤ 0.33
entonces *CRITERIO* [i] = 0.33
Si *Prom* [*Clase_ilocutiva*] = 0 entonces *CRITERIO* [i] = 0

pertenece. Las palabras contenidas en la frase se someten a un análisis para identificar cuál de ellas es un verbo. Una vez identificado el verbo(s) en la frase, se emplea un diccionario [20] con la función de devolver el significado de las palabras.

Las palabras encontradas en la definición resultante del verbo se analizan para identificar las *FI* establecidas en la Tabla 3. Para determinar la caracterización de los valores de los atributos de los *TCI* se utilizan los *criterios de impacto de la información* para obtener un promedio de la existencia de cierto TCI, ya que, es posible que una frase tenga varios verbos que apunten a una misma *clase ilocutiva*. El valor de los primeros cinco criterios que representan las clases ilocutivas se determina con el siguiente algoritmo.

En la Tabla 7 se muestran dos ejemplos de frases caracterizadas mediante las *clases ilocutivas* (CI) y el *tipo de acto ilocutivo* (TAI) que van desde el criterio C_1 hasta el criterio C_7 definidos de acuerdo a la Tabla 6.

Tabla 7. Caracterización por medio las clases y actos ilocutivos.

	CI					TAI	
Frase caracterizada	c_1	c_2	c_3	c_4	c_5	c_6	c_7
Tengo problemas con un alumno en la materia de cálculo integral	0	0	0	0	1	1	0
Hola, tengo problemas en la materia de cálculo, poner atención es difícil	0.7	0	0.7	0	1	1	0

Tabla 8. Criterios seleccionados para representar las fuerzas ilocutivas.

VI	VP	TCI	FI	I	Cl	EL
Saludar	Saludar	Expresivo	Saludo	0.3	0.3	0.7
Prometer	Prometer	Compromisivo	Promesa	0.7	0.7	0.7
Buscar	Buscar	Compromisivo	Examinar	0.7	0.7	0.3
Predecir	Predecir	Asertivo	Adivinar	0.7	1	0.7

Tabla 9. Ejemplos de frases caracterizadas por medio las fuerzas ilocutivas.

	FI					
Frase caracterizada	c_8	c_{14}	c_{15}	c_{16}	c_{17}	c_{18}
Tengo problemas con un alumno en la materia de cálculo integral	0	0	0	0	0	1
Hola, tengo problemas en la materia de cálculo, poner atención es difícil	0.7	0.7	0.7	0	1	0

La primera frase "**Tengo** problemas con un alumno en la materia de cálculo integral", se asocia al verbo "**Tener**" e indica la confirmación de un hecho. La *FI* obtenida mediante las definiciones del diccionario es: "**Tolerar**".

El siguiente paso es obtener el peso de los criterios de impacto asociados a la *FI* (Tabla 3). En este ejemplo, los pesos asociados corresponden a $CI = 1$, $CL = 1$ y $EL = 0.66$. Como esta frase sólo tiene una *FI* asociada al criterio C_5 (asertiva), el algoritmo determinó que el valor del atributo sea 1, debido a que el promedio de los *criterios de la información* de es *TCI* se encuentra en el rango de 0.66 y 1.

La segunda frase "**Hola**, **tengo** problemas en la materia de cálculo, **poner** atención es difícil" se considera una frase compuesta debido a que contiene varias *FI* en una sola oración. Las *FI* encontradas en ésta oración son: "**saludo**", "**promesa**", "**examinar**" **y** "**adivinar**". A continuación se presenta la Tabla 8, la cual contiene los valores de los criterios de los *actos del habla* y sus *criterios de impacto de la información* de las *FI* identificadas. En este ejemplo, las cuatro *FI* pertenecen a las *TCI* identificadas con los criterios (C_1 expresivo, C_3 compromisivo y C_5 asertivo).

Como esta frase sólo tiene una *FI* asociada al criterio C_1 (Expresiva), el algoritmo determinó que el valor del atributo sea 0.66, debido a que el promedio de los *criterios de la información* de acuerdo al número de *TCI* se encuentra en el rango de 0.33 y 0.66. El valor del tercer criterio C_3 (compromisivo), resultó de calcular el promedio de los seis criterios de la información correspondientes a sus dos *FI*, determinándose que el valor del atributo sea 0.66, debido a que se encuentra en el rango de 0.33 y 0.66.

Tabla 10. Caracterización de frases mediante los tipos de preguntas poderosa.

	TPP				
Frase caracterizada	c_{30}	c_{31}	c_{32}	c_{33}	c_{34}
Tengo problemas con un alumno en la materia de cálculo integral	0	0	0	0	1
Hola, tengo problemas en la materia de cálculo, poner atención es difícil	0.7	0	0	0.7	1

Tabla 11. Caracterización mediante los criterios de la dialéctica socrática.

	PPBD		PIM		PII	
Frase caracterizada	c_{35}	c_{36}	c_{37}	c_{38}	c_{39}	c_{40}
Tengo problemas con un alumno en la materia de cálculo integral	0	1	0	0	1	1
Hola, tengo problemas en la materia de cálculo, poner atención es difícil	1	0	1	1	0	0

Una vez caracterizadas las clases ilocutivas se procede a determinar el grupo de atributos *TAI* (C_6 directo y C_7 indirecto). El valor nominal asignado en los criterios $C_6 = 1$ y $C_7 = 0$ ya que se considera un *acto ilocutivo directo*.

En la Tabla 9 se muestra la caracterización del siguiente grupo correspondientes a *FI* (C_8 a C_{18}), en esta tabla solo serán mostrados los criterios relacionados a los *TCI* de la Tabla 7.

En la segunda frase caracterizada, para la *FI* C_8 ("**saludo**") se utiliza el valor establecido en el criterio $C_1 = 0.66$, ya que su clase ilocutiva es expresiva con un valor de 0.66. Las *FI*, C_{14} y C_{15} ("**promesa**" y "**examinar**") pertenecen a los *TCI* compromisiva por lo que las dos *FI* adquieren el mismo valor establecido en el criterio $C_3 = 0.66$.

Finalmente la *Fl* C_{17} ("**adivinar**") deberá obtener el valor del *TCI* C_5 (asertiva) con un valor de 1. En la Tabla 10 se muestran las dos frases caracterizadas con los atributos pertenecientes al *TPP* que van de C_{30} a C_{34}.

Estos atributos toman directamente los valores nominales pertenecientes a los *TCI* expuestos en la Tabla 7 siguiendo la relación especificada de la Tabla 4. Por ejemplo, en la frase dos el valor de C_{30} (pregunta aclarativa) toma el valor de 0.66 porque según la Tabla 4 el tipo de *clase ilocutiva* es asertiva (C_3).

En la Tabla 11 se muestra un ejemplo de las frases caracterizadas con los atributos pertenecientes a la dialéctica socrática. Los valores nominales mostrados a continuación dependerán de la relación de los valores proporcionados en los *VP* de la Tabla 5, es decir, si el verbo es ofrecer, prometer, buscar o predecir, se puede hacer una PPBD, por lo tanto, el peso en todos estos atributos estará en 1.

Por ejemplo, la primera frase se caracteriza como *VP*: "admitir". Por lo tanto, el atributo el valor nominal es 1 para los criterios C_{36} (Ironía), C_{39} (¿Por qué?) y C_{40} (¿Qué?) ya que son los tipos *PPBD* y pronombres interrogativos asociado a ese tipo de verbo.

Tabla 12. Caracterización de frases del coachee con los actos del ilocutivos.

	CI			VP						
Frase	c_1	c_4	c_5	c_{19}	c_{20}	c_{25}	c_{26}	c_{27}	c_{28}	c_{29}
Hola, buen día. Tengo problemas con un alumno en la materia de cálculo integral.	x	x	w	x	x	z	x	w	w	x
Posiblemente la falta de atención dentro del aula de clases.	x	x	z	z	x	x	z	z	z	z

Tabla 13. Caracterización con atributos asociados a la dialéctica socrática.

Frase	TPP			TPBD		PIM		PII	
	c_{30}	c_{32}	c_{34}	c_{35}	c_{36}	c_{37}	c_{38}	c_{39}	c_{40}
Hola, buen día. Tengo problemas con un alumno en la materia de cálculo integral.	x	x	x	w	z	w	w	z	z
Posiblemente la falta de atención dentro del aula de clases.	z	x	x	w	z	w	w	z	z

4. Ejemplo de la caracterización de las frases del coachee

En esta sección se presentan los resultados obtenidos por la aplicación de la metodología. Las frases caracterizadas son una extracción de las frases enunciadas por el *coachee* en la Tabla 1. La carcaterización como se describió en la Tabla 6, está conformada por tres atributos principales:

1. Los actos del habla,
2. La dialéctica socrática,
3. Los criterios de impacto de la información.

La caracterización basada en los *actos del habla* quedó conformada en 4 grupos de atributos que integran a los 29 criterios, basados en los actos ilocutivos y perlocutivos vistos en la sección 3.2. En la Tabla 12 se muestran los atributos de las clases ilocutivas y los verbos perlocutivos que tienen una interrelación.

Los valores nominales de cada uno de los atributos oscilan entre el 0 y 1 como es mostrado en la sección 3.3. Por simplificación se muestran los valores nominales de la siguiente manera, $w=1$, $x= 0.66$, $y = 0.33$, $y\ z = 0$.

En la Tabla 13 se muestran los atributos seleccionados de la dialéctica socrática (TPP, TPBD, PIM y PII) con las CI y VP mostrados en la Tabla 12. Los TPP tienen una interrelación con las CI y los VP identifican el pronombre que puede ser usado para formular una PPBD con el PIM o el PII.

5. Conclusiones y trabajo a futuro

En esta investigación se presentó una *caracterización de frases* mediante atributos asociados a los *actos del habla* y la *dialéctica socrática*. La categorización de los

atributos fue establecida de acuerdo con la clasificación de los *actos ilocutivos* y los tipos de *preguntas poderosas* que pueden ser formuladas en los *diálogos socráticos*.

Para poder caracterizar las frases mediante los *actos ilocutivos* se utilizó un diccionario para obtener el significado de los verbos que fueron encontrados en las frases que se deseaban caracterizar. Posteriormente se realizó una búsqueda en la base de conocimiento para establecer los valores de los atributos correspondientes que fueron elegidos para la representación de la intención de la frase y así poder establecer los valores de los atributos correspondientes a la *dialéctica socrática*.

Esto puede ayudar a estructurar una *pregunta poderosa directa* generada por un *agente conversacional* basado en *dialéctica socrática* en un contexto eduactivo, ya que el *tipo de pronombre* asociado a la intención de la frase puede utilizarse para buscar preguntas directas que involucren los conceptos y los mismos *pronombres interrogativos* en una base de conocimientos. Esta caracterización permitirá modelar otros contextos con la misma interacción e intención de un *diálogo socrático*.

Como trabajos futuros, se espera el desarrollo de un modelo de selección de frases para un *coach virtual* que utiliza la técnica de la *dialéctica* a partir de las frases caracterizadas por esta propuesta para crear *preguntas poderosas*.

Dichas frases contribuirán en el asesoramiento educativo, donde el asesorado pueda encontrar de manera paulatina una respuesta a su problemática mediante la reflexión. El desarrollo de un juego serio enfocado a la capacitación que aplique una estrategia de solución basada en la dialéctica, puede ser un ejemplo de aplicación y podría tener un impacto directo en el aprendizaje del *coachee*.

Referencias

1. Castillero-Mimenza, O.: Método Socrático: qué es y cómo se aplica en la psicología (2018)
2. Sobrado, J. D.: Sócrates y la mayéutica: cómo hacer preguntas clave para facilitar el aprendizaje (2019)
3. Geis, M. L.: Speech Acts and Conversational Interaction. Cambridge University Press, Cambridge (1995)
4. Hatim, B., Mason, I.: Discourse and the Translator. Routledge (2014)
5. Castro-Rivera, J.: Modelado de la Personalidad en Modelos Preferenciales Multicriterio a través de Agentes Virtuales Inteligentes (2018)
6. Austin, J. L.: How to do things with words. Cambridge: Harvard University Press (1962)
7. Gómez, G. J., Zea-Restrepo, C. M.: Incorporación de Agentes Inteligentes en Ambientes de Aprendizaje (1998)
8. Aguirre, G. C.: Especificación de los actos del habla en sistemas multi-agente (2018)
9. Morales-Rodriguez, M. L.: Modèle d'interaction sociale pour des agents conversationnels animés. Application à la rééducation de patients cérébro-lésés (2007)
10. Sperber, D., Wilson, D.: La pertinence: communication et cognition. Les Editions de Minuit, Paris (1989)
11. Delgado-Hernández, X. S.: Aplicación de modelos de programación matemática a la selección de expresiones verbales en agentes virtuales socio-emocionales (2021)
12. Niewerburgh, C. V., Giráldez Hayes, A.: Coaching educativo. Colección: Didáctica y Desarrollo. Ediciones Paraninfo, S.A. (2016)
13. Habilidades del coach 2. Las preguntas poderosas.
14. De La Fuente-Morales, E.: Enseñanza de la matemática por la mayéutica. Praxis Investigativa ReDIE, vol. 9, no. 8 (2017)

15. Searle, J. R.: Expression and Meaning: Studies in the Theory of Speech Acts. Cambridge University Press, Cambridge (1979)
16. Searle, J. R.: Speech Acts: An Essay in the Philosophy of Language. Cambridge University Press (1969)
17. Lozano-Bachioqui, E.: La interpretación y los actos de habla. Mutatis Mutandis. vol. 3, pp. 333–348 (2010)
18. Borzi, C.: Actos del Habla Directos e Indirectos el Caso de la Pregunta (1999)
19. Delgado-Hernández, X. S., Morales-Rodríguez, M. L., Rangel-Valdez, N., Cruz-Reyes, L., Gómez-Santillán, C., González-Barbosa, J. J.: Analysis of Speech Acts for the Design of a Corpus of Phrases used in an Intelligent Learning Environment, vol. 9 (2019)
20. meetDeveloper: API de diccionario gratuito (2022)

ISSN 1870-4069

Red neuronal convolucional ramificada con atención para la mejora de voz

Noel Zacarias-Morales, José Adán Hernández-Nolasco,
Pablo Pancardo

Universidad Juárez Autónoma de Tabasco,
División Académica de Ciencias y Tecnologías de la Información,
México

{adan.hernandez,pablo.pancardo}@ujat.mx
201H18002@alumno.ujat.mx

Resumen. La mejora de la voz es un proceso que implica eliminar o atenuar el ruido presente en una señal de voz. En este sentido, muchos autores han empleado las redes neuronales para extraer la voz de manera inteligible y de calidad. A diferencia de los artículos que se revisaron, este trabajo propone una red convolucional ramificada con atención de múltiples encabezados que hace posible reducir el número de parámetros entrenables, así como incrementar la precisión en la extracción de la voz. Los resultados obtenidos demostraron que la incorporación del mecanismo de atención basado en múltiples encabezados mejoró en términos generales la capacidad del modelo convolucional ramificado, conforme a los valores de las métricas de calidad PESQ, STOI y SI-SDR. Los valores logrados confirman que incorporar un mecanismo de atención permite la mejora de los modelos de redes neuronales convolucionales ramificadas para extraer voz inteligible y de calidad.

Palabras clave: Red neuronal, convolución, atención, voz, ruido.

Branched Convolutional Neural Network with Attention for Voice Enhancement

Abstract. Speech enhancement is a process that involves removing or attenuating noise present in a speech signal. In this sense, many authors have used neural networks to extract the voice in an intelligible and quality way. Unlike the articles that were reviewed, this work proposes a branched convolutional network with attention to multiple headers that makes it possible to reduce the number of trainable parameters, as well as increase the precision in voice extraction. The results obtained showed that the incorporation of the attention mechanism based on multiple headers generally improved the capacity of the branched convolutional model, according to the values of the PESQ, STOI and SI-SDR quality

metrics. The values obtained confirm that incorporating an attention mechanism allows the improvement of branched convolutional neural network models to extract intelligible and quality voice.

Keywords: Neural network, convolution, attention, voice, noise.

1. Introducción

Un reto fundamental en la audición es escuchar selectivamente diferentes sonidos en una mezcla de señales acústicas. Es decir, la extracción de parámetros de una sola fuente de sonido es especialmente difícil en las grabaciones de un solo canal. El mejoramiento de la voz es la tarea de eliminar o atenuar el ruido añadido en una señal de voz, y generalmente se ocupa en mejorar la inteligibilidad y la calidad de la voz que sufre degradación por incluir ruido. El mejoramiento de la voz se emplea como procesado previo en aplicaciones como en el reconocimiento automático de la voz.

El propósito de la mejora de voz monocanal es proporcionar una solución al problema en el que se utilizan grabaciones hechas con un único micrófono. La mejora del habla monocanal se considera un problema muy difícil, ya que no se tienen pistas direccionales del origen de las distintas señales de audio que componen los ruidos presentes. En el mundo real, las señales de voz se ven fácilmente corrompidas por ruido. Los ruidos pueden agruparse en ruidos estacionarios (que no cambian en función del tiempo) y ruidos no estacionarios (que cambian cuando transcurre el tiempo).

Algunos ruidos que pertenecen a la categoría de no estacionarios son los ruidos de la calle, el ruido de un tren, el ruido de balbuceo (la voz de otras personas) y los sonidos de instrumentos musicales. Algunos que pertenecen a la categoría de estacionarios son los procedentes de acondicionadores de aire, ventiladores, compresores o bombas impulsoras. La relación entre la voz y el ruido en el dominio del tiempo puede escribirse como (1):

$$y(t) = x(t) + n(t), \tag{1}$$

donde *x(t)* es la señal de voz limpia y *n(t)* es el ruido añadido, dando como resultado que *y(t)* sea la señal de voz con ruido. Ahora bien, sea t el índice de tiempo, la señal puede representarse como $y = [y(1), \ldots, y(T)]$, donde T es la longitud del fragmento de audio. Al aplicar la transformada de Fourier de tiempo corto (STFT), podemos representar la señal acústica de (1) en el dominio de tiempo-frecuencia (TF) como (2):

$$Y(k,l) = X(k,l) + N(k,l), \tag{2}$$

donde k es el índice de la banda de frecuencias, l denota el índice de la trama temporal, *Y(k,l)*, *X(k,l)*, y *N(k,l)* son los coeficientes STFT de la señal de voz

ruidosa, la señal objetivo y la señal de ruido, respectivamente. Las definiciones anteriores son válidas únicamente para un micrófono de un solo canal.

En este caso, la tarea de mejora de voz tiene como objetivo recuperar la señal de voz objetivo x de la señal de voz ruidosa y [20]. Nuestra propuesta se basa en el mapeo del espectrograma de magnitud, y en este método basado en el mapeo, el objetivo de entrenamiento del modelo es mapear una función no lineal F desde la señal de voz con ruido *y(t)*, a una señal de voz limpia mejorada *x(t)*, como se escribe en (3):

$$y(t) \rightarrow^{F} x(t). \tag{3}$$

Debido a que existen problemas de variación rápida cuando se usa la señal de voz sin procesar (en el dominio del tiempo), el método basado en el mapeo se aplica habitualmente al espectrograma de magnitud de la señal de voz (dominio de la frecuencia), que se crea aplicando la transformada de Fourier de tiempo corto en una ventana temporal de un banco de filtros. Posteriormente, se realiza la operación inversa de la transformada de Fourier de tiempo corto para reconstruir el espectrograma de vuelta a la señal en el dominio del tiempo utilizando la información de fase de la señal de voz original con ruido.

Las redes neuronales basadas en el método de mapeo se entrenan para reconstruir los datos de salida a partir de los datos de entrada. Los datos de salida se obtienen de la señal de voz limpia *x(t)*, mientras que los datos de entrada se extraen de la señal de voz mezclada con ruido *y(t)*. En concreto, la red neuronal aprende una función F minimizando la pérdida del error cuadrático medio (MSE) entre el espectrograma de entrada y su entrada reconstruida, como en (4):

$$L_{MSE} = \|Y - F(X)\|^2, \tag{4}$$

o la pérdida de error medio absoluto (MAE) entre la entrada de la señal de voz y su entrada reconstruida, como en (5). Recientemente, se ha avanzado en la resolución de problemas de mejora de la voz en mezclas acústicas monocanal en escenarios cada vez más difíciles, gracias a los métodos de aprendizaje profundo:

$$L_{MAE} = \|Y - F(x)\|. \tag{5}$$

Un tipo de red neuronal muy utilizado en el problema de mejora de la voz es la red neuronal convolucional (CNN, por sus siglas en inglés), que tienen la capacidad de capturar patrones en los fotogramas vecinos mediante un conjunto de conexiones locales.

Se ha reportado de que las redes neuronales convolucionales son más eficaces que las redes neuronales perceptrón multicapa [1] y más eficiente que las redes neuronales recurrentes [11]. De los trabajos donde ha sido relevante el uso de las redes neuronales convolucionales destacan los siguientes.

En [11], Park & Lee demuestran que una red neuronal convolucional puede lograr un mejor rendimiento con una red 12 veces más pequeña que una red neuronal recurrente. La red neuronal convolucional es capaz de tratar las estructuras temporales y espectrales locales de la voz, por lo que es eficaz para separar los elementos de la voz y del ruido de las señales ruidosas.

Las redes neuronales convolucionales ha demostrado su eficacia para mejorar la voz tanto en el dominio de la frecuencia como en el del tiempo (forma de onda). Kinoshita et al. [5] emplearon la eliminación de ruido de la voz basada en la estimación del enmascaramiento utilizando una red neuronal convolucional. Este trabajo fue motivado por el éxito de las redes de convolución temporal para la separación de voz (Conv-TasNet) [19].

Ellos adaptaron la arquitectura de la red para la tarea de reducción de ruido de una señal, que se realiza tanto en el dominio temporal como en el de la frecuencia. En este trabajo los autores también investigaron la pérdida multitarea que predice dos salidas, la voz y el ruido. Además, se propuso una versión ampliada de la red neuronal convolucional usando una red residual (ResNet) [12], con la que se puede conseguir un mejor resultado, ya que la arquitectura de ResNet se ajusta a la tarea de la mejora de voz, que es reconstruir la señal de entrada eliminando la señal ruidosa residual.

La atención es un mecanismo cognitivo de procesamiento de señales de nuestro cerebro. Permite a nuestros cerebros captar eficazmente varias características informativas de los distintos estímulos sensoriales. La fusión de los modelos basados en el aprendizaje profundo y el mecanismo de atención ha ayudado a los modelos a enfatizar las características más informativas y suprimir las menos útiles.

Uno de los mecanismos de atención más utilizados recientemente es la atención de múltiples encabezados (del inglés Multi-Head Attention), que es un módulo que ejecuta varios mecanismos de atención en paralelo [18]. Las salidas de atención independientes se concatenan y se transforman linealmente en la dimensión esperada. Intuitivamente, los encabezados de atención múltiples permiten atender a partes de la secuencia de forma diferente, y se puede expresar como (6):

$$\text{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = [\text{head}_1, ..., \text{head}_h]W_0, \tag{6}$$

donde (7):

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V), \tag{7}$$

donde W son todas las matrices de parámetros entrenables. La atención de múltiples encabezados es un módulo que utiliza la atención de producto punto escalado, que es un mecanismo de atención en el que los productos de puntos se escalan de forma $\sqrt{d_k}$.

Formalmente tenemos una consulta $\boldsymbol{Q}$, una clave $\mathbf{K}$ y un valor $\mathbf{V}$, y calculamos la atención como (8):

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}})\mathbf{V}. \tag{8}$$

La figura 1 muestra la representación gráfica de la atención de producto-punto escalado y atención de múltiples encabezados, y los detalles se pueden consultar en [18]. Existen trabajos en los que se implementaron las redes neuronales convolucionales con mecanismos de atención exitosamente; por ejemplo, Sun

et al. [15] proponen una red neuronal convolucional recurrente que combine las ventajas de ambas, y optimizar aún más el rendimiento de la separación de una señal de voz con ruido mediante el uso de un mecanismo de atención.

Lan et al. [7] introducen un mecanismo de atención en un modelo con la arquitectura codificador-decodificador convolucional para enfatizar explícitamente la información útil, sus resultados experimentales mostraron que los mecanismos de atención que propusieron pueden emplear una pequeña fracción de parámetros para mejorar eficazmente el rendimiento de los modelos basados en redes neuronales convolucionales en comparación con sus versiones normales, además que su modelo logro generalizar bien a los ruidos no vistos durante el proceso de entrenamiento.

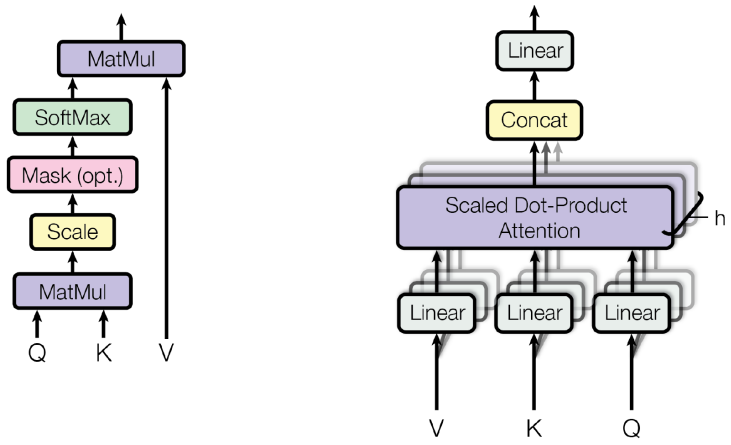


Fig. 1. (Izquierda) Atención de producto-punto escalado. (Derecha) La atención de múltiples encabezados consiste en varias capas de atención que funcionan en paralelo [18].

Motivados por estos trabajos, se propone una red convolucional ramificada con atención para solucionar el problema de la mejora de la voz en el dominio de la frecuencia. La propuesta se basa en una arquitectura convolucional con dos ramificaciones, un módulo de atención y capas densas.

El resto de este artículo se organiza como sigue. En la sección 2 se describe el modelo propuesto. La configuración experimental de los datos y el entrenamiento se presentan en la Sección 3. La sección 4 presenta los resultados obtenidos. Y la sección 5 concluye este trabajo.

2. Descripción del modelo

Se utilizó una red neuronal que se compone de una serie de capas convolucionales y densas, así como un módulo de atención. A continuación, primero se

Tabla 1. Hiperparámetros utilizados en el modelo implementado.

Tipo	Filtros	Nodos	Kernel	Activación	Strides	Dropout	Conexión
1D-Conv_01	64	-	16	PReLu	1	0.1	-
1D-Conv_02	64	-	16	PReLu	1	0.1	-
1D-Conv_03	32	-	16	PReLu	1	0.1	-
1D-Conv_04	2	-	16	PReLu	1	-	-
Rama_01							
1D-Conv_05	64	-	16	PReLu	1	0.1	1D-Conv_04
1D-Conv_06	64	-	16	PReLu	1	0.1	-
1D-Conv_07	32	-	16	PReLu	1	0.1	-
1D-Conv_08	1	-	16	PReLu	1	-	-
Rama_02							
1D-Conv_09	64	-	16	PReLu	1	0.1	1D-Conv_04
1D-Conv_10	64	-	16	PReLu	1	0.1	-
1D-Conv_11	32	-	16	PReLu	1	0.1	-
1D-Conv_12	1	-	16	PReLu	1	-	-
Modulo Atención	-	-	-	-	-	0.1	1D-Conv_08 1D-Conv_12
Dense	-	512	-	ReLu	-	-	-
Dense	-	512	-	ReLu	-	-	-
Lineal	-	256	-	-	-	-	-

describen las operaciones de convolución en la arquitectura ramificada y luego se describe el módulo de atención.

2.1. Red neuronal convolucional ramificada

La arquitectura general de la red neuronal convolucional ramificada (B-CNN, por sus siglas en inglés) se muestra en la figura 2, y los detalles de los hiperparametros utilizados se pueden consultar en la tabla 1. Se construyó un modelo convolucional ramificado que se puede dividir en cinco componentes:

1. Primeramente, la capa de entrada que alimenta este modelo con vectores de tamaño (256, 1).
2. A continuación, un primer bloque compuesto de 4 capas convolucionales con 64, 64, 32 y 2 filtros respectivamente, utilizando un tamaño de kernel = 16, stride = 1, padding = same, PReLu como función de activación. Se empleó dropout = 0.1 únicamente en las 3 primeras capas.
3. Posteriormente se encuentran las dos ramificaciones con configuraciones similares al primer bloque convolucional mencionado; cada una con 64, 64, 32 y 1 filtros respectivamente, usando un tamaño de kernel = 16, stride = 1, padding = same y PReLu como función de activación, y con dropout = 0.1, únicamente en las tres primeras capas.

4. Seguidamente se encuentra el módulo de atención, el cual recibe como datos de entrada dos vectores de tamaño (256,1) provenientes de las dos ramificaciones, los cuales son concatenados antes de ingresar en el módulo de atención.
5. Por último, se encuentran dos capas densas con 512 nodos cada una que emplean ReLu como función de activación, más una capa final de tipo lineal con 256 nodos que genera datos de salida de dimensión (256,).

La tabla 2 resume las dimensiones de los datos de salida de las capas sucesivas en la red propuesta, así como la cantidad de parámetros entrenables de cada capa. Cabe hacer mención que la última capa convolucional del primer bloque (1D-Conv_04) genera datos de salida de dimensión (256, 2).

Esto se debe a que estos datos son divididos para generar dos vectores de dimensiones (256, 1) cada uno, que sirven como datos de entrada para cada una de las dos ramas del modelo propuesto. El inicializador de la matriz de pesos del kernel de todas las capas convolucionales es el inicializador uniforme Glorot (también llamado inicializador uniforme Xavier).

Tabla 2. Dimensiones y cantidad de parametros entrenables del modelo implementado.

Capa (tipo)	Dimencion de salida	Cantidad de Parametros
Input	(None, 256, 1)	-
1D-Conv_01	(None, 256, 64)	17,472
1D-Conv_02	(None, 256, 64)	81,984
1D-Conv_03	(None, 256, 32)	40,992
1D-Conv_04	(None, 256, 2)	1,538
Rama_01		
1D-Conv_05	(None, 256, 64)	17,472
1D-Conv_06	(None, 256, 64)	81,984
1D-Conv_07	(None, 256, 32)	40,992
1D-Conv_08	(None, 256, 1)	769
Rama_02		
1D-Conv_09	(None, 256, 64)	17,472
1D-Conv_10	(None, 256, 64)	81,984
1D-Conv_11	(None, 256, 32)	40,992
1D-Conv_12	(None, 256, 1)	769
Modulo Atención	(None, 512)	273,420
Dense	(None, 512)	262,656
Dense	(None, 512)	262,656
Lineal	(None, 256)	131,328
Total de parámetros entrenables		1,354,480

El código del modelo utiliza internamente capas auxiliares de la librería Keras que no generan parámetros entrenables adicionales para el modelo, sino que

ayudan con la modificación de las dimensiones de los datos, estas capas auxiliares son:

1. Reshape: capa para separar la matriz de la capa Conv1D_04 de dimensión (None, 256, 2) en dos vectores de dimensión (None, 256, 1) utilizadas en cada una de las dos ramificaciones del modelo.
2. Concatenate: capa para unir los datos de salida de las dos ramificaciones con dimensión (None, 256, 1), en un vector con dimensión (None, 512, 1).
3. Flatten: capa para aplanar los datos de entrada de dimensión (None, 512, 1) a (None, 512).

2.2. Modulo de atención

Los métodos de aprendizaje profundo basados en mecanismos de atención alcanzaron el éxito en muchas tareas como la traducción automática [21], el reconocimiento de voz [6] y el procesamiento de imágenes [3]. Los mecanismos de atención son eficaces, ya que pueden ayudar al modelo a obtener mejores resultados mediante la identificación de características importantes. Considerando esto, se construyó e incorporó un módulo de atención para ayudar a identificar las características más importantes para mejorar el rendimiento del modelo convolucional ramificado.

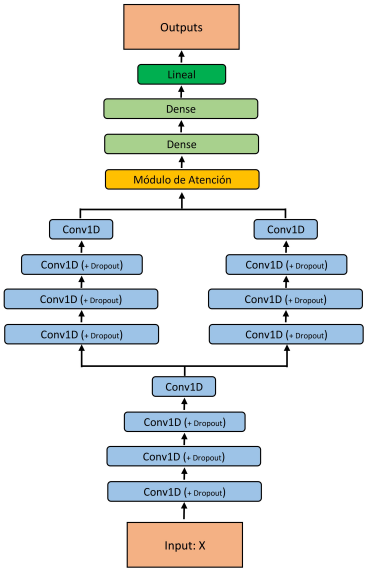


Fig. 2. Diagrama del modelo convolucional ramificado propuesto.

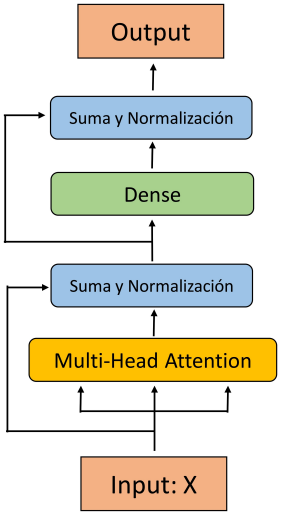


Fig. 3. Diagrama del modulo de atención.

El módulo de atención se basa en el uso de la atención de múltiples encabezados (multi-head attention), que ha resultado más eficaz que otros tipos

de atención. Por lo tanto, se integró la atención de múltiples encabezados en el modelo convolucional ramificado para identificar las características más relevantes. El módulo de atención está compuesto de las siguientes cuatro capas:

1. Una capa de atención de múltiples encabezados, con heads = 4 y dropout = 0,1.
2. Una capa de normalización, la cual suma los datos de salida de la capa de atención con los datos de entrada del módulo de atención, para despues aplicar la normalización con epsilon = 1e-6.
3. Una capa densa con 512 nodos y ReLu como función de activación.
4. Una capa de normalización, la cual suma los datos de salida de la capa densa con los datos de entrada de la capa densa, para despues aplicar la normalización con epsilon = 1e-6.

3. Configuración experimental

3.1. Datos

Se entrenó y evaluó el modelo propuesto realizando mezclas de audio con tres conjuntos de datos; como conjunto de datos de voz se utilizó el TIMIT [2], y como conjunto de datos de ruido se combinó NoiseX-92 [17] y DEMAND [16]. El TIMIT es un conjunto de datos que contiene grabaciones de enunciados de 630 hablantes que representan 8 divisiones dialectales del inglés americano, cada uno de ellos hablando 10 frases con diferente fonética de hablantes masculinos y femeninos.

El material del TIMIT está subdividido originalmente en porciones equilibradas para el entrenamiento y las pruebas (los criterios de subdivisión se describen en [2]). NoiseX-92 es un conjunto de datos compuesto de grabaciones de varios tipos de ruidos acústicos; entre ellos: ruidos de equipos de corte y soldadura eléctrica, ruido blanco, ruidos militares y de vehículos.

DEMAND contiene grabaciones de varios tipos de ruidos acústicos en entornos interiores (domésticos, oficina, público y transporte), y entornos al aire libre (calle y naturaleza). Para el conjunto de entrenamiento y validación se crearon cinco horas de mezclas de audio en clips de un minuto con SNRs uniformemente muestreados entre -10 dB y 10 dB (con lo que la señal de voz se corrompió con los diferentes ruidos).

Tanto los clips de voz como los de ruido fueron elegidos al azar. Posteriormente, se muestrearon los audios a 8 kHz para alimentar el modelo con las bandas de frecuencias más relevantes; se calculó el espectro de potencia de la magnitud de la señal utilizando la transformada de Fourier de corto tiempo (STFT) con un tamaño de 256 FFT, una ventana de longitud de trama de 32 ms (256 muestras), y con solapamiento del 50 % (128 muestras).

Por último, los datos de entrenamiento y validación se normalizaron a media cero y varianza unitaria para facilitar el proceso de entrenamiento. Para evaluar el modelo se crearon mezclas de audio con SNRs uniformemente muestreados de -10 dB, -5 dB, 0 dB, 5 dB y 10 dB. La fase de la señal únicamente se conservó

durante el proceso de predicción del modelo, para luego añadirla a la señal limpia estimada, de forma similar a lo mostrado en [9] y [8].

3.2. Métricas de evaluación

Para evaluar el desempeño del modelo propuesto se utilizaron como métricas de evaluación: la evaluación perceptiva de la calidad de la voz (PESQ) [13]; la inteligibilidad objetiva a corto plazo (STOI) [4]; y la relación señal-distorsión invariable en escala (SI-SNR) [14], que son las métricas estándares más utilizadas para evaluar el desempeño de las propuestas para el problema de mejora de voz. Los valores de PESQ oscilan entre -0.5 y 4.5 (cuanto más alto sea el valor, mejor será la calidad de voz); los valores de STOI suelen oscilar entre 0 y 1 (por lo general se convierte como un porcentaje de inteligibilidad).

3.3. Estrategia de entrenamiento

La estrategia de entrenamiento consistió en el mapeo del espectrograma de magnitud como el objetivo de entrenamiento. Se implementó el modelo convolucional ramificado utilizando la librería Keras con Tensorflow. La función de pérdida usada durante el proceso de entrenamiento fue el Error Cuadrático Medio (MSE), ya que el objetivo fue mejorar todas las métricas de evaluación, no una específica.

Se empleó Adam como optimizador con tasa de aprendizaje = 0.0001, b1 = 0.9, b2 = 0.999 y epsilon=1e-08. Se empleó un tamaño de lote de 64, y se utilizó el 10 % de los datos de entrenamiento para la validación, con el propósito de monitorear y controlar el rendimiento de la red y evitar el sobreajuste. Se eligió la precisión como métrica a monitorizar.

Tabla 3. Resultados de STOI (%), PESQ y SI-SDR de los modelos bajo diferentes ruidos.

Modelo	SNR	STOI (%)	PESQ	SI-SDR
B-CNN (sin atención)	-10	69.62	2.44	6.26
	-5	81.59	2.79	11.47
	0	89.85	3.14	15.96
	5	94.48	3.53	22.17
	10	97.38	3.79	22.90
B-CNN (con atención)	-10	69.71	2.45	6.30
	-5	81.76	2.78	11.51
	0	90.06	3.15	16.09
	5	95.43	3.54	24.20
	10	97.48	3.79	23.34

La duración del entrenamiento se estableció en 50 épocas; y se implementaron dos estrategias: una estrategia de detención anticipada (con lo que

el entrenamiento se detenía si después de 6 épocas consecutivas la métrica monitorizada dejaba de mejorar); así como una estrategia de reducción de la tasa de aprendizaje, con factor = 0.8 y una tasa de aprendizaje mínima = 0.00001 (aplicada cada época que la métrica monitorizada dejaba de mejorar).

4. Análisis y resultados

La tabla 3 muestra los resultados de las tres metricas estándar de mejora de la voz utilizadas habitualmente: la evaluación perceptiva de la calidad de la voz (PESQ), la inteligibilidad objetiva a corto plazo (STOI), y relación señal-distorsión invariable en escala (SI-SDR). Los resultados se basan en cinco niveles de SNR: -10 dB, -5 dB, 0 dB, 5 dB y 10 dB. La evaluación se realizó en el modelo convolucional ramificado con y sin el módulo de atención para contrastar el resultado del impacto del módulo de atención.

En los resultados de la tabla 3 se aprecia que la inclusión del módulo de atención basado en la atención de múltiples encabezados mejoró en términos generales la capacidad del modelo convolucional ramificado, con excepción del valor obtenido con PESQ en SNR de -5 dB. El audio con SNR de 5 dB fue el que mejor resultado mostró al incorporar el módulo de atención, pasando de 94.48 a 95.43 en STOI y de 22.17 a 24.20 en SI-SDR.

Respecto al entrenamiento de los modelos, las figuras (4) y (5) muestran las curvas de pérdida de los datos de entrenamiento y validación para los dos modelos (el modelo sin el módulo de atención y con el módulo de atención) con el fin de mostrar cómo la complejidad afectó al proceso de entrenamiento.

Aunque se estableció la duración del entrenamiento en 50 épocas, la estrategia de detención anticipada detuvo el entrenamiento en la época 38 en ambos modelos, ya que la fue en la época 32 en donde se alcanzó el valor más alto de la métrica monitorizada.

La estrategia de reducción de la tasa de aprendizaje también contribuyó en la rapidez de la convergencia de ambos modelos; en el caso del modelo sin el módulo de atención, la tasa de aprendizaje se modificó en 11 ocasiones, y en el caso del modelo con el módulo de atención, en 12 ocasiones.

Durante el entrenamiento se pudo observar que la incorporación de Dropout en ambos modelos (con y sin el módulo de atención) contribuyó significativamente a evitar el sobreajuste de ambos. Se identificó que, aunque ReLU es la función de activación más comúnmente utilizada; en este modelo, ReLU refleja un mejor rendimiento en las capas densas del modelo, mientras que PReLU es la función de activación con mejor rendimiento para las capas convolucionales, similar a lo mencionado por [10].

También se encontró que la normalización a media cero y varianza uno mejoró el proceso de entrenamiento de ambos modelos (con y sin el módulo de atención), esto es, la precisión mejoró alrededor de 10 % cuando se aplicó la normalización.

En cuanto a los datos utilizados en el proceso de entrenamiento de los modelos convolucionales ramificados, se identificó que se requiere de una gran cantidad de datos para predecir una mejor señal de voz limpia; y en este caso

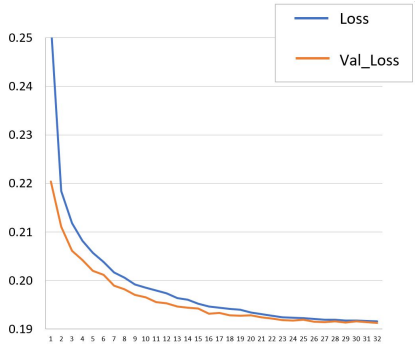


Fig. 4. Gráfica de la curva de pérdida de entrenamiento del modelo sin el módulo de atención para los datos de entrenamiento y validación.

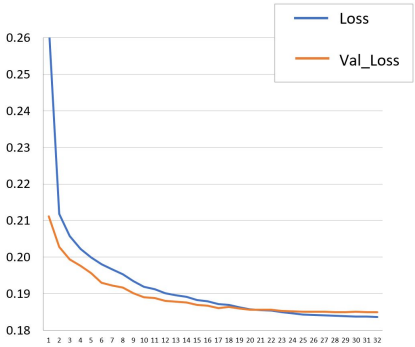


Fig. 5. Gráfica de la curva de pérdida de entrenamiento del modelo con el módulo de atención para los datos de entrenamiento y validación.

se identificó que aproximadamente cinco horas de audio fueron suficientes para que los modelos convergieran.

Usar más de cinco horas no mostró mejora en la velocidad de convergencia del entrenamiento ni en el incremento de su precisión, por lo menos no en estos modelos convolucionales ramificados con menos de 1.36 millones de parámetros entrenables.

5. Conclusiones

Aunque la mejora de la voz basada en el aprendizaje profundo ha demostrado ser muy eficiente al generar una señal de voz limpia con una calidad e inteligibilidad relativamente alta, todavía se considera que algunos entornos de ruido son muy difíciles de tratar para una red neuronal. En este trabajo se incorporó exitosamente un módulo de atención basado en la atención de múltiples encabezados en un modelo de red neuronal convolucional ramificado, el cual mostró mejoras en la tarea de atenuar y eliminar ruido de la voz, tal como se demuestra cuando se evalúa con las métricas STOI, PESQ, SI-SDR.

En un futuro se planea continuar con la investigación en mejora de voz, incorporando otras arquitecturas convolucionales como las codificador-decodificador, o modificar los datos para procesarlos como matrices en vez de vectores, e incluso incorporar una mayor variedad de señales de ruidos no estacionarios para permitir que el modelo se pueda evaluar en ambientes más complejos.

Agradecimientos. Los autores agradecen al Laboratorio Nacional de Supercómputo del Sureste de México (LNS), perteneciente al padrón de laboratorios nacionales CONACYT, por los recursos computacionales, el apoyo y la asistencia técnica brindados, a través del proyecto No. 202103086N.

Referencias

1. Fu, S., Tsao, Y., Lu, X., Kawai, H.: Raw waveform-based speech enhancement by fully convolutional networks. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, pp. 6–12 (2021)
2. Garofolo, J. S., Lamel, L .F., Fisher, W .M., Fiscus, J. G., Pallett, D. S., Dahlgren, N. L., Zue. V.: TIMIT acoustic-phonetic continuous speech corpus (1993)
3. Glaser, T., Ben-Baruch, E., Sharir, G., Zamir, N., Noy, A., Zelnik-Manor, L.: PETA: Photo albums event recognition using transformers attention (2021)
4. Jensen, J., Taal, C. H.: An algorithm for predicting the intelligibility of speech masked by modulated noise maskers. IEEE/ACM Trans. Audio, Speech and Lang. Proc., vol. 24, pp. 2009–2022, IEEE Press (2016)
5. Kinoshita, K., Ochiai, T., Delcroix, M., Nakatani, T.: Improving noise robust automatic speech recognition with single-channel time-domain enhancement network. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 7009–7013, IEEE (2020)
6. Kipyatkova, I.: End-to-end russian speech recognition models with multi-head attention. Speech and Computer, pp. 327–335, Springer International Publishing (2021)
7. Lan, T., Lyu, Y., Ye, W., Hui, G., Xu, Z., Liu, Q.: Combining multi-perspective attention mechanism with convolutional networks for monaural speech enhancement. IEEE Access, vol. 8, pp. 78979–78991 (2020)
8. Li, L., Lu, Z., Watzel, T., Kürzinger, L., Rigoll, G.: Light-weight self-attention augmented generative adversarial networks for speech enhancement. Electronics, vol. 10 (2021)
9. Nossier, S. A., Wall, J., Moniri, M., Glackin, C., Cannings, N.: An experimental analysis of deep learning architectures for supervised speech enhancement. Electronics, vol. 10 (2021)
10. Nossier, S., Wall, J., Moniri, M., Glackin, C., Cannings, N.: An experimental analysis of deep learning architectures for supervised speech enhancement. Electronics, vol. 10 (2021)
11. Park, S., Lee, J.: A fully convolutional neural network for speech enhancement. Proc. Interspeech, pp. 1993–1997 (2017)
12. Plantinga, P., Bagchi, D., Fosler-Lussier, E.: An exploration of mimic architectures for residual network based spectral mapping. IEEE Workshop on Spoken Language Technology (2018)

13. Rix, A. W., Beerends, J. G., Hollier, M. P., Hekstra, A. P.: Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In: IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 2, pp. 749–752 (2001)
14. Roux, J., Wisdom, S., Erdogan, H., Hershey, J. R.: SDR - half-baked or well done?. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 626–630 (2019)
15. Sun, C., Zhang, M., Wu, R., Lu, J., Xian, G., Yu, Q., Gong, X., Luo, R.: A convolutional recurrent neural network with attention framework for speech separation in monaural recordings. Scientific Reports, vol. 11, pp. 1–14 (2021)
16. Thiemann, J., Ito, N., Vincent, E.: The diverse environments multi-channel acoustic noise database (DEMAND): A database of multichannel environmental noise recordings. In: Proceedings of Meetings on Acoustics ICA2013, vol. 19 (2013)
17. Varga, A., Steeneken, H. J .M.: Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems. speech communication, vol. 12, pp. 247–251 (1993)
18. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, pp. 6000–6010 (2017)
19. Yi L., Nima M.: Conv-TasNet: Surpassing ideal time frequency magnitude masking for speech separation. IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, pp. 1256–1266 (2019)
20. Yuliani, A. R., Faizal, M., Suryawati, E., Ramdan, A., Ferdinandus, H.: Speech enhancement using deep learning methods: A review. Jurnal Elektronika dan Telekomunikasi, vol. 21, pp. 19–26 (2021)
21. Zhang, T., Huang, H., Feng, C., Cao, L.: Enlivening redundant heads in multi-head self-attention for machine translation. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 3238–3248, Association for Computational Linguistics (2021)

ISSN 1870-4069

Detección de nefropatía como complicación en pacientes diabéticos de tipo II mediante el uso de la regresión logística

Man Kit Liao-Li[1], José María Celaya-Padilla[1], Carlos E. Galván-Tejada[1], Jorge I. Galván-Tejada[1], Huizilopoztli Luna-García[1], Hamurabi Gamboa-Rosales[1], Miguel Cruz[2]

[1] Universidad Autónoma de Zacatecas,
Maestría en Ciencias del Procesamiento de la Información,
México

[2] Instituto Mexicano del Seguro Social (IMSS),
Centro Médico Nacional Siglo XXI,
México

{27805072, jose.celaya, ericgalvan, gatejo,
hlugar, hamurabigr}@uaz.edu.mx

Resumen. Datos de la Organización Mundial de la Salud (OMS) muestran que en 2019 la Diabetes fue la novena causa más importante de muertes en el mundo, un efecto común de esta enfermedad no controlada derivada de la hiperglucemia, es que con el tiempo va dañando gravemente muchos órganos y sistemas del cuerpo, como los nervios y vasos sanguíneos, dando paso al desarrollo de complicaciones más comunes como infartos de miocardio o accidentes cerebrovasculares, neuropatías, retinopatías e insuficiencia renal, siendo la última nuestro interés en particular. En el estudio se realizó un análisis sobre un conjunto de datos para determinar las variables más significativas y predecir el desarrollo de Nefropatía en pacientes con diagnóstico confirmado de Diabetes a fin de generar un modelo de clasificación utilizando Regresión Logística como algoritmo de aprendizaje automático.

Palabras clave: Diabetes, nefropatía, regresión logística, selección de características, predicción.

Detection of Nephropathy as a Complication in Diabetic Patients Type II Using Logistic Regression

Abstract. According to data from the World Health Organization (WHO) in 2019, Diabetes was the ninth leading cause of death in the world, a common

effect of this uncontrolled disease derived from hyperglycemia, which over time seriously damages many organs and systems of the body, especially the nerves and blood vessels, giving way to the development of more common complications such as myocardial infarctions or strokes, neuropathies, retinopathy and kidney chronic disease, the last of which is of interest to us in this article. This research performed an analysis on a data set to determine the most significant variables to predict the development of nephropathy in patients with a confirmed diagnosis of Diabetes in order to generate a model using logistic regression as a Machine Learning (ML) algorithm.

Keywords: Diabetes, kidney disease, logistic regression, feature selection, prediction.

1. Introducción

Según datos de la Organización Mundial de la Salud (OMS) el periodo comprendido entre 2000 y 2016, la mortalidad prematura (que comprende a personas antes de los 70 años) originada por Diabetes aumentó en un 5%, y en 2019 este padecimiento causó la muerte en 1.5 millones de personas ocupando la novena causa de defunciones en el mundo [1].

La Federación Internacional de la Diabetes (FID) estimó que para ese mismo año había cerca de 463 millones de casos confirmados con esta enfermedad y se espera que para 2045 las cifras lleguen a los 700 millones, en México para el 2018 según la Encuesta Nacional de Salud y Nutrición se tenía un registro de 82,767,605 sujetos mayores de 19 años con Diabetes, además la tasa de mortalidad pasó de 8.60 en 2016 a 11.95 por cada 10 mil habitantes en 2020.

Para este mismo año, la diabetes pasó a ser la tercera causa de muertes en el país con 151,019 casos en el que 52% fueron hombres y 48% mujeres [2]. La Diabetes Mellitus (DM) mal controlada causa microangiopatía, una afección en los pequeños vasos sanguíneos que contribuye a cambios patológicos y lesión de múltiples órganos produciendo principalmente daños en el sistema nervioso, ocular y renal.

Comúnmente, para diagnosticar en pacientes diabéticos si tienen daño renal se realiza un análisis de orina y medir sus niveles de albuminuria y creatinina con el fin de determinar el grado de filtración glomerular de los riñones

Las manifestaciones de un daño renal se detectan con estos parámetros, sin embargo, es importante considerar que no siempre se seguirá este patrón ya que solo se verá en pacientes con Diabetes Mellitus de tipo 1 (DM1), esta situación no es exactamente así en el caso de la Diabetes Mellitus tipo 2 (DM2) dado que muestran ausencia de albuminuria y a su vez experimentan un deterioro progresivo en la función renal [3].

Debido a las consecuencias a largo plazo que puede conducir esta patología y al impacto en las finanzas en los gobiernos nacionales e internacionales; la comunidad científica, médica y gubernamental han enfocado esfuerzos para reducir los índices de predisposición de este padecimiento en la población, tratando de diagnosticar prematuramente o en todo caso prevenirlo.

Esta investigación pretende impactar en el desarrollo de nuevas herramientas tecnológicas que contribuyan en esta necesidad, con las técnicas de Inteligencia Artificial es posible predecir con anticipación no solo a personas con altas probabilidades de contraer Diabetes si no, también en el desarrollo de complicaciones derivadas de esta enfermedad como Insuficiencia Renal dando la oportunidad de tratar con eficacia y antelación a dichos pacientes.

1.1. Trabajos relacionados

Recientemente la Inteligencia Artificial (IA) ha comenzado a incorporarse en la medicina para mejorar la atención de pacientes logrando una mayor precisión en el diagnóstico y abriendo camino a brindar una mejor atención médica en general. Existen proyectos en la actualidad dedicados a explorar las aplicaciones de la IA en la medicina, una de ellas es la asistencial que busca la prevención, diagnóstico, tratamiento y seguimiento de todo tipo enfermedades, la cual abordaremos con mayor énfasis en esta investigación [4].

En la siguiente sección se muestra un compendio de trabajos relacionados y colaboraciones sobre Nefropatía diabética para tener una perspectiva amplia de lo que se ha investigado en los últimos años. M. A. Makroum en su trabajo presentado en 2022 titulado "Machine Learning and Smart Devices for Diabetes Management: Systematic Review", presenta una revisión sistemática de investigaciones realizadas en los últimos 10 años, comparando propuestas de aplicaciones para el control de Diabetes que tienen aplicación tanto de escasas técnicas de IA como los que si tienen implementación de algoritmos de aprendizaje automático.

Todas las implementaciones están limitadas exclusivamente al control de Diabetes, no se incluye para el seguimiento y predicción de complicaciones derivadas de esta patología [5]. En 2021 P. Chittora presenta su trabajo en el artículo "Prediction of Chronic Kidney Disease - A machine learning perspective", una comparativa sobre 7 de diferentes modelos de clasificación: Redes Neuronales, C5.0, Regresión Logística, CHAID (Chi-square automatic interaction detection), LSVM, KNN y Random Forest. Genera en primera instancia un modelo con todas las características del conjunto de datos y posteriormente hace una selección de ellas por medio de diferentes métodos (filter, wrapper & embedded).

Evalúa cada algoritmo con las métricas de desempeño: Accuracy, Classification Error, Precisión, Recall, F-Measure, Coeficiente GINI, Curva ROC y Área bajo la Curva (AUC). El mejor modelo de clasificación fue LSVM con SMOTE (Sobre-muestreo de las clases minoritarias) con todas las variables del conjunto alcanzando un Accuracy de 98.86%, seguido del modelo con el método embebido (LASSO FS SMOTE) utilizado para la selección de características, obtuvo un Accuracy de 98.46%.

También identificó que el estudio se realiza en un conjunto de datos de la Universidad de California Irvine de su repositorio de Machine Learning, en los que incluyen confirmación de padecimientos como Diabetes e hipertensión y estudios de laboratorios que suelen ser necesarios para el control de pacientes con Enfermedad Crónica Renal (ERC).

Sin embargo, no incluyen medicamentos o tratamiento que sigue cada paciente. En la selección de características se limita a considerar solamente 6 más significativas de cada método [6]. M. A. Islam muestra en su artículo "Risk Factor Prediction of Chronic Kidney Disease Based on Machine Learning Algorithms" del 2021, la implementación de algoritmos de clasificación como: Regresión Logística, Random Forest y Naive Bayes. Establece a la hemoglobina como la característica fundamental para el modelo de predicción y la menos significativa a la hipertensión. El mejor modelo de clasificación fue Random Forest con un Accuracy de 98.88%. El conjunto de datos que se analizó solo se limitó a estudiar los casos ya confirmados con ERC.

No hay antecedentes de Diabetes ni de tratamiento que siguen los pacientes, solo cuenta con historial de hipertensión [7]. M. Vásquez en su Tesis de Maestría de 2019 propone un modelo basado en Redes Neuronales (RN) complementando el análisis con SVM y Random Forest. Pronostica el desarrollo de Enfermedad Renal Crónica a partir de registro de historial médico de pacientes, no se incluyen pruebas de laboratorio solo padecimientos comunes que desarrollan las personas con ERC entre ellos la Diabetes e hipertensión.

El mejor modelo fue Redes Neuronales con un AUC de 98%, aunque Random Forest reporta un comportamiento similar muestra más falsos positivos. En el estudio no hay información referente a tratamientos o medicamentos que toman los pacientes ni datos sobre muestras de laboratorio [8].

H. Polat en su trabajo de investigación de 2017 titulado "Diagnosis of Chronic Kidney Disease Based on Support Vector Machine (SVM) by Feature Selection Methods", realiza un modelo de clasificación en 5 diferentes perspectivas. El primero sin selección de características, y en las otras restantes con distintos métodos de selección con el fin de comparar las métricas de desempeño y el número de características significativas.

El mejor modelo de clasificación fue utilizando la selección por evaluador de subconjunto filtrado con un Accuracy del 98.5%. Su limitación es que utiliza el conjunto de datos de la Universidad de California Irvine que tiene el registro de 400 pacientes con 24 características para realizar el análisis, pero enfoca solamente a SVM como modelo clasificador [9].

Se puede observar que todas las propuestas se centran en comparar los algoritmos de predicción existentes, sin tomar en cuenta la importancia que tienen las características significativas resultantes, para determinar si el contexto del estudio es suficiente para abordar el tema en una perspectiva funcional.

Este articulo busca mejorar la evaluación del modelo de clasificación implementando validación cruzada, concepto que no se aborda en ninguna investigación anterior, y nos aseguran si las predicciones muestran algún sesgo o sobreajuste. Además, se pretende contribuir en el diagnóstico prematuro para disminuir el índice de incidencia de ERC en etapas terminales.

1.2. Objeto de estudio

De acuerdo con lo anterior el estudio busca predecir y diagnosticar prematuramente a pacientes con Diabetes a desarrollar potencialmente una Enfermedad Renal Crónica

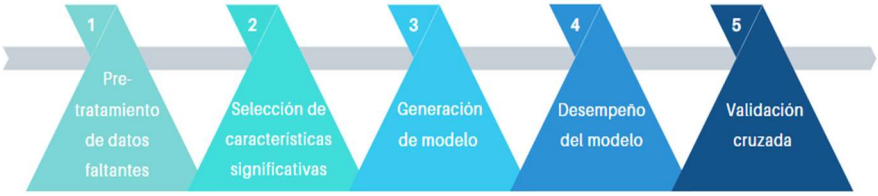


Fig. 1. Diagrama que describe el flujo del desarrollo de la investigación para obtener el modelo de predicción de ERC.

(ERC) utilizando algoritmos de aprendizaje automático presentando un modelo categórico con las características más significativas previamente seleccionadas, a partir de un análisis sobre un conjunto de datos proporcionada por el Centro Médico Nacional Siglo XXI del Instituto Mexicano del Seguro Social (IMSS).

2. Materiales y metodología

Esta investigación se realizó mediante de software estadístico R Studio desarrollado en diferentes etapas que se describirá a continuación. En primera instancia, se efectuó un tratamiento y limpieza a los datos, en segunda se abordará la selección de características significativas, en tercera tendremos la generación del modelo de clasificación implementando los algoritmos más comunes encontrados en la literatura como Regresión Logística, Random Forest y KNN obteniendo su eficiencia en la predicción con las métricas de desempeño tema que se contempla en la cuarta etapa, por último se evaluará el modelo con la validación cruzada con el fin evitar un sesgo en las predicciones y la precisión del modelo; todo este desarrollo propuesto se puede apreciar en el diagrama de flujo que se muestra en la Fig. 1.

2.1. Preprocesamiento

El conjunto de datos utilizado para este estudio, fue proporcionado por la Unidad de Investigación Médica en Bioquímica del Hospital de Especialidades "Bernardo Sepúlveda" del Centro Médico Nacional Siglo XXI del Instituto Mexicano del Seguro Social (IMSS) por un convenio de colaboración con la Universidad Autónoma de Zacatecas, es un conjunto de datos, contiene 46 características y 1787 pacientes con diagnósticos confirmado de Diabetes y personas sanas; tomando en consideración solo las características que implican médicamente en el desarrollo de cualquier complicación derivado de la Diabetes como lo es la edad, índice de masa corporal, así como niveles de creatinina, urea, glucosa, lípidos y colesterol en sangre.

De la misma manera se añadió al análisis los medicamentos más utilizados por los pacientes para el control de la Diabetes para observar su comportamiento en la participación de complicaciones.

Este conjunto de datos ha sido utilizado en investigaciones previas para abordar la predicción de Diabetes [10] el cual, presenta valores faltantes por lo que se le sometió a un proceso de imputación de datos con la librería de R missForest, este método

iterativo consiste en asignar valores inexistentes realizando múltiples árboles de decisión sobre las muestras del del conjunto de datos, los valores asignados son un promedio de las observaciones, obteniendo un error en las predicciones del 1.55%, corroborando este dato se obtuvo el promedio de la variable HDLU antes y después de la imputación con 44.23 y 44.21 respectivamente.

Nuestro conjunto de datos tiene además de la Insuficiencia Renal otras complicaciones como Retinopatía, Neuropatía, Cardiopatías Isquémicas, entre otros; y para este estudio solo se consideraron a pacientes Nefrópatas con antecedente de diabetes descartando las demás, obteniendo de esta manera un conjunto de datos con un total de 26 pacientes de los cuales 13 son casos (Nefropatía) y 13 de control (Diabetes sin complicación).

A pesar de que el dataset final es demasiado pequeño, este estudio buscar obtener un primer acercamiento en la predicción con técnicas de Inteligencia Artificial desde un contexto aplicativo en la sociedad mexicana, todo con la premisa de generar una herramienta que mejore las estrategias de detección en etapas iniciales de la ERC derivada de la Diabetes.

2.2. Selección de características significativas

Para determinar las características más significativas para el modelo de clasificación se implementó el proceso de selección hacia adelante y hacia atrás. En el caso del primer método el proceso comienza con un modelo vació que posteriormente va añadiendo variables significativas a partir de un criterio de similitud, el cual seguirá sumando términos al modelo hasta ya no encontrar variables que impacten en la respuesta del mismo.

En el segundo pasa lo contrario, parte de un modelo que tiene todas las características que pueden influir en la respuesta, y va eliminando términos menos influyentes dejando las variables más significativas, cabe resaltar que este procedimiento se realiza exclusivamente con Regresión Logística.

El criterio de similitud es la información de Akaike (AIC), un estimador muestral de la esperanza de log-verosimilitud, el cual se define en función de la máxima verosimilitud de las observaciones, cuanto más aumenta el número de parámetros aumenta el valor de AIC y el mejor modelo se va ajustando conforme este valor sea menor [11]. Con este proceso no se pretende identificar al modelo verdadero, sino al mejor de los modelos candidatos que se ajustan a las observaciones.

2.3. Generación de modelo de predicción

Una vez obtenido las características significativas se procede a generar el modelo de predicción implementando los algoritmos más usados en la literatura como Regresión Logística (RL), Random Forest y KNN.

De este proceso RL resultó ser el más sencillo y eficiente de implementar mientras los otros dos presentaban un sobre ajuste en las métricas de desempeño por lo que el estudio se centrará en este modelo, además de que resulta muy útil para los casos en que se desea predecir la presencia o ausencia de una característica según los valores de

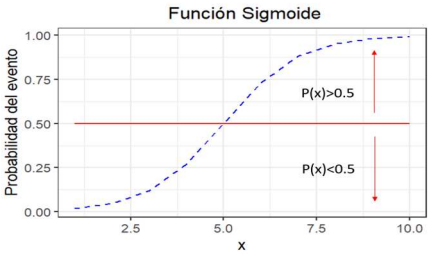


Fig. 2. Gráfica de la función sigmoide que describe al modelo de Regresión Logística.

un conjunto de predictores previamente analizadas y seleccionadas en donde se eliminan atributos poco relacionados con la salida.

Este algoritmo es similar al modelo de regresión lineal, pero está adaptado para modelos en que la variable dependiente es dicotómica o binaria [12]. Este modelo predice los eventos en probabilidades de ocurrencia de un suceso, en función del valor que toman las variables independientes de 0 a 1. Matemáticamente es posible utilizar el modelo de la regresión lineal de mínimos cuadrados, pero los valores extremos del predictor serían menores que 0 o mayores a 1.

Por ello es común utilizar una función que responda a estos requisitos que en la práctica suele ser sigmoide, dada por la ecuación 1. Si en la función anterior, le damos valores a z obtendremos una curva como la que se muestra en la Fig. 2, la cual es utilizada para clasificar 1 para cuándo y > 0.5 y 0 para y < 0.5:

$$p(x) = \frac{1}{1 + e^{-z}}. \tag{1}$$

Por esto, z viene siendo la representación de los coeficientes del modelo de regresión que después de realizar un desarrollo algebraico sobre la ecuación sigmoide obtenemos la expresión matemática que describe el modelo de Regresión Logística que se muestra en la ecuación 2 [13]:

$$p(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}. \tag{2}$$

2.4. Medidas de desempeño

Un algoritmo se evalúa para determinar si realizará una buena predicción, basándose en cómo funciona el modelo con nuevos y futuros datos, para ello se utilizan varias métricas que se emplearon en la investigación para validar de manera empírica el desempeño del modelo y que se describen a continuación:

- **Accuracy (Exactitud)**: Mide la fracción de predicciones correctas y se expresa debido a probabilidad de 0 a 1, entre más se acerque a 1, mayor será su exactitud predictiva [14].
- **Curva ROC:** La curva característica operativa del receptor (Receiver Operating Characteristic) es una representación de la sensibilidad frente a la

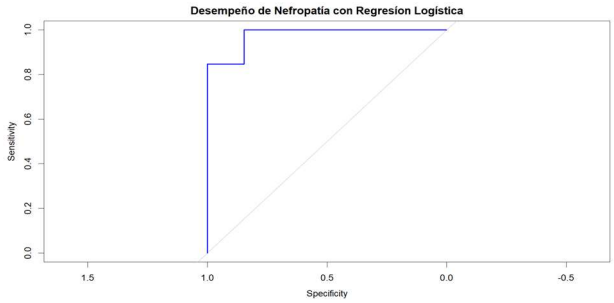


Fig. 3. Curva ROC del modelo de Regresión Logística con Creatinina, Insulina y Sexo.

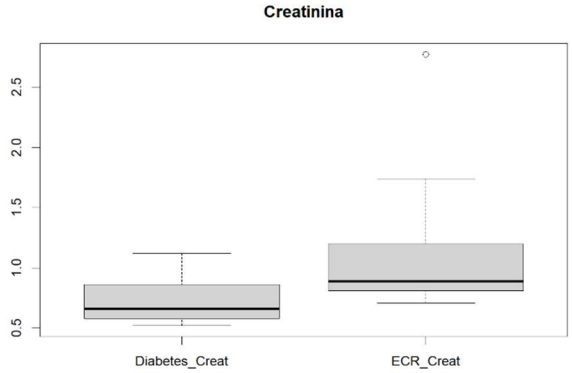


Fig. 4. Gráficas de caja y bigotes que describen la creatinina en pacientes diabéticos y los que tienen ERC.

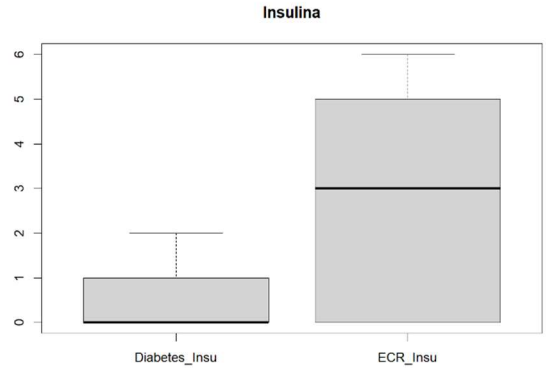


Fig. 5. Gráficas de caja y bigotes que describen el comportamiento de la Insulina en pacientes diabéticos y ERC.

especificidad, expresada como una probabilidad de predicción de un evento con los verdaderos positivos y los falsos positivos dada por el Área Bajo la Curva (AUC) establecida entre valores de 0 al 1.

- **Sensibilidad:** Se define como la capacidad de un algoritmo para predecir un resultado positivo verdaderos, es decir que tanto la predicción como el resultado real es positivo. Se dice que un algoritmo tiene una sensibilidad perfecta cuando predice que todo es positivo pase lo que pase. Por tanto, esta métrica no es suficiente para evaluar el desempeño del algoritmo.
- **Especificidad:** Esta métrica se define como la capacidad que tiene un algoritmo para predecir un resultado en positivo falso, es decir que teniendo el resultado real en negativo la predicción obtiene un positivo [15].

2.5. Validación cruzada

Para garantizar un buen modelo de clasificación es necesario asegurar que las predicciones sean precisas y evitar sobreajustes en sus resultados por ello la validación cruzada cumple un papel importante.

Este proceso consiste en dividir el conjunto de datos original en k subconjuntos (en nuestro caso k=3) en el que cada uno se les hará el entrenamiento y prueba con el modelo de predicción resultante, el proceso se repetirá k veces y en cada iteración se tomará un conjunto de prueba diferente.

Al finalizar las iteraciones se calculan las métricas de desempeño de cada subconjunto y se obtienen sus promedios, con esto nos da un estimado de la precisión del modelo sin un sobreajuste en la predicción [16].

3. Resultados

Para el análisis de predicción a desarrollar una ERC con precedentes de Diabetes se tomaron en cuenta las 42 características y al implementar la selección hacia adelante y hacia atrás se obtuvo un modelo con 32 variables de las cuales la mayoría no aportaba una significancia importante en el modelo.

Por lo que se procedió a realizar una reducción manual eliminando las variables menos significativas descritas por la prueba t de student, un parámetro que determina si las medias de dos grupos (características) tienen una diferencia importante [17], consiguiendo de esta manera solamente 3 características significativas (Creatinina, Insulina y Sexo) con un desempeño AUC de 97.63% cuya Curva ROC se muestra en la Fig. 3.

Creatinina e Insulina fueron las variables que más aportan en la predicción, implementando la primera característica en un modelo univariado con Regresión Logística nos da un AUC de 80.47%, se puede ver su comportamiento en las gráficas de caja y bigotes que muestran una presencia mayor en los pacientes con Enfermedad Renal Crónica (ERC) que en los diabéticos (ver Fig. 4).

Realizando el mismo procedimiento con Insulina, el AUC del modelo univariado fue de 77.51% se puede observar en la Fig. 5 las gráficas de caja muestran una diferencia considerable en los pacientes con ERC respecto a los diabéticos.

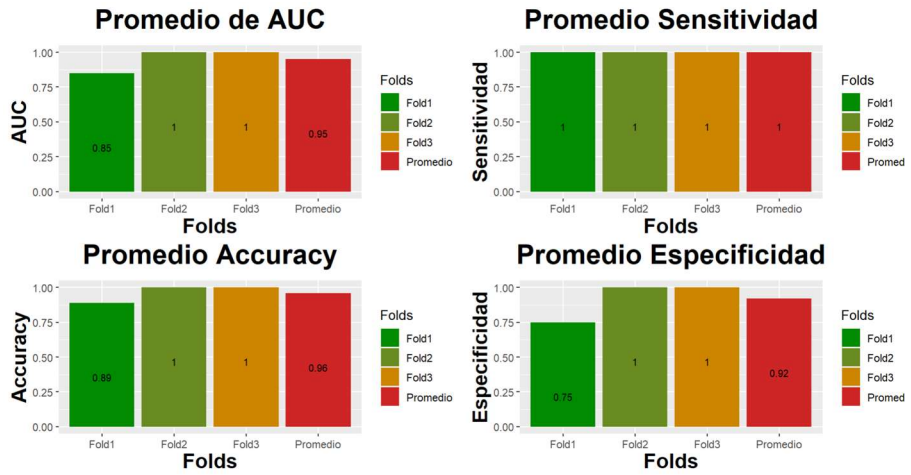


Fig. 6. Promedios generales de las métricas de desempeño en la validación cruzada.

3.1. Validación cruzada

Esta evaluación se realizó con la librería de Caret para hacer la separación del conjunto de 26 pacientes en subconjuntos k-folds (k=3) obteniendo así particiones diferentes de las observaciones originales y con ello realizar entrenamientos y pruebas del modelo propuesto, en este caso se dividió el data set en 3 particiones, en cada una se evaluaron las métricas de desempeño en prueba y se obtuvo el promedio general de cada métrica, mismas que se muestran en la Fig. 6.

4. Conclusiones y trabajo a futuro

Después de aplicar la metodología propuesta, esta generó un modelo conformado por Creatinina, Insulina y Sexo, el cual obtuvo un desempeño en la prueba de 97.63% para AUC, 92.31% en Accuracy, 84.62% de Especificidad y 100% en Sensibilidad; el nivel de precisión fue un tanto cercano a los mencionados anteriormente en la sección de trabajos relacionados presenta una ventaja, nuestro modelo tiene únicamente 3 variables en comparación al mejor modelo abordado en la sección 1.1 que utiliza 24 características.

Se encontró a la Creatinina por sí sola como una variable muy buena para la predicción de una ERC, en un modelo univariado esta característica representa más del 80% de eficiencia; aunque no es algo desconocido por la comunidad médica tener a la Creatinina como factor determinante en la detección de una ERC se puede afirmar que este parámetro resultó ser más significativa que la misma Urea (característica también presente en nuestro conjunto de datos), desecho que comúnmente también suele presentar niveles muy altos en pacientes con daño renal [18], además considerando la investigación médica referida al inicio de este artículo en donde describen que la

presencia de albuminuria como medida de diagnóstico no es constante en todos los pacientes con Diabetes de tipo II [2] y con los resultados obtenidos en este estudio, su propone un seguimiento constante de niveles de este residuo en el torrente sanguíneo como prioritario, más que la presencia de proteína en la orina, para detectar el nivel de riesgo a predisponer o en su caso diagnosticar una Nefropatía Diabética en pacientes con este tipo de Diabetes.

Este argumento se tendría que ser valorada y validada por un experto en el área para ser considerar como aportación a la comunidad médica y académica.

Otra variable importante encontrada fue el medicamento de Insulina el cual se complementa en el modelo de clasificación multivariado, y según la práctica médica este parámetro suele ser más alto en pacientes con Insuficiencia Renal, debido a este fármaco se metaboliza en el riñón y al existir una falla considerable en este órgano, la Insulina permanece más tiempo en el sistema, por ello, considerar esta variable permite tener una perspectiva complementaria en la predicción de este padecimiento.

Cabe mencionar que el conjunto de datos que se utilizó para el estudio no cuenta con niveles albuminuria para añadir en el análisis, sería de gran interés considerarla en un trabajo a futuro para observar su comportamiento con el modelo de predicción ya que es muy utilizada en la comunidad médica para diagnosticar a los pacientes con Enfermedad Renal Crónica seguido de la creatinina, así también incluir el tiempo de diagnóstico de Diabetes para ver la relación del daño paulatino de los vasos sanguíneos con la falla renal.

Otro aspecto importante para considerar es contar con un conjunto de datos más amplio de casos confirmados de ERC dado que en este estudio solo se tenía 13 pacientes positivos y 13 con solo Diabetes, contar con más observaciones ayudaría tener un modelo con mayor precisión. Es importante resaltar que este estudio es un pequeño acercamiento en la predicción de la Enfermedad Renal Crónica Diabética, en próximas investigaciones se pretende implementar otros algoritmos de predicción como Support Vector Machine (SVM) y Redes Neuronales, para comparar sus comportamientos respecto a los obtenidos con Regresión Logística y obtener un mejor modelo de clasificación.

Agradecimientos. Esta investigación pudo ser realizada gracias al apoyo otorgado por el Consejo Nacional de Ciencia y Tecnología al alumno Man Kit Liao-Li con el número de becario 901279 de la Maestría en Ciencias del Procesamiento de la Información, que sin su aporte el desarrollo sobre este tema en la comunidad científica no sería posible.

Referencias

1. Organización Mundial de la Salud: Diabetes. Centro de Prensa, Organización Mundial de la Salud (2021)
2. Estadísticas a Propósito del Día Mundial de la Diabetes. Datos Nacionales (2021)
3. Górriz, J., Terrádez, L.: Clínica y Anatomía Patológica de la Nefropatía Diabética 1. Clínica de la Enfermedad Renal Diabética Introducción, Valencia (2021)

4. Ávila-Tomás, J.F., Mayer-Pujadas, M.A., Quesada-Varela, V.J.: Artificial intelligence and its applications in medicine II: Current importance and practical applications Atencion Primaria, vol. 53, no. 1, pp. 81–88 (2021)
5. Makroum, M.A., Adda, M., Bouzouane, A., Ibrahim, H.: Machine Learning and Smart Devices for Diabetes Management: Systematic Review. Sensors, vol. 22, no. 5, p. 1843 (2022)
6. Chittora, P, et.al.: Prediction of Chronic Kidney Disease - A Machine Learning Perspective. IEEE Access, vol. 9, pp. 17312–17334 (2021)
7. Islam, M. A., Akter, S., Hossen, M.S., Keya, S.A., Tisha, S.A., Hossain, S.: Risk factor prediction of chronic kidney disease based on machine learning algorithms. In: Proceedings of the 3rd International Conference on Intelligent Sustainable Systems, ICISS 2020, pp. 952–957 (2020)
8. Vásquez, M. Gabriel, R.: Clasificador con redes neuronales para el pronóstico de la enfermedad renal crónica en la población colombiana (2022)
9. Polat, H., Danaei Mehr, H., Cetin, A.: Diagnosis of Chronic Kidney Disease Based on Support Vector Machine by Feature Selection Methods. Journal of Medical Systems, vol. 41, no. 4 (2017)
10. Alcalá-Rmz, V.: Identification of People with Diabetes Treatment through Lipids Profile Using Machine Learning Algorithms (2021)
11. González, A.: Selección de variables: una revisión de métodos existentes (2015)
12. IBM: Regresion Logística. SPSS Statistics (2021)
13. Páez, O., Sangrador, O., Arias, M.: Regresión logística binaria simple (2022)
14. Amazon Web Services: Amazon Machine Learning Guía para desarrolladores (2016)
15. Irizarry, R. A.: Introducción a la ciencia de los datos. pp. 529–582 (2021)
16. Laura-Ochoa, L.: Evaluation of classification algorithms using cross validation. In: Proceedings of the LACCEI international Multi-conference for Engineering, Education and Technology, vol. 2019 (2019)
17. Portal de formación estadística: La prueba t (2022)
18. Salabert, E.: Urea alta: causas, síntomas, y como bajar sus niveles (2018)

ISSN 1870-4069

Detección y clasificación de neumonía en imágenes de Rayos X usando técnicas de preprocesamiento y Deep Learning

Victor Hugo Galindo-Ramirez, José Agustín Almaraz-Damian, Clara Cruz-Ramos, Volodymyr Ponomaryov, Rogelio Reyes-Reyes

Instituto Politécnico Nacional, México

vgalindor1601@alumno.ipn.mx, vponomar@ipn.mx

Resumen. La radiografía de tórax es una herramienta utilizada en el diagnóstico médico, este tipo de imágenes se destacan por ser accesibles para la comunidad, proporcionando suficiente información de la región del tórax, para ser evaluadas por los especialistas, en caso de presentar alguna enfermedad. Además, debido a su simplicidad, la radiografía de tórax es la mejor opción contra las imágenes de tomografía computarizada (CT), ultrasonido (US) o resonancia magnética (MRI) en pacientes pediátricos. En este trabajo se propone un esquema de preprocesamiento el cual elimina las regiones de la imagen de Rayos X que no pertenecen a la caja torácica, como etiquetas generadas al realizar el estudio, entre otras. Después de procesar las antes mencionadas las imágenes, se procede a entrenar un modelo basado en la arquitectura de las Redes Neuronales Convolucionales. Finalmente, el sistema diseñado emplea el algoritmo Grad-CAM con el fin de proporcionar una imagen que contenga la representación perceptual de las características relevantes que fueron obtenidas para cada clase. El sistema ha demostrado un rendimiento similar en comparación con los métodos más avanzados, empleando métricas de calidad como Exactitud (Accuracy), Precisión, Sensibilidad y Métrica-F1.

Palabras clave: Clasificación, rayos x, aprendizaje profundo, neumonía, CNN.

Detection and Classification of Pneumonia in X-Ray Images Using Preprocessing Techniques and Deep Learning

Abstract. Chest X-ray imaging is a tool used in medical diagnosis, this type of image stands out against computed tomography (CT), ultrasound (US), or magnetic resonance imaging (MRI) in paediatric patients. Moreover, is accessible to the community and provides enough

information about the chest region to be evaluated by physicians. In this work, a preprocessing scheme is proposed to eliminate regions of the X-Ray image that do not belong to the thorax area, also known as the rib cage. This region also contains objects, such as labels, generated when the study is performed. After processing the aforementioned images, we proceed to train two models based on Convolutional Neural Network architectures. Finally, the designed system uses the Grad-CAM algorithm to provide an image that includes the perceptual representation of the relevant features obtained for each class. The system has shown similar performance compared with State-Of-The-Art methods, using quality metrics such as Accuracy, Precision, Sensibility, and F1-Score. Detection and Classification of Pneumonia in X-Ray images using preprocessing techniques and Deep Learning.

Keywords: Classification, x-ray, deep learning, pneumonia, CNN.

1. Introducción

En la actualidad, la tasa de mortalidad por Neumonía ha ido en aumento, principalmente en niños menores de cinco años, además la Organización Mundial de la Salud (OMS) estima que se reportan alrededor de 156 millones de casos en todo el mundo por año [10]. En México, el Instituto Nacional de Salud Pública (INSP) informó que se diagnosticaron 117 mil casos y 21 mil personas fallecieron por neumonía en los años 2017 y 2018 [31].

El Instituto Nacional de Estadística y Geografía (INEGI) en la Figura 1 presenta un estudio de las defunciones por neumonía por cada 10 mil habitantes entre los meses de Enero-Agosto (2011-2020) y de los años (2011-2019) donde el total de las muertes por influenza y neumonía que ascendieron a 29 mil casos, el 99 % es representado por defunciones causadas por la neumonía [4].

Los sistemas asistidos por computadora (CAD) se utilizan para mejorar, detectar y extraer una Región de Interés (ROI) de la imagen digitalizada de un estudio clínico. Estos sistemas pueden ayudar a los médicos radiólogos especialistas a analizar dicha ROI de una manera rápida y precisa, reduciendo el tiempo de procesamiento y mejorando el tiempo de respuesta durante el tratamiento de un paciente.

El aprendizaje máquina o mejor conocido como Machine Learning [30] tiene el objetivo de desarrollar sistemas que aprendan a reconocer patrones automáticamente, por otro lado, el aprendizaje profundo o Deep Learning [12], es empleado para desarrollar sistemas computacionales inteligentes que demandan una gran cantidad de datos con el fin de clasificar imágenes.

Los sistemas CAD requieren de métodos que logren proporcionar ofrecer segunda opinión al especialista y poder emitir un diagnóstico. Dependiendo del objetivo del sistema, las técnicas de aprendizaje automático requieren características de tipo *handcraft*, las cuales están basadas en patrones perceptuales, estadísticos o médicos. Por el contrario, las técnicas de aprendizaje profundo

utilizan la información contenida en los datos ingresados, lo que permite extraer y aprender características que no se caracterizan como perceptuales, pero son relevantes para determinar si la imagen pertenece a una clase o no.

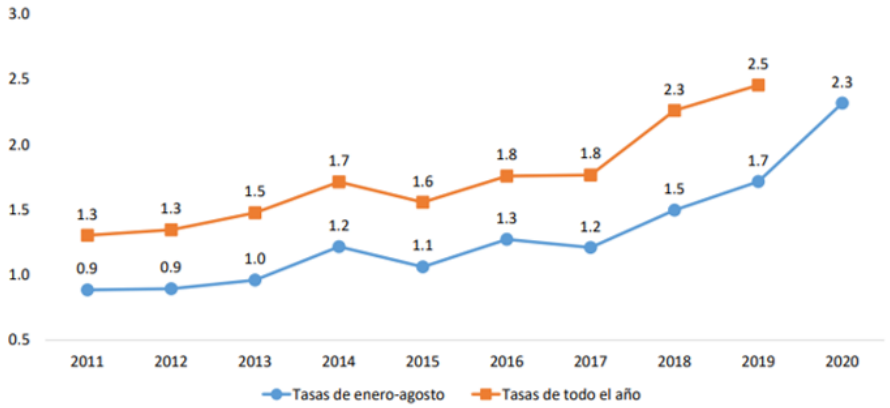


Fig. 1. Tasa de defunciones por influenza y neumonía por cada 10 000 habitantes Enero-Agosto (2011-2020) y cierre de año (2011-2019) obtenido de [4].

2. Estado del arte

El sistema propuesto por Jiang et al. [14] presenta un sistema el cual clasifica imágenes de Rayos X como entrada, para las arquitecturas InceptionResNetV2, Xception, DenseNet201, VGG19, el autor reporta que se obtiene el mejor resultado utilizando la arquitectura InceptionResNetV2 con un 94 % de exactitud.

Sánchez et al.[23] proponen un sistema el cual clasifica neumonía en imágenes de Rayos X, utilizando la arquitectura Xception y un subespacio basado en el análisis de componentes principales (PCA) el cual incrementa la exactitud a partir de pequeños dataset de entrenamiento, los autores reportan una exactitud del 96 % utilizando únicamente 600 imágenes.

Lujan et al.[17] desarrollan un sistema CAD basado en la arquitectura Xception empleando pesos previamente entrenados con el dataset de ImageNet Large Scale Visual Recognition Challenge (ILSVRC), mejor conocido como Imagenet [22]. Además, utilizan una técnica de preprocesamiento, la cual consiste en eliminar los bordes negros de las imágenes de Rayos X para mantener la mayor cantidad de información de la ROI.

Varela et al. [28] proponen un sistema que mediante una red neuronal artificial clasifica imágenes de Rayos X y de tomografía computarizada en tres clases: COVID-19, Neumonía y Normal, se extraen 129 características de tipo handcraft por cada una de las 750 imágenes, finalmente los autores reportan una exactitud del 97 %.

El sistema propuesto por Varela et al.[29] clasifica Neumonía utilizando una segmentación que por medio de algoritmos de tipo handcraft elimina regiones de la imagen quedando únicamente con ambos pulmones, posteriormente se utiliza una red neuronal modular, que consiste en entrenar y clasificar 40 características divididas en: histograma, Matriz de co-ocurrencia de niveles de grises (GLCM) y patrones locales binarios (LBP) respectivamente, la clasificación final es promediada de acuerdo con cada salida obtenida por los 3 tipos de características.

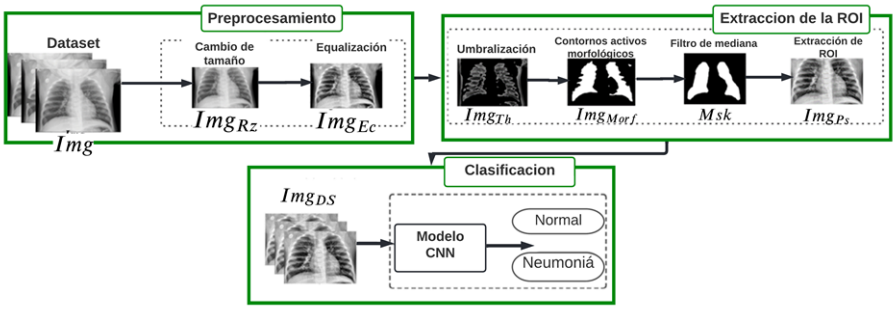


Fig. 2. Diagrama a bloques del sistema propuesto.

En este trabajo se presenta un sistema CAD de Rayos X para la detección de Neumonía, en el cual, se utiliza una técnica de procesamiento diseñada para extraer la región pulmonar con el fin de reforzar la extracción de patrones pertenecientes a la ROI. Posteriormente, la Imagen de la ROI se empleará para entrenar la arquitectura de Red Neuronal Convolucional (CNN) Xception.

El sistema CAD propuesto se implementó en una PC con una GPU Nvidia®RTX 3090. Los experimentos emplearon el dataset 'Imágenes de Rayos X de tórax para la clasificación' propuesto en [15], se evalúa el rendimiento del sistema propuesto empleando las métricas de calidad: exactitud, especificidad, sensibilidad y métrica-F1 frente a los sistemas consultados en el Estado del Arte.

3. Sistema propuesto

El sistema CAD propuesto se conjunta de tres etapas principales: (a) Preprocesamiento, (b) Extracción de la ROI y (c) Clasificación. En la primera etapa se redimensionan las imágenes y se aplica una mejora de contraste.

En la segunda etapa, la caja torácica se segmenta y se extrae mediante el uso del algoritmo de contornos activos morfológicos y el algoritmo bounding box respectivamente. Finalmente, en la tercera etapa la imagen ROI es procesada por la Arquitectura CNN elegida, cuyas características son extraídas y aprendidas para clasificar imágenes Normales y de Neumonía. El diagrama a bloques conceptual del sistema CAD diseñado se presenta en la Figura 2.

3.1. Dataset Chest X Ray Images

El dataset proporcionado por la Universidad de California [15] contiene 5232 imágenes de Rayos X de tórax, las cuales se encuentran divididas en dos clases; neumonía y normal, proponiendo 234 imágenes normales y 390 de neumonía, las cuales son utilizadas para validar el desempeño del sistema.

En la Figura 3 se muestran dos imágenes de Rayos X, donde una de ellas contiene patrones radiológicos que se presentan en la neumonía, los cuales son: neumonía lobar, bronconeumonía, neumonía intersticial, neumonía redonda y neumonía por aspiración [21], por otro lado, la imagen normal no presenta patrones radiológicos.

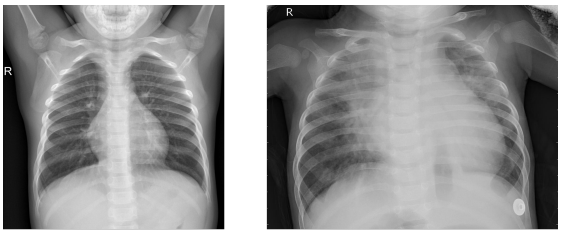


Fig. 3. Muestras aleatorias del dataset para la clase: A) Imagen normal; B) Imagen de neumonía.

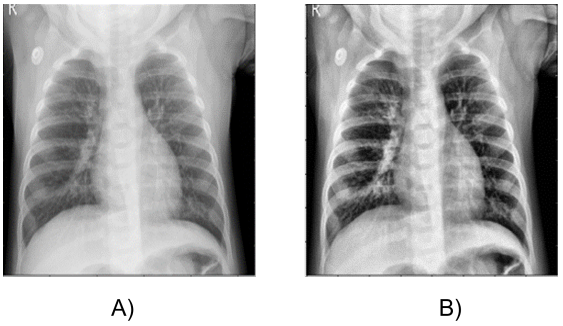


Fig. 4. Muestra aleatoria de la etapa de preprocesamiento: A) Imagen Original; B) Imagen Ecualizada.

3.2. Preprocesamiento

Las imágenes originales de Rayos X de tórax $\mathrm{Img}(i, j)$ proporcionadas por [15] no presentan un tamaño y el contraste de estas imágenes son relativamente bajos lo que afecta la calidad de la imagen en el diagnóstico. En este estudio, se

propone cambiar el tamaño de las imágenes $\mathrm{Img}(i,j)$ a 400×400 píxeles con el fin de utilizar un tamaño de imagen homogéneo, además se aplica una ecualización adaptativa del histograma por contraste mejor conocido como CLAHE[20] para mejorar la calidad de la imagen. En la Figura 4 se observa la diferencia de contraste entre una imagen original y la misma imagen ecualizada $\mathrm{Img}_{\mathrm{Ec}}$ que presenta ligera mejora de contraste.

3.3. Extracción de la ROI

La radiografía de tórax normalmente presenta etiquetas que contienen información adicional como por ejemplo, el nombre del paciente, la fecha del examen, el indicador del lado derecho-izquierdo de la radiografía, etc. En este trabajo se propone extraer la región de la caja torácica, la cual es donde está presente la neumonía. Se extrae esta área, con el propósito de obtener una imagen de región de interés (ROI) para garantizar que las características extraídas pertenezcan a la región pulmonar que es donde se presenta la neumonía.

Para generar la imagen procesada $\mathrm{Img}_{\mathrm{Ps}}$ se deben de aplicar diferentes algoritmos *handcraft* a la imagen $\mathrm{Img}_{\mathrm{Ec}}$. Se requieren cuatro pasos para obtener las imágenes procesadas, primero se aplica una umbralización para generar una imagen binaria $\mathrm{Img}_{\mathrm{Bin}}$ donde las regiones más claras de la imagen tomaran tonalidades blancas y las partes oscuras de la imagen tomaran tonalidades negras, esta umbralización se puede definir como:

$$V(i,j) = \begin{cases} 1, & \text{si } X(i,j) >= th, \\ 0, & \text{si } X(i,j) < th, \end{cases} \tag{1}$$

donde $X(i,j)$ es el nivel de intensidad de la imagen de Rayos X en la posición (i,j), $V(i,j)$ es el valor asignado en la posición i,j y th está definido por:

$$th = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} X(i,j), \tag{2}$$

donde $X(i,j)$ es el nivel de intensidad de la imagen de Rayos X en la posición (i,j), a esta umbralización se le conoce en la literatura como umbralización por la media estadística. En el siguiente paso de la etapa de extracción de la ROI, se emplea el algoritmo de contornos activos morfológicos [5].

Para aplicar este algoritmo se emplean operaciones morfológicas como la erosión y la dilatación en una imagen binaria en lugar de resolver una PDE (Partial Differential Equations) y se denota como:

$$C_t = F \times N, \tag{3}$$

donde F es un escalar y N una superficie normal ya que desea observar la evolución de la curva C. La evolución de cualquier función $U(x,y)$, que incorpore la curva como uno de sus conjuntos de nivel es:

$$\frac{\partial u}{\partial t} = F \times | \bigtriangledown u|. \tag{4}$$

La PDE anterior para la curva de evolución es la siguiente:

$$\frac{\partial u}{\partial t} = \pm| \bigtriangledown u|, \qquad \text{Cuando} \quad F = \pm 1. \tag{5}$$

Algunos operadores morfológicos pueden ser expresados como PDE y se presentan de la siguiente manera:

– **Dilatación**

$$\lim_{h \to 0} \left(\frac{D_h u - u}{h} \right) = | \bigtriangledown u|, \tag{6}$$

$$\frac{\partial u}{\partial t} = | \bigtriangledown u|. \tag{7}$$

– **Erosión**

$$\lim_{h \to 0} \left(\frac{D_h u - u}{h} \right) = | \bigtriangledown u|, \tag{8}$$

$$\frac{\partial u}{\partial t} = -| \bigtriangledown u|. \tag{9}$$

Por lo tanto, de la imagen Img_{Bin} se genera una máscara del área donde se encuentra contenido el pulmón Msk. A continuación, se aplica un filtro de media con una ventana de 21×21 pixeles a la imagen Msk eliminando regiones que no forman parte del área pulmonar Img_{Mn}.

Finalmente, se extrae la caja torácica obteniendo Img_{Ps}. En la Figura 5 se muestra una comparación entre una imagen original y una procesada, donde se han eliminado las áreas insignificantes, es decir, que no aportan información; estas imágenes de procesadas ayudan a la CNN a extraer características de la región pulmonar.

3.4. Etapa de clasificación

En la siguiente sección, se presenta una breve discusión de la arquitectura Xception utilizada en el sistema y la metodología utilizada al emplearlas.

Transferencia de aprendizaje. Es una técnica que utiliza los pesos obtenidos por un sistema que fue entrenado para una tarea en general, estos pesos son transferidos a otro sistema para resolver una tarea particular y se define como:

Sea un dominio general D_g entrenado para una tarea general T_g y un dominio particular D_p con una tarea particular T_p. Transfer Learning mejora el aprendizaje de la función predictiva particular $f_p(.)$ en el dominio D_p usando los conocimientos de D_g y la tarea T_g donde el dominio $D_g \neq D_p$ o la tarea $T_g \neq T_p$. En la Figura 6 se presenta un ejemplo visual del transfer learning para una tarea en particular.

Xception. Chollet et al.[6] propone la arquitectura Xception la cual está basada en la arquitectura Inception-V3, la diferencia radica en que Xception

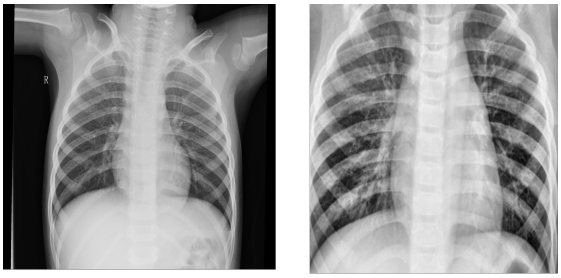


Fig. 5. Comparación de las Imágenes de Rayos X A) Imagen original y B) Imagen procesada.

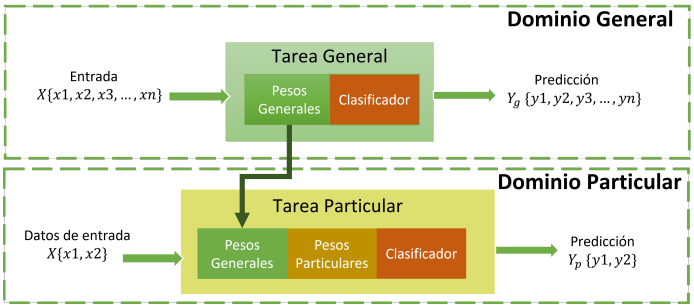


Fig. 6. Esquema de la técnica transfer learning.

Tabla 1. Hiperparametros usados en la Arquitectura Xception CNN.

Arquitectura	Learning Rate	Decay	Batch size	Optimizador	Épocas
Xception	5×10^{-2}		16	SGD	100

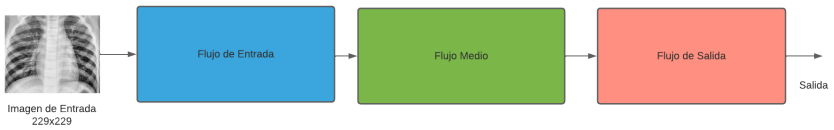


Fig. 7. Arquitectura simplificada de Xception.

realiza convoluciones separables (Depth Wise Separable Convolutions) por cada canal de la imagen de entrada por lo que reduce las operaciones y el costo computacional de la red, el autor reporta una exactitud del 94 % utilizando el conjunto de datos llamado Imagenet.

4. Resultados experimentales

4.1. Entorno experimental

El sistema CADx propuesto se implementó en una PC con una CPU Intel®Xeon E5 1230-V5, 24 GB de RAM, con una GPU NVIDIA RTX ®3090 con 24 GB de RAM, ejecutando un sistema operativo Linux de 64 bits, Python 3.6, y las bibliotecas: TensorFlow 2.7[18] y scikit-image[3].

4.2. Hiperparámetros

Se utilizan las imágenes procesadas para entrenar la arquitectura CNN anteriormente mencionada. Además, obtenemos el mapa de activación de características mediante el algoritmo Grad-CAM para validar que las características extraídas pertenezcan a la región pulmonar respectivamente a cada clase. La Tabla 1 contiene los hiperparámetros empleados para la arquitectura CNN. El Learning Rate decay es expresado como:

$$Lrd = Lr_i \times D^{(s/ds)}, \tag{10}$$

donde Lr_i es la taza de aprendizaje inicial, D es la taza de caída s son los pasos, que aumentan en cada época y ds es el umbral de la taza de caída de los pasos.

4.3. Métricas de evaluación

Evaluamos el rendimiento del sistema CAD propuesto considerando las siguientes métricas de evaluación: **Exactitud (Accuracy)**, **Precisión**, **Sensibilidad**, **Especificidad** y **Metrica-F1**. Estos criterios se describen en términos de tp, tn, fp y fn, que denotan verdadero positivo, verdadero negativo, falso positivo y falso negativo:

$$\text{Exactitud} = \frac{tp + tn}{tp + tn + fp + fn}. \tag{11}$$

El valor de la exactitud mide los casos en que el sistema clasifica correctamente utilizando el total de elementos evaluados, y se calcula como:

$$\text{Sensibilidad} = \frac{tp}{tp + fn}. \tag{12}$$

El valor de sensibilidad mide el número de elementos positivos que se clasifican correctamente, y se calcula como:

$$\text{Especificidad} = \frac{tn}{tn + fp}. \tag{13}$$

El valor de especificidad mide el número de elementos negativos que se clasifican correctamente, y se calcula como:

$$\text{Precision} = \frac{tp}{tp + fp}. \tag{14}$$

El valor de precisión mide el número de elementos que están correctamente clasificados entre el total de elementos positivos a evaluar, y se calcula como:

$$\text{Métrica-F1} = \frac{2tp}{2tp + fp + fn} \tag{15}$$

El valor de la Metrica-F1 mide la media armónica entre precisión y la sensibilidad.

4.4. Resultados y comparación

Para comparar el rendimiento del sistema propuesto con los sistemas descritos en el Estado del Arte, utilizamos métricas de evaluación y el algoritmo Grad-CAM para demostrar que utilizando las imágenes procesadas se puede extraer características significativas de neumonía en lugar de tomar características insignificantes en áreas como etiquetas, brazos, cuellos, etc. Los resultados expuestos en la Tabla 2 muestran que el sistema implementado mantiene un rendimiento similar al de otros sistemas existentes.

Finalmente, se emplea el algoritmo Grad-CAM [24], en los sistemas CAD presentados por Sánchez [23] y Lujan [17] presentan los mapas de calor para tener una aproximación a las regiones o patrones que se están tomando en cuenta para ser clasificadas como neumonía o normal. Los sistemas propuestos por Jiang [14] y Varela [28,29] no presenta ninguna validación de este tipo, por lo que la única comparación se realiza de forma cuantitativa.

La Figura 8 muestra la comparación entre cuatro imágenes Grad-CAM donde la imagen obtenida por Luján concentra su atención al área del corazón para clasificar neumonía, mientras que la imagen presentada por Sánchez [23] concentra toda la atención en una región perteneciente al hombro/brazo y lóbulo superior, del metodo propueso se puede observar que la imagen procesada enfoca la extracción de características dentro de la caja torácica tomando regiones como los alvéolos, lóbulo superior e inferior.

Tabla 2. Comparación del desempeño del sistema propuesto frente a sistemas encontrados en el Estado del Arte.

	Sistema Propuesto	**Jiang[14]**	**Sánchez[23]**	**Lujan[17]**	**Varela[28]**	**Varela[29]**
	Xception	**Xception**	**Xception**	**Xception**	**ANN**	**ANN**
Exactitud	91 %	92 %	96 %	87 %	97 %	99 %
Sensibilidad	91 %	-	95 %	99 %	-	98 %
Especificidad	86 %	-	95 %	84 %	-	99 %
Precision	87 %	-	95 %	84 %	-	-
F-score	91 %	-	91 %	95 %	-	-
#Imagenes-Val	624	500	200	624	112	624
Epocas	100	100	100	100	-	-

Finalmente, en el dataset de imágenes originales (sin procesar) se puede observar que las características extraídas son generales en la imagen, es decir,

el comportamiento del sistema clasifica neumonía tomando extrayendo características contenidas de las etiquetas, brazos y los bordes laterales de la imagen, regiones donde la neumonía no está presente fisiológicamente.

La comparación del desempeño entre diferentes sistemas CAD muestra que el sistema CAD propuesto es competitivo con los sistemas encontrados en la literatura. Jiang et al.[14] entrena, valida y prueba con 500 imágenes respectivamente, obteniendo una precisión del 92 % con la arquitectura Xception, pero su estudio no sustenta qué características se toman en cuenta para llevar esta clasificación.

Sánchez et al.[23] propone generar un subespacio para entrenar la arquitectura Xception utilizando únicamente 400 imágenes para entrenamiento y 200 para prueba, obteniendo una exactitud del 96 %, el mapa de calor mostrado toma características de la caja toráxica y alrededores, además el algoritmo PCA reduce la información de una imagen.

Lujan et al.[17] empleó un esquema de transferencia de aprendizaje la cual resultó en una exactitud del 87 %, pero al emplear el algoritmo Grad-CAM se puede ver que el sistema extrae características del área del corazón como significativas para clasificar una imagen en la clase de neumonía.

Varela et al [28,29] emplea una clasificación utilizando técnicas de tipo handcraft, los resultados presentan un sobre entrenamiento al obtener un ACC del 97 % y 99 % respectivamente, sin embargo, los autores no presentan las gráficas de Precisión/Perdida las cuales sirven para verificar el comportamiento de una red, además del uso de tres redes neuronales las cuales tendrán diferente comportamiento, a comparación de utilizar una la cual sea capaz de reconocer los diferentes conjuntos de patrones y así, poder entregar una sola clasificación al contrario de promediar el resultado de tres.

En este trabajo se propuso un método de procesamiento, el cual contribuye a la extracción de características de la caja toráxica por parte de la arquitectura Xception, debido a que se obtiene la región de interés. Se logra una precisión del 91 %. Para validar la eficiencia del sistema, se utilizó Grad-CAM, lo que confirma que utilizando las imágenes procesadas se pueden extraer características significativas del ROI, lo que resulta en un alto rendimiento de clasificación.

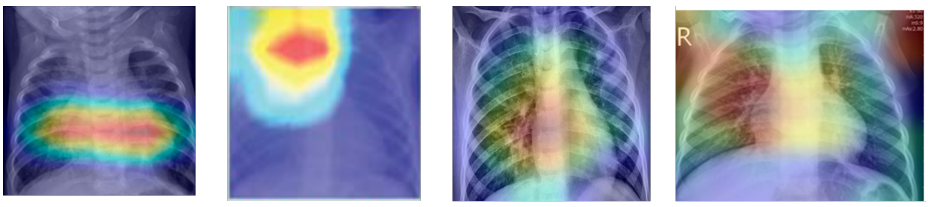


Fig. 8. Comparativa de los diferentes Grad-CAM obtenidos por A) Lujan[17], B) Sánchez[23], C) Sistema propuesto-imagen procesada, y D) Sistema propuesto-imagen original.

5. Conclusiones

El sistema CAD propuesto clasifica neumonía en imágenes de Rayos X, demuestra un buen desempeño en cuanto a las métricas de evaluación utilizadas: Exactitud, Especificidad y Sensibilidad, obteniendo 91 %, 91 %, 86 % respectivamente, garantizando la extracción de los patrones contenidos en la caja toráxica y demostrando un desempeño similar en comparación con los sistemas CAD presentados en el estado del arte.

Nuestro trabajo futuro consistirá en utilizar arquitecturas CNN entrenadas desde cero *(from scratch)* para mejorar la clasificación perteneciente al área pulmonar, desarrollando un sistema multiclase donde se emplee la técnica de procesamiento propuesta.

Agradecimientos. Se le agradece cordialmente al Instituto Politécnico Nacional y al Consejo Nacional de Ciencia y Tecnología por el apoyo brindado para la realización de este proyecto.

Referencias

1. ACR-RSNA: Rayos X del Tórax (radiografía de tórax) (2019)
2. Bisong, E.: Google colaboratory. Building Machine Learning and Deep Learning Models on Google Cloud Platform. Apress, Berkeley, CA.
3. Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., et al.: Design for machine learning software: Experiences from the scikit-learn project (2013)
4. Características de las defunciones registradas en México (2020)
5. Caselles, V., Ron, K., Guillermo, S.: Geodesic active contours. In: Proceedings of IEEE international conference on computer vision. IEEE (1995)
6. Chollet, F.: Xception: Deep learning with depthwise separable convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition (2017)
7. Ciregan, D., Ueli, M., Jürgen, S.: Multi-column deep neural networks for image classification. In: 2012 IEEE conference on computer vision and pattern recognition. IEEE (2012)
8. Cush, J. J.: Approach to articular and musculoskeletal disorders. Harrisons Principles Of Internal Medicine, vol. 2, pp. 1979–1986 (2001)
9. Das, A., Agrawal, H., Zitnick, L., Parikh, D., Batra, D.: Human attention in visual question answering: Do humans and deep networks look at the same regions?. Computer Vision and Image Understanding (2017)
10. WHO: Estimación Mundial de La Incidencia de Neumonía Clínica entre los menores de 5 años (2013)
11. Gonzalez, R. C., Richard, E. W.: Image processing. Digital Image Processing 2.1 (2007)
12. Goodfellow, I., Yoshua, B., Aaron, C.: Deep learning. MIT press, (2016)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition (2015)
14. Jiang Z.: Chest x-ray pneumonia detection based on convolutional neural networks. In: 2020 International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE) pp. 341–344. IEEE (2020)

15. Kermany, D., Zhang, K., Goldbaum, M.: Labeled optical coherence tomography (OCT) and chest X-Ray images for classification (2018)
16. Krizhevsky, A., Ilya, S., Geoffrey, E. H.: Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, vol. 25 (2012)
17. Luján-García, J. E., et.al.: A transfer learning method for pneumonia. Applied Sciences, vol. 8, no. 2908, pp. 10 (2020)
18. Martín, A., Ashish, A., Paul, B., Eugene, B., Zhifeng, C., et.al.: TensorFlow: Large-scale machine learning on heterogeneous systems, software available from tensorflow.orgm (2015)
19. Mueller, J. P., Luca, M.: Deep Learning for dummies (2019)
20. Pizer, S. M., et al.: Adaptive histogram equalization and its variations. Computer vision, graphics, and image processing 39.3, pp. 355–368 (1987)
21. Reynolds, J. H., Arpan, K. B.: Imaging pneumonia in immunocompetent and immunocompromised individuals. Current opinion in pulmonary medicine 18.3, pp. 194–201 (2012)
22. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A. et.al.: Imagenet large scale visual recognition challenge. International Journal of Computer Vision, vol. 115, no. 3, pp. 211–252 (2015)
23. Sanchez, K., et.al.: Subspace-based domain adaptation using similarity constraints for pneumonia diagnosis within a small chest x-ray image dataset. In: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI) pp. 1232–1235 IEEE (2021)
24. Selvaraju, R. R., Das, A., Vedantam, R., Cogswell, M., Parikh, D., Batra, D.: Grad-Cam: why did you say that? visual explanations from deep networks via gradient-based localization (2016)
25. Sokolova, M., Guy, L.: A systematic analysis of performance measures for classification tasks. Information processing & management, vol. 45, no. 4, pp. 427–437 (2009)
26. Sons, J. W., Pitas, I.: Digital image processing algorithms and applications (2000)
27. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. The journal of machine learning research. vol. 1929, no. 58 (2014)
28. Varela-Santos, S., Melin, P.: A new approach for classifying coronavirus COVID-19 based on its manifestation on chest x-rays using texture features and neural networks. Information sciences, vol. 545, pp. 403–14 (2021)
29. Varela-Santos, S., Melin, P.: A new modular neural network approach with fuzzy response integration for lung disease classification based on multiple objective feature optimization in chest Xray images, Expert Systems with Applications (2020)
30. Wang, H., Lei, Z., Zhang, X., Zhou, B., Peng, J.: Machine learning basics. Deep learning, pp. 98–164 (2016)
31. INSP: ¡El invierno se acerca! Y la neumonía lo sabe (2020)

ISSN 1870-4069

Predicción de enfermedades cardíacas derivadas de diabetes, mediante algoritmos genéticos: caso de estudio

Isamar Aparicio-Montelongo[1], José M. Celaya-Padilla[2], Huizilopoztli Luna-García[2], Carlos E. Galván-Tejada[2], Jorge I. Galván-Tejada[2], Hamurabi Gamboa-Rosales[2]

[1] Consejo Nacional de Ciencia y Tecnología,
Unidad Académica de Ingeniería Eléctrica,
México

[2] Universidad Autónoma de Zacatecas,
México

{20204469, jose.celaya, hlugar, ericgalvan,
gatejo, hamurabigr}@uaz.edu.mx

Resumen. En México, según datos del INEGI las principales causas de muerte son Enfermedades Cardíacas (EC) y Diabetes Mellitus (DM). En el primer semestre de 2021 se registraron 579,586 decesos, ocupando los primeros lugares solo después de COVID-19. Correspondiendo a las EC, como la segunda más importante con un 19.7%. Gracias al avance tecnológico, en la actualidad, es posible crear modelos para el diagnóstico oportuno de patologías, mediante técnicas de Inteligencia Artificial (IA). El objetivo de esta investigación es implementar algoritmos genéticos (AG) como método de selección de características posteriormente, generar un modelo multivariado con Regresión Logística para conocer si es posible considerarlo como una herramienta eficaz en la detección de pacientes propensos a sufrir un episodio cardíaco derivado de DM.

Palabras clave: Enfermedades cardíacas, inteligencia artificial, predicción, diabetes, algoritmos genéticos, selección de características.

Prediction of Diabetes-Related Heart Disease Using Genetic Algorithms: A Case Study

Abstract. In Mexico, according to INEGI data, the main causes of death are Heart Disease (CD) and Diabetes Mellitus (DM). In the first semester of 2021, 579,586 deaths were registered, occupying the first places only after COVID-19. Corresponding to CD, as the second most important with 19.7%. Thanks to

technological advances, it is now possible to create models for the timely diagnosis of pathologies, using Artificial Intelligence (AI) techniques. The objective of this research is to implement genetic algorithms (GA) as a feature selection method and subsequently generate a multivariate model with Logistic Regression to find out if it is possible to consider it as an effective tool in the detection of patients prone to suffer a cardiac episode derived from DM.

Keywords: Heart disease, artificial intelligence, prediction, diabetes, genetic algorithms, feature selection.

1. Introducción

Según la Organización Mundial de la Salud (OMS), la Diabetes es una enfermedad metabólica crónica caracterizada por niveles elevados de glucosa en sangre (o azúcar en sangre), que con el tiempo es una causa importante de complicaciones cardíacas, retinopatía, nefropatía, accidente cerebrovascular y amputación de miembros inferiores [1]. Cada año mueren más personas por Enfermedades Cardíacas (EC) que por cualquier otra causa. Más de tres cuartas partes de los fallecimientos relacionados con cardiopatías ocurren en países en proceso de desarrollo [2].

Lamentablemente, en México, los datos del INEGI muestran que tan solo en 2019 fallecieron 156,041 personas, por distintas enfermedades del corazón. Así mismo, en 2020 el número fue de 218,704 individuos, es decir, hubo un incremento del 40% aproximadamente de personas fallecidas entre ambos años [3] y para el primer semestre de 2021 ya existían 113,899 muertes [4]. La probabilidad de padecer una Enfermedad Cardiovascular (ECV) aumenta conforme se desarrolla la obesidad y sobrepeso inducido por malos hábitos alimenticios, inactividad física, consumo nocivo de alcohol, sal, azúcares, grasas, entre otros [1].

Hoy en día, la Ciencia y la Tecnología han concentrado sus esfuerzos para tratar de mitigar el aumento de decesos por estas enfermedades, por tal motivo, en el presente estudio, se realizó una aproximación precisa para determinar las principales causas de complicaciones cardíacas, derivadas de Diabetes y presentes en grupos de edad (Jóvenes, Adultos y Adultos Mayores) proponiendo a algoritmos genéticos como método de selección de características, dado que están inspirados en la selección natural, simulando el proceso evolutivo de los organismos vivos para resolver problemas de optimización y búsqueda. A continuación, se describen algunas investigaciones, en donde se apreciarán sus contribuciones, en el ámbito detección de estos padecimientos por medio de técnicas de Inteligencia Artificial.

2. Trabajos relacionados

El uso de técnicas de Inteligencia Artificial (IA) se refiere a la combinación de algoritmos para crear máquinas o sistemas que imitan la inteligencia humana siendo capaces de analizar gran cantidad de datos, identificar patrones, formular predicciones

de manera automática y precisa, automatizar actividades como la toma de decisiones y resolución de problemas. En esta sección abordaremos el aporte de este ámbito desde un aspecto médico y científico analizando las contribuciones en la detección o diagnósticos de las enfermedades de interés de este estudio.

La comunidad científica ha tratado de incorporar estas técnicas de IA en el proceso de diagnóstico temprano de Enfermedades del Corazón, por ejemplo en 2019 González-Cedillo et al [5], desarrolló un sistema predictivo para detectar pacientes propensos a sufrir alguna EC, a través del uso de Naive Bayes, el autor reporta una exactitud en el modelo de 86.81%, sin embargo, no fueron exploradas otras opciones de clasificación, así mismo, el modelo propuesto incluía 75 atributos, de los cuales, 14 fueron seleccionados en base a investigaciones efectuadas en otros estudios. Además, la cantidad de instancias analizadas (303), limita la efectividad del resultado.

Posteriormente en 2020, Chicco et al [6], realizaron una comparativa de diez clasificadores de aprendizaje automático (Random Forest, Gradient boosting, Decision tree, Regresión Lineal, Naive Bayes, SVM, Redes Neuronales y KNN, por mencionar algunos) para predecir la supervivencia de pacientes (299) con Insuficiencia Cardíaca y categorizar las características clínicas correspondientes a los factores de riesgo más importantes, comparando los resultados de las predicciones a través de índices comunes de la matriz de confusión, como el coeficiente de correlación de Matthews (MCC), área bajo la curva (AUC), accuracy, entre otros.

Obteniendo el 74% de exactitud mediante Random Forest. Así mismo, Alí et al [7] en 2020, proponen un sistema sanitario inteligente para la predicción de Enfermedades Cardíacas utilizando enfoques de fusión de características y aprendizaje profundo. Llevaron a cabo una comparativa del modelo propuesto, contra seis clasificadores existentes (Naive Bayes, Redes Neuronales, Decision tree, Random Forest, Regresión Logística y SVM).

El dataset utilizado fue una combinación de dos conjuntos diferentes sobre enfermedades del corazón, obteniendo un total de 597 pacientes y 90 variables, realizan la eliminación de características ruidosas mediante el método de la ganancia de información y entropía, reduciendo a 14 atributos más significativos. Manejan métricas de evaluación como accuracy, recall, error cuadrático medio (por mencionar algunas).

Los autores informan que su modelo propuesto muestra una exactitud del 83.5%, siendo mayor a otros algoritmos convencionales. Por otro lado, en 2021 Gallego Valcárcel & Lucas Monsalve et al [8], implementaron modelos de clasificación utilizando técnicas de aprendizaje automático (Redes Neuronales, Máquina de soporte de vectores y Random Forest), para predecir el riesgo de fallecer por Insuficiencia Cardíaca a partir de datos clínicos de pacientes recopilados (299). Apoyándose de técnicas de reducción de variables (análisis de componentes principales y eliminación hacia atrás), se obtienen 17 características más relevantes.

Realizaron la evaluación de los modelos por medio de validación cruzada, siendo las Redes Neuronales, el mejor algoritmo, con una exactitud del 82.63%. Mientras tanto, Faiayaz Waris & Koteeswaran et al [9] en 2021, proponen mejorar el algoritmo de vecinos cercanos (KNN, por sus siglas en inglés) y, con ello, corroborar que es más preciso que el KNN normal, en predicción temprana de Enfermedades Cardíacas.

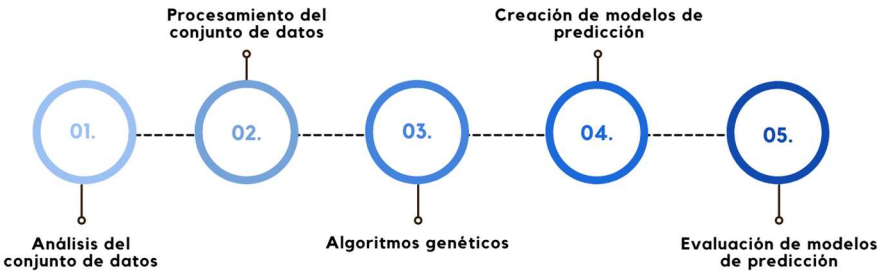


Fig. 1. Proceso general de ejecución de estudio.

El conjunto de datos analizado consta de 303 pacientes y 14 características, mismas que fueron incluidas en los modelos planteados. Alcanzando un 93% de exactitud mediante KNN modificado. Como se puede observar, la mayoría de las propuestas utilizan una cantidad muy pequeña de datos y no todas realizan una selección de características, por lo que se requiere un análisis más profundo.

Es por ello, que en esta investigación se plantea trabajar con un conjunto de datos de mayor dimensión considerando información de personas que radican en Estados Unidos, dado que la mayoría de sus pacientes tiene etnicidad similar a la de los mexicanos por lo que nos permitiría obtener conocimiento más amplio en la detección de enfermedades, además de que la infraestructura y políticas de aquel país permiten tener acceso a este tipo de datos de manera abierta para la comunidad científica y académica. De acuerdo con lo anterior, se optó por realizar un primer acercamiento a la predicción de Enfermedades Cardíacas con datos ya existentes en la comunidad médica de aquel país lo suficientemente amplio para generar un modelo preciso y eficaz y poder ser implementado en un contexto como el de México.

Explorando la selección de las variables más significativas, por medio de algoritmos genéticos (basados en la teoría evolutiva funcionan con una población de cromosomas que por sus diversas características abarcan un gran campo de distintas soluciones al mismo tiempo, por el contrario, los algoritmos tradicionales manipulan un solo punto de búsqueda) [10], para después generar modelos predictivos con Regresión Logística, realizar una evaluación mediante validación cruzada y así evitar el sobreajuste.

3. Metodología

Por medio de un modelo se busca determinar cuáles son los factores de riesgo para clasificar pacientes diabéticos propensos a sufrir una Enfermedad Cardíaca. La metodología propuesta se muestra en la Fig. 1.

En primera instancia, el conjunto de datos es separado en distintas clases (sujetos con Diabetes, Pre-diabetes, No Diabéticos) así como por grupos de edad (Jóvenes, Adultos, Adultos mayores), posteriormente, se realiza la selección de características mediante un algoritmo genético, para encontrar el mejor subconjunto de variables que permita detectar una EC, por último, este modelo es evaluado mediante validación cruzada.

Tabla 1. Datos demográficos del conjunto de datos original.

Sector poblacional			Enfermos	Sanos
Hombres	Jóvenes	6736	10,205	131,769
	Adultos	66,125		
	Adultos Mayores	69,113		
Mujeres	Jóvenes	6562	13,688	98,018
	Adultos	51,943		
	Adultos Mayores	53,201		

3.1. Datos utilizados en el estudio

Se trabajó en un conjunto de datos disponible de manera abierta obtenidos de la plataforma Kaggle en su repositorio de datasets, que lleva por nombre "Indicadores de salud de Enfermedades Cardíacas (Health Indicators for Heart Disease)" [11], consta de 22 características entre binarias u ordinales de las cuales se tomarán en cuenta los atributos clínicos como edad, sexo, índice de masa corporal (IMC), control de colesterol, presión arterial alta, colesterol alto, diabetes, características sobre hábitos como actividad física, dificultad para caminar, consumo de frutas y verduras, estado de salud física, mental y general, así mismo, vicios tales como consumo excesivo de alcohol y cigarrillos.

Tiene un total de 253,680 pacientes entre sanos (57% hombres y 43% mujeres) y enfermos (predominando las mujeres con el 57%) como se muestra en la Tabla 1, de los cuales, 141,974 son hombres, 111,706 mujeres, con un rango de edad que van desde 18 a 80 años en adelante; sobresaliendo el grupo de 60 a 64 años en ambos géneros con el 13% para cada uno. El 91% de las observaciones son personas sanas (229,787) y 9% son pacientes con alguna Enfermedad Cardíaca (23,893).

3.2. Procesamiento de datos

Con el propósito de probar si hay diferencia entre grupos de edad, se determinó segmentar tres sectores de la población, los cuales, son personas Jóvenes entre 18 y 29 años (13,298), Adultos entre 30 y 59 años (118,068) y Adultos mayores de 60 años (122,314),así mismo, a fin de explorar la posibilidad de padecer alguna Enfermedad Cardíaca en pacientes Diabéticos, se dividió en tres agrupaciones: No diabéticos (213,703), Pre-diabéticos (4631) y Diabéticos (35,346), encontrando los factores de riesgo más significativos entre cada segmento de la población.

3.3. Algoritmos genéticos (AG)

Para la selección de características se implementó un AG, el cuál, es un método de búsqueda de variables que se basa en el principio de la evolución por selección natural. El procedimiento funciona haciendo evolucionar conjuntos de atributos (cromosomas) que se ajustan a determinados criterios a partir de una población aleatoria inicial

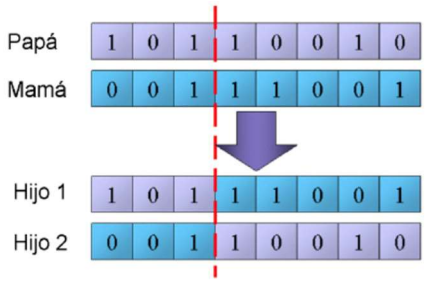


Fig. 2. Intercambio de información genética entre dos individuos.

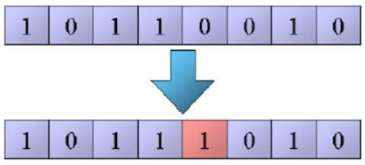


Fig. 2. Una mutación modifica al azar parte del cromosoma de los individuos.

mediante ciclos de replicación diferencial, recombinación y mutación de los cromosomas más aptos [12].

Por imitación de este proceso, los algoritmos genéticos son capaces de ir creando soluciones para problemas del mundo real. El proceso de evolución tiene como único objetivo mejorar la población de soluciones mediante la aplicación repetitiva de las operaciones de cruzamiento (sinónimo de apareamiento entre dos individuos de diferente sexo) mutación (alteraciones ocasionales del cromosoma) y selección (los cromosomas del individuo más fuerte o mejor adaptado son transferidos a su descendencia).

El intercambio de la información genética del par de individuos también se puede llevar a cabo de diferentes formas (una de ellas se ilustra en la Fig. 2), donde aleatoriamente se ha seleccionado un punto de corte común a ambos padres, y que sirve como referencia para intercambiar su información genética para producir dos hijos con características diferentes a los padres, aunque éstos hereden parte de su información genética.

Una vez establecida la frecuencia de mutación, se genera un número entre 0 y 1 de manera aleatoria y si ese número es menor que la frecuencia de mutación se permite que un gen del cromosoma cambie su información; si no, se dejará como está. La mutación modifica al azar parte del cromosoma de los individuos (ver Fig. 3), y permite alcanzar zonas del espacio de búsqueda que no estaban cubiertas por los individuos de la población actual.

Finalmente, una vez aplicados los operadores genéticos, se seleccionan los mejores individuos para conformar la población de la generación siguiente. Este proceso se realiza por medio de la evaluación de cada individuo con la función de aptitud y se reemplaza la población original. El algoritmo genético se deberá detener cuando se

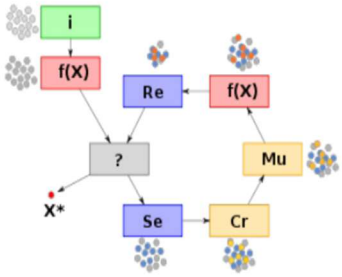


Fig. 3. Ciclo del algoritmo genético.

alcance la solución óptima, por lo general ésta se desconoce, así que se deben utilizar otros criterios de detención.

Normalmente se usan dos criterios: 1) correr el AG un número máximo de iteraciones (generaciones), y 2) detenerlo cuando no haya cambios en la población. Mientras no se cumpla la condición de término se repite el ciclo: gSelección (Se) → Cruzamiento (Cr) → Mutación (Mu) → Evaluación ($f(x)$) → Reemplazo (Re). Ver Fig. 4, donde (?) es la condición de término y $x*$ es la mejor solución [13].

En este estudio se utilizaron 500 soluciones máximas también conocidas como cromosomas, conteniendo 5 genes (chromosomeSize=5) que corresponden a modelos desarrollados utilizando un clasificador del centroide más cercano (classification.method=”nearcent”) con una precisión de clasificación del 100% (goalFitness=1) el cual también se le conoce como función objetivo ya que representa el valor de precisión que desea obtener el cromosoma siendo este el criterio de detención, para la selección de características se empleó el método selección hacia adelante, el cual es iterativo comenzando con un modelo en la que no tiene ninguna variable y en cada iteración se va añadiendo una variable hasta que la adición de nuevas características no mejore el rendimiento del modelo.

3.4. Modelos de predicción y métricas de desempeño

Después de obtener las características más significativas, se procede a la creación de modelos de predicción implementando Regresión Logística (RL) ya que en primera instancia es un algoritmo sencillo, fácil de entrenar sobre gran cantidad de datos y rara vez existe un sobreajuste en comparación con otros algoritmos, por tal motivo se optó por aplicar RL en esta investigación.

Es una técnica de aprendizaje automático que proviene del campo de la estadística, mide la relación entre la probabilidad de la característica dependiente con una o más variables independientes y el conjunto de atributos disponibles para el modelo. Lo que se busca en estos problemas es una clasificación, por lo que se obtiene un resultado binario entre 0 y 1.

Se utiliza un valor umbral para asignar los valores de probabilidad, cuando es mayor a 0.5 el resultado es positivo, de lo contrario, será negativo. A la función que relaciona la variable dependiente con las independientes se le llama función sigmoidea, la cual

Tabla 1. Matriz de confusión.

		Predicción	
		Positivos	**Negativos**
Observación	**Positivos**	**Verdaderos Positivos (VP)**	**Falsos Negativos (FN)**
	Negativos	**Falsos Positivos (FP)**	**Verdaderos Negativos (VN)**

es una curva en forma de S que puede tomar cualquier valor entre 0 y 1, pero nunca valores fuera de estos límites [14], la Ecuación 1 define esta función:

$$p(x) = \frac{1}{1 + e^{-z}}, \tag{1}$$

donde z es la representación de los coeficientes del modelo de regresión que después de realizar un proceso algebraico sobre la ecuación sigmoide se obtiene la expresión matemática que representa el modelo de Regresión Logística, como se muestra en la ecuación 2:

$$p(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}. \tag{2}$$

Con ayuda de algunas métricas se evalúan los modelos midiendo el desempeño de las predicciones realizadas. A continuación, se describen las métricas utilizadas en el presente estudio:

- Curva ROC: herramienta estadística utilizada para clasificar a los individuos de una población en dos grupos (uno que represente un evento de interés y otro que no). Dada por el área bajo la curva (AUC), métrica de precisión estándar para modelos de clasificación binaria el cual mide la capacidad de predecir eventos positivos en comparación con negativos, devuelve un valor decimal comprendido entre 0 y 1; los valores cercanos a 1 indican un modelo de aprendizaje automático muy preciso.
- Matriz de confusión. Es una herramienta que permite obtener el desempeño de un algoritmo, se aplica en problemas de clasificación binaria (2 clases). Está compuesta por verdaderos positivos (VP), falsos negativos (FN) es decir, casos que en realidad fueron positivos pero el modelo lo clasificó como negativo, falsos positivos (FP) y verdaderos negativos (VN) los cuales son casos que en realidad fueron negativos pero el modelo lo clasificó como positivo [15]. Como se muestra en la Tabla 2, cada columna de la matriz representa el número de predicciones de cada clase y las filas interpretan los valores reales. Mediante la matriz de confusión se pueden obtener algunas métricas de evaluación como exactitud, sensibilidad y especificidad.

- Exactitud (Accuracy). Métrica que indica el porcentaje de predicciones clasificadas correctamente, tanto como para positivos y negativos [16], dada por la Ecuación 3:

$$\text{Exactitud} = \frac{VP + VN}{VP + FP + FN + VN}. \tag{3}$$

- Sensibilidad. Indica la tasa de clasificación positiva, es decir, la proporción de casos positivos que el modelo predijo correctamente (verdaderos positivos), está dada por la Ecuación 4: (4)

$$\text{Sensibilidad} = \frac{VP}{VP + VN}. \tag{4}$$

- Especificidad. Es la capacidad de un algoritmo para predecir un falso positivo, es decir, el resultado real de la predicción es negativo y el modelo lo clasificó como positivo [17], dada por la Ecuación 5:

$$\text{Especificidad} = \frac{VN}{VN + FP}. \tag{5}$$

3.5. Evaluación de modelos de predicción

Para determinar qué tan eficaces son nuestros modelos se evaluaron el rendimiento de cada uno con Validación cruzada K Fold (K-Fold Cross-Validation), la cual nos permite estimar la capacidad predictiva de estos cuando se utilizan nuevas observaciones diferentes a los usados en el entrenamiento.

Este proceso consiste en dividir al conjunto de datos de forma aleatoria en K particiones aproximadamente del mismo tamaño cada uno, mientras que k-1 folds se usan para entrenar el modelo y una partición se utiliza como prueba.

Este proceso se repite k veces utilizando una partición distinta como validación en cada iteración, generando k estimaciones del error cuyo promedio se emplea como estimación final [18].

4. Diseño experimental

Con la intención de generar una herramienta que contribuya en mejorar las estrategias de detección temprana de Enfermedades y a la necesidad de fortalecer la atención primaria de salud en México por las altas tasas de mortalidad prematura, deterioro en la calidad de vida de los pacientes y los altos costos de atención de sus complicaciones, este estudio busca contribuir con esta finalidad realizando una adaptación de diagnósticos más personalizados.

Para ello fue necesario dividir el conjunto de datos procedente de pacientes de Estados Unidos en diferentes sectores poblacionales por edades y por grupos de casos positivos y grupos de control cómo se menciona en la sección 4.2, seguido de la implementación de algoritmos genéticos (descrito en la sección 4.3) mediante Galgo,

Tabla 2. Características más significativas y porcentaje de desempeño en grupos de Diabetes.

Subconjunto	Características más significativas	Desempeño del modelo mediante AUC
No Diabéticos	Dificultad para caminar, control de colesterol, consumo excesivo de alcohol y sexo	68.64%
Prediabéticos	Salud física, edad, frutas y salud mental	69.76%
Diabéticos	Dificultad para caminar, colesterol alto, sexo, salud física, salud en general, consumo excesivo de alcohol, edad, presión arterial alta, IMC, salud mental y fumador	73.62%

librería del software estadístico R que utiliza AG para resolver problemas de optimización, especialmente en conjuntos de datos con grandes dimensiones [12].

Seguido de la creación de modelos clasificatorios por Regresión Logística, centrándose en métricas de desempeño vistas en la sección 4.4 por último, evaluar las predicciones a través de Cross-Validation (sección 4.5) con k=3, para evitar el sobreajuste en las predicciones, esta etapa es fundamental para obtener un modelo adaptable al contexto de la sociedad mexicana.

5. Resultados

Después de revisar el diseño experimental (descrito en la sección 4), se ejecutaron dos análisis diferentes obteniendo los siguientes resultados:

5.1. Grupos de diabetes

Para el primer análisis se emplearon tres subconjuntos: No diabéticos, Pre-diabéticos y Diabéticos. Los algoritmos genéticos obtuvieron entre 4 y 11 características más significativas para cada uno de los grupos (ver Tabla 2).

La evaluación de los modelos propuestos se realiza con la validación cruzada con la intención de evitar un sesgo en las predicciones y efectúa particionando el conjunto de datos original en subconjuntos aleatorios conformado por el 70% para entrenamiento y 30% para pruebas, realizando este proceso 3 veces (k=3), es decir 3 particiones diferentes.

Se obtuvieron los resultados de las métricas para después conseguir el promedio general de cada una, como se muestra en la Fig. 4. Logrando de esta manera una exactitud (accuracy) del 55.6% para personas No diabéticas, 67.4% en Pre-diabéticos y 65.8% para Diabéticos.

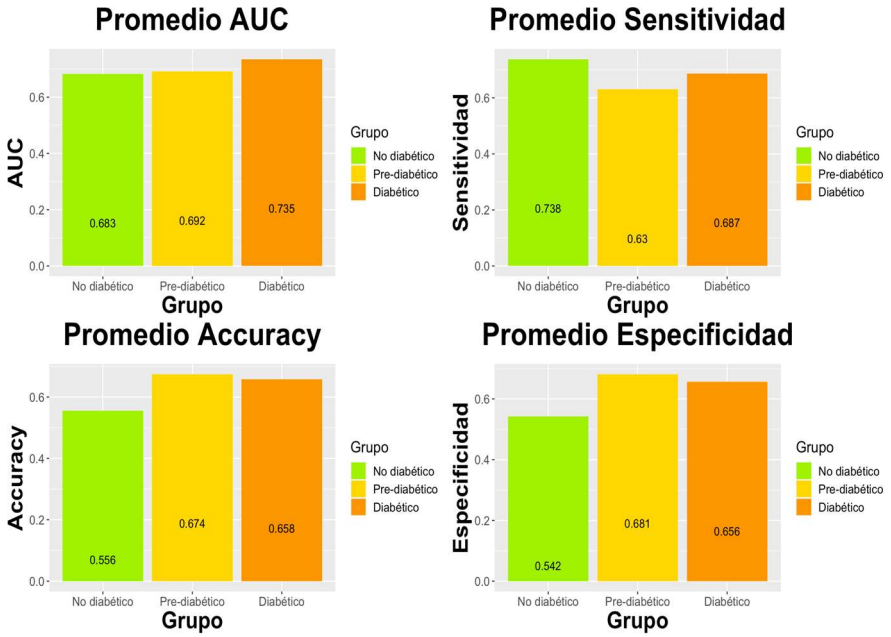


Fig. 4. Promedios de métricas de desempeño, obtenidas en grupos de Diabetes mediante validación cruzada.

Tabla 3. Características más significativas en grupos de Edad.

Subconjunto	Características más significativas	Desempeño del modelo mediante AUC
Jóvenes	Diabetes, control de colesterol, frutas	58.4%
Adultos	Edad, Diabetes	71.87%
Adultos Mayores	Salud física, fumador	63.68%

5.2. Grupos de edad

Para el segundo análisis, las variables más significativas obtenidas por el algoritmo genético se muestran en la Tabla 3, al igual que el desempeño del modelo descrito por AUC de cada modelo multivariado, siendo el grupo de adultos el de mayor precisión en las predicciones con el 71.87%, seguido de adultos mayores con 63.68% y 58.4% para jóvenes.

Un buen modelo de clasificación debe proporcionar predicciones precisas, por lo cual, es necesario aplicar una validación cruzada sobre los datos y de esta manera evitar un sesgo o sobreajuste. El promedio general de las métricas de desempeño utilizadas se muestra en la Fig. 5, abordando solamente exactitud se obtuvo 97.7% para Jóvenes, 68.3% en Adultos y 72.1% en Adultos Mayores.

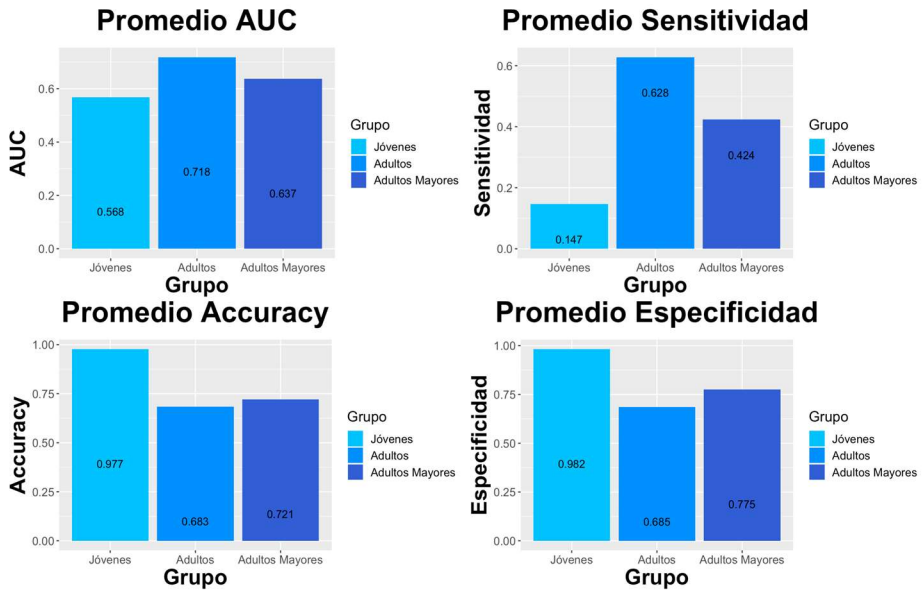


Fig. 5. Promedios de métricas de desempeño, obtenidas en grupos de Edad mediante validación cruzada.

6. Conclusiones y trabajo a futuro

La mayoría de los trabajos que incorporan IA tratan de generar un modelo que permita una correcta clasificación y no consideran las características que cambian en base a la etapa del sujeto, sin embargo, en esta investigación se busca estudiar si las variables que se están encontrando siguen siendo significativas a lo largo de la vida de los pacientes, con el fin de localizar modelos adaptativos dependiendo de su edad. De esta manera, contribuir en el diagnóstico preventivo y certero de padecimientos cardíacos proporcionando un tratamiento personalizado.

Este estudio muestra que la dificultad para caminar, un mal control de colesterol, género y un alto consumo de alcohol, son factores de riesgo que nos pueden ayudar a predecir Enfermedades Cardíacas en personas No Diabéticas, mientras que un bajo consumo de frutas, edad, mal estado de salud física y mental son características importantes para detectar una EC en pacientes Pre-diabéticos y Diabéticos, además, para este último grupo, es importante mencionar que el riesgo aumenta considerando otras variables como la dificultad para caminar, colesterol alto, género, mal estado de salud en general, alto consumo de alcohol, presión alta, índice de masa corporal y haber consumido más de 100 cigarrillos a lo largo de la vida.

En cuanto a los grupos de edad, en específico para Jóvenes y Adultos, es muy fundamental saber si el paciente es Diabético, ya que es una enfermedad clave para que un individuo sea propenso a sufrir un episodio cardíaco, debido a un mal control de

colesterol y malos hábitos alimenticios, por lo tanto, muestra un alto índice de padecer alguna complicación cardíaca.

La edad, también es un atributo importante que hay que tomar en cuenta en los Adultos, mientras que el mal estado de salud física y ser fumador son indicios que se deben considerar en Adultos Mayores. Las técnicas de Inteligencia Artificial y algoritmos de aprendizaje automático permiten predecir con un alto grado de precisión cualquier tipo de enfermedad en etapas tempranas, de manera no invasiva.

Como trabajo a futuro, se propone mejorar el desempeño del modelo de clasificación implementando otro método de selección de características dentro de los algoritmos genéticos, además de realizar un análisis de este mismo conjunto de datos, aplicado en hombres y mujeres, para determinar los factores de riesgo presentes en cada segmento y relacionar sus similitudes siguiendo la misma metodología que en el presente estudio.

Además, sería de gran interés realizar una comparativa de los resultados obtenidos en esta investigación con datos de pacientes mexicanos, como se mencionó en la sección de trabajos relacionados en la actualidad no existe un conjunto de datos abiertos con acceso al público de información de esta índole en México, debido a que nuestro país sigue en proceso de desarrollo.

Agradecimientos. Este artículo fue desarrollado gracias al apoyo de las Becas de Posgrado otorgado por Consejo Nacional de Ciencia y Tecnología (CONACYT) a la alumna Isamar Aparicio-Montelongo con el número de becario 1110281 de la Maestría en Ciencias del Procesamiento de la Información.

Referencias

1. Organización Mundial de la Salud: Diabetes (2021)
2. Organización Mundial de la Salud: Enfermedades cardiovasculares (2017)
3. Instituto Nacional de Estadística Geografía e Informática: Comunicado de prensa núm. 592/21 28 de octubre de 2021 pp. 2/4 (2021)
4. Instituto Nacional de Estadística Geografía e Informática: Estadísticas de defunciones registradas, enero-junio 2021, vol. 2021, pp. 1–40 (2022)
5. González-Cedillo, C. D.: Diagnóstico de enfermedades cardíacas con los algoritmos supervisados Naives Bayesian (2019)
6. Chicco, D., Jurman, G.: Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. BMC Med. Inform. Decis. Mak., vol. 20, no. 1, pp. 1–16 (2020)
7. Ali, F., El-Sappaghb, S., cS.M. Riazul-Islam, S. M., Kwak, D., Ali, A., Imran, M., Kwak, K. S.: A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion. Inf. Fusion, vol. 63, pp. 208–222 (2020)
8. Gallego-Valcárcel, D., Delly-Fabián, L. M.: Modelos De Aprendizaje Automático Para La Predicción Del Riesgo De Fatalidad Por Insuficiencia Cardiaca Con Datos Clínicos (2021)
9. Faiayaz-Waris, S., Koteeswaran, S.: Heart disease early prediction using a novel machine learning method called improved K-means neighbor classifier in python. Mater. Today Proc., pp. 1–7 (2021)

10. Álvarez, J., Hurtado, S., Trujillo, H.: Algoritmos genéticos (2010)
11. Kaggle, Teboul, A.: Heart Disease Health Indicators Dataset (2021)
12. Trevino, V., Falciani, F.: GALGO An R package for Genetic Algorithm Searches (Customized for Variable Selection in Functional Genomics) (2006)
13. Garduño Juárez, R.: Algoritmos genéticos (2018)
14. Rodríguez, D.: La regresión logística (2018)
15. Barrios, J.: La matriz de confusión y sus métricas (2019)
16. González, L.: Amazon Machine Learning, Guía para desarrolladores. pp. 94–96 (2016)
17. Irizarry, R.: Introducción a la Ciencia de Datos - Análisis de datos y algoritmos de predicción con R. CRC Press, pp. 538–539 (2021)
18. Amat, R.: Validación de modelos predictivos: Cross-validation, OneLeaveOut, Bootstraping (2020)

ISSN 1870-4069

Diseño IoT de invernadero para el control de variables mediante técnicas de inteligencia artificial

Lucero Ortiz-Aguilar, Luis Hernández-Silva,
Bernardo Muñoz-López, Alan Cortes-Ruiz

Tecnológico Nacional de México,
Instituto Tecnológico Superior de Purísima del Rincón,
México

lucero.oa@purisima.tecnm.mx

Resumen. En la actualidad el tema de optimización en recursos naturales es una preocupación que compete a diferentes áreas de la ciencia. En específico en lo que son los invernaderos o también conocidos como GreenHouse han sido una tendencia en la última década. Tanto tendencias de invernaderos en casa como a gran escala son de interés en la comunidad científica. En este trabajo con el objetivo de adaptar una metodología de control de invernaderos que pueda ser usado en un contexto local y regional se propuso un diseño basado en IoT y Técnicas de IA. Nuestra investigación tuvo como parte de los objetivos la investigación de diferentes diseños de invernaderos, diseños de riego, estudio de diferentes sensores y técnicas de IA. Primero se hizo un modelo a escala y posteriormente se tiene proyectado aplicarlo a una escala en el Instituto Tecnológico Superior de Purísima del Rincón (ITSPR). En este trabajo reportamos la metodología desempeñada y los resultados obtenidos del diseño propuesto.

Palabras clave: IoT, cloud computing, automatización, aplicaciones de la IA, greenhouse.

Greenhouse IoT Design for Variable Control through Artificial Intelligence

Abstract. One of the main topics of optimization in natural resources is related to the responsibility of different areas of science. Specifically, greenhouses have been a trend in the last decade. Both home and large-scale greenhouse trends are of interest in the scientific community. In this work, a proposal of a model for a greenhouse control methodology that can be used in a local and regional context, and a design based on IoT and AI techniques were proposed. Our research had as part of the objectives the investigation of different greenhouse designs, irrigation

designs, and study of different sensors, and AI techniques. First, a scale model was made, and later it is planned to apply it to a scale at the Instituto Tecnológico Superior de Purísima del Rincón (ITSPR). In this paper we report the methodology used and the results obtained from the proposed design.

Keywords: IoT, cloud computing, automation, artificial intelligence application, greenhouse.

1. Introducción

En la actualidad el tema de optimización en recursos naturales es una preocupación que compete a diferentes áreas de la ciencia. Por lo general, las áreas que se dedican a la agricultura plantean modelos a prueba y error de como aprovechar diferentes recursos como el agua. Una forma de coadyuvar en la medición y solución de este tipo de problemas es generar un simulador de invernadero que permita controlar las variables y además tenga la capacidad de escalamiento para medir un entorno real.

Los ambientes inteligentes, han crecido en popularidad en los últimos años [29,21,24]. De manera particular, la industria de los invernaderos y agricultura (Greenhouse conocido en inglés) ha ganado popularidad debido a su habilidad de producir vegetales frescos de manera muy rápida y a gran escala. De acuerdo con la ONAA (Organización de las Naciones Unidas para la Agricultura y la Alimentación) , la agricultura es una de las actividades que se busca sea sostenible, sustentable y con mayor productividad [12].

El tener invernaderos implica que diferentes expertos estén involucrados de manera activa para tengan un correcto funcionamiento. Lo anterior debido a que el invernadero necesita de ajustes de parámetros con valores óptimos y esto permita la producción de alimentos. Por lo tanto, un invernadero que sea capaz de simular o controlar las variables puede mejorar la producción de los alimentos y el costo de estos. La primera problemática a la que nos enfrentamos en la optimización de los recursos para los invernaderos es el tener una correcta medición de todos los factores influyentes, posteriormente un monitoreo y finalmente un control.

En este sentido, la parte de monitoreo y control se puede apoyar fuertemente en las nuevas tecnologías de Internet de las Cosas (IoT) e Inteligencia Artificial (IA). Es sumamente importante que a partir de una medición de datos se pueda continuar con un tratamiento adecuado de la información y generar una clasificación. En el estado del arte son pocas las bases de datos publicas relacionadas a cultivo de plantas que puedan dar una guía para el tratamiento de estos problemas.

Es por eso por lo que la medición y monitoreo depende de la infraestructura y de las plantas que se tengan en el invernadero, por lo tanto, la solución de estos también. En este caso, dado que se pretende que el diseño del invernadero

pueda contener diferentes especies de plantas, por lo cual es necesario considerar el incluir cada una como una clase diferente. En primera aproximación se necesita un sistema que recolecte los datos y los almacene en una base de datos. Posteriormente, la parte que es de nuestro interés es el generar o identificar estados o patrones dentro de la información obtenida, es decir, hacer un análisis de datos.

Ya que diferentes especies de plantas convivirán dentro de un mismo entorno que se desea este controlado. Una vez caracterizado el entorno e identificando (en clases), podemos hacer en una etapa posterior hace una clasificación y predicción de los ajustes en el invernadero. Es por lo anterior que se vio en la necesidad de aplicar una técnica de agrupamiento no supervisado, para identificar esos grupos y hacer posteriormente los ajustes adecuados para el correcto funcionamiento del ambiente. Como limitante nuestro proyecto es el buscar como generar escenarios en los cuales se optimicen los recursos como el agua.

En este trabajo mostramos un prototipo de un modelo basado en Inteligencia Artificial e Internet de las cosas que permite controlar el sistema de riego para un invernadero. Nuestra propuesta contiene diferentes etapas relacionadas al diseño, monitoreo, agrupación, clasificación y control. En primera instancia y como parte de este trabajo de investigación se creo un modelo a escala que nos permitió probar el sistema con las técnicas de inteligencia artificial.

Posteriormente considerando una infraestructura perteneciente al Instituto Tecnológico Superior de Purísima del Rincón (ITSPR), se hará una adaptación con la base de datos e interfaces electrónicas. Finalmente, las diferentes variables se guardaron en una base de datos que será el insumo primero de una etapa de clasificación no supervisada con el K-means y posteriormente se podrá predecir los valores adecuados para mantener el invernadero en óptimas condiciones.

2. Estado del arte

En esta sección se da una visión general de estado del arte en las investigaciones relacionadas a los invernaderos, Sistemas de Monitoreo de Riego y aplicaciones de la inteligencia Artificial a Invernaderos.

2.1. Sistemas de monitoreo de riego

Los invernaderos modernos requieren varios puntos de medición para la monitorización de parámetros climáticos internos como lo son temperatura, humedad, luminosidad en diferentes partes del invernadero en general para garantizar el funcionamiento adecuado del sistema y la automatización de este [3].

Por otra parte para mejorar el rendimiento agrícola con menos recursos y esfuerzos laborales, se han realizado innovaciones sustanciales a lo largo de la historia humana. Sin embargo, la alta tasa de población nunca permitió que la demanda y la oferta coincidieran durante todos estos tiempos.

Según las cifras pronosticadas, en 2050, se espera que la población mundial alcance los 9.800 millones, un aumento de aproximadamente el 25 % con respecto

a la cifra actual [5]. Dentro de las investigaciones y aplicaciones más recientes de sistemas de riego están la realizada por Guijarro-Rodríguez et al. en [15], cuyo trabajo se enfoca en realizar un sistema de riego automatizado con Arduino.

Su problemática inicial es el aprovechar el agua, para mantener una humedad adecuada para un conjunto de plantas. Su diseño es básico, con pocos elementos como: electroválvula, modulo relay, sensores DHT, pantallas, LCD, y sensores de humedad en tierra.

El sistema fue probado por un periodo corto de dos semanas, cuyas lecturas sirvieron para calibrar los sensores y que la automatización fuera lo más adecuada posible. Sin embargo, dicho diseño es para un huerto pequeño doméstico, el cual no es susceptible a ser escalable. En el trabajo realizado por [8] menciona que los invernaderos son estructuras de clima controlado con paredes y techo especialmente diseñadas para el cultivo de plantas fuera de temporada.

La mayoría de los sistemas de invernadero utilizan sistemas manuales para monitorear la temperatura y la humedad, lo que puede causar incomodidad al trabajador, ya que debe visitar el invernadero todos los días y controlarlos manualmente. Los sensores utilizados en el trabajo desarrollado por [8] son el sensor de humedad YL69 y el DHT11 (sensor de temperatura y humedad).

A partir de los datos recibidos, Raspberry PI3 controla automáticamente la humedad y la temperatura de manera eficiente dentro del invernadero al accionar una tubería de irrigación, un ventilador de enfriamiento y ventanas corredizas, respectivamente, de acuerdo con las condiciones requeridas de los cultivos para lograr el máximo crecimiento y rendimiento.

La temperatura y la humedad registradas se almacenan en una base de datos en la nube (ThingSpeak), y los resultados se muestran en una página web, desde donde el usuario puede verlos directamente. En este proyecto guiado por [25], usando un Arduino Mega como controlador para integrar datos del sensor y enviar la información a la plataforma IoT a través del módulo WiFi ESP8266-01 para el monitoreo de un jardín de hongos.

En este sistema, hay tres parámetros que se monitorean a una frecuencia de muestreo de 5 minutos; temperatura ambiente, humedad relativa e intensidad de la luz. Además de usar un navegador web, los datos se pueden duplicar y el usuario puede monitorearlos a través de aplicaciones de terceros en Android.

El objetivo principal del estudio sistemático realizado por [11] es la recopilación de toda la investigación relevante sobre aplicaciones agrícolas, sensores/dispositivos, protocolos de comunicación y tipos de red de IoT. También en dicho trabajo analizan los principales temas y desafíos que se están investigando en el campo de la agricultura.

Además, se ha presentado un marco de agricultura IoT que contextualiza la representación de una amplia gama de soluciones actuales en el campo de la agricultura. Del mismo modo, también se han presentado las políticas de los países para la agricultura basada en IoT . Por último, se han presentado problemas abiertos y desafíos para proporcionar a los investigadores direcciones futuras prometedoras en el dominio de la agricultura de IoT.

En la investigación [26] se muestra el desarrollo de un sistema de monitoreo de invernadero basado en sensores. El sistema está relacionado a un control visual inteligente de invernadero controlado por red inalámbrica que se basa en la Web e Internet, comentan sobre una posibilidad de que el sistema aproxime la tasa de crecimiento de la planta mediante el uso de datos recopilados de sensores que traten las características de los parámetros ambientales.

2.2. IoT aplicado a GreenHouse

De acuerdo con [31] con las tendencias tecnológicas moviéndose hacia el Internet de las cosas (IoT), la mayoría de las aplicaciones de monitoreo continuo están migrando de redes cableadas a comunicaciones inalámbricas de corto alcance y baja potencia. El artículo presenta el diseño y la evaluación del desempeño de un sistema modular y flexible para el monitoreo remoto y continuo de invernaderos. La red de sensores inalámbricos tiene las principales ventajas de estructura simple, alta eficiencia, bajo costo, seguridad y confiabilidad en comparación con la tecnología de monitoreo por cable existente.

El Internet de las Cosas es una tecnología que conecta dispositivos electrónicos, sensores, y dispositivos electrónicos, sensores que pueden ser utilizados en la agricultura para el manejo de cultivos porque es fácil de comercializar y puede diseñarse para ser implementado por técnicos calificados que tienen un conocimiento limitado de la información, tecnología, sensores, dispositivos que pueden detectar lo que ocurre en el invernadero y tomar decisiones con base a los datos obtenidos [20].

La tolerancia a la falla es un aspecto distintivo de la arquitectura propuesta de los sistemas anteriores descritos en la nube utilizando el puerto web para garantizar un sistema, la escuela informática, la industria mecánica y la industria de la agricultura puede exclusivamente como una combinación de plantas de invernadero, sensores para detectar invernadero y los parámetros de microcontrolador como sistemas de recolección de datos, las características de este modelo se pueden realizar mediante el desarrollo de la orientación en la agricultura.

El principal objetivo de investigaciones recientes como la realizada por [5] es desarrollar un invernadero inteligente práctico con sistemas de control inteligente para obtener las circunstancias adecuadas. El sistema propuesto tiene la capacidad de monitorear y controlar el invernadero desde cualquier lugar del mundo. El siglo XXI se convirtió en el inicio del desarrollo de las tecnologías de la información, donde una de las revoluciones fue la Presencia del Internet de las Cosas. Internet de las Cosas es una tecnología que combina dispositivos electrónicos, sensores [4].

Esta tecnología puede adoptarse en la agricultura para el manejo de cultivos como son los invernaderos que utilizan un microcontrolador Arduino o un microordenador Raspberry Pi. Estos dispositivos se utilizan porque su precio es bajo y es fácil de comercializar y puede diseñarse para que los técnicos con conocimientos limitados de tecnología de la información puedan ejecutarlo.

El hardware puede detectar lo que sucede en el invernadero y tomar decisiones basadas en los datos adquiridos, se utilizan a menudo en la agricultura de

precisión sensores de temperatura y humedad, sensores de humedad del suelo y sensores de luz, donde los datos adquiridos por el hardware se transmitirán de forma inalámbrica. En algunos lugares con climas extremos como el Reino de Arabia Saudita (KSA) enfrenta varias limitaciones, que incluyen temperaturas extremas, escasez de agua, costos de desalinización del agua de mar y suelos no fértiles.

En el clima desértico donde el verano dura más de la mitad del año, como en Arabia Saudita. La temperatura media en julio ronda los 43 ˚C, y la media en enero ronda los 14 C. Es imposible potenciar la producción de verduras y frutas como tomates, pepinos, pimientos dulces y fresas, ya que la temperatura óptima para su crecimiento se encuentra en el rango de 11 C a 28 C.

El sistema propuesto en la investigación realizada por [30] se considera inteligente porque es capaz de monitorizar, de forma autónoma, la temperatura exterior y el consumo de energía en horas punta, para generar con precisión la temperatura de referencia adecuada y garantizar que la temperatura del invernadero alcance esta temperatura de referencia. Además, este sistema puede identificar el ángulo de los rayos del sol para controlar la apertura y el cierre de los toldos, lo que se traduce en reducir los efectos de las altas temperaturas.

Fig. 1. Invernadero.

Todos estos parámetros capturados relacionados con la temperatura y la energía se registran para futuros análisis y predicciones en un modelo de datos de gráficos dinámicos que se utiliza para diseñar el almacenamiento de back-end del sistema.

3. Marco teórico

Algunos conceptos básicos para la compresión de los Invernaderos con Iot los veremos de forma breve a continuación.

3.1. Invernadero

De acuerdo con [6] establece que: un invernadero está constituido por una estructura metálica o de plástico, la cual se le cubre con materiales translucidos para obtener la suficiente luminosidad en su interior. Un ejemplo de lo anterior se puede observar en la figura 1.

3.2. Riego por aspersión

De acuerdo con [1] existen dos tipos de riego por aspersión:

1. Sistemas estacionarios que permanecen en la misma posición mientras dura el riego. Es decir,se basan en un sistema el cual simula la forma de lluvia intensa en el cultivo y tiene por objetivo que se infiltre el agua en un punto donde debe caer. Este tipo de sistema requiere de una red que distribuya el agua y llegue con suficiente presión a cada aspersor para que realicen su función adecuada.
2. Sistemas mecanizados que se desplazan mientras aplican el agua de riego. El movimiento de estos sistemas es de forma circular y necesitan de una toma de corriente eléctrica para poder funcionar.

Las ventajas y desventajas de los sistemas de riego son:

– Son capaces de cubrir grandes distancias de terreno con bajos costos.
– Se adaptan al tipo de parcela.
– La vida útil es mucho mayor a la de otros sistemas.
– La instalación de este sistema es mucho más compleja.
– Incrementa la aparición de maleza.
– La mano de obra es más elevada.

3.3. Sensores

Una parte primordial de un diseño de IoT es el uso de sensores conectados a una red. De acuerdo con [33], un sensor es un dispositivo que mide un fenómeno dado y da como resultado una señal de salida. Un sensor nos permite adquirir o detectar cantidades físicas que por su naturaleza o tamaño no pueden ser percibidas directamente por el ser humano. Los sensores usados en el sistema de esta propuesta fueron los siguientes:

Sensor de temperatura DHT11. El DHT11 está compuesto por un sensor de humedad y temperatura con una señal digital de salida ofrece una alta fiabilidad y una excelente estabilidad a largo plazo [14]. Se compone de un sensor tipo NTC el cual presenta un rango de temperatura que va desde los (0°C a 50°C).

Para poner a funcionar el sensor DHT11 es necesario realizar las conexiones, debido a que es un sensor sencillo solo cuenta con cuatro pines donde uno es de voltaje V/cc, otro es de señal E/S, un pin N/C que no se conecta y un GND que es la conexión a tierra.

La salida del sensor DHT11 debe ser conectada a las entradas digitales de la placa Arduino y de acuerdo con el código de programación se pueden analizar los datos con los cuales se establecerán acciones como activar un sistema de ventilación para reducir el aumento de temperatura.

Sensor de humedad de suelo YL69. De acuerdo con [2] dicho sensor tiene la capacidad de medir la humedad del suelo. Lo anterior se hace midiendo la tensión entre dos terminales que permiten el paso de corriente de acuerdo con la resistencia del suelo. Cuando la tierra reduce o aumenta la humedad que exista en ella, el sensor comienza a detectar la diferencia de capacitancia por lo que manda cada valor censado como dato a la salida.

Para poder analizar los valores obtenidos por el sensor es necesario conectar la salida del sensor a las entradas analógicas o digitales de la placa Arduino ya que con el código de programación se podrá apreciar las mediciones censadas y poder establecer funciones como la activación del riego a partir de los rangos requeridos por cada cultivo.

3.4. K-Means

Los algoritmos de agrupamiento son herramientas que permiten la extracción y compresión de datos, estimación de densidades de probabilidad, entre otras cosas [17]. El K-Means es un algoritmo de agrupamiento, que ha sido utilizado en diferentes aplicaciones de Machine Learning[10], reconocimiento de patrones, entre otros. Como se ha mencionado anteriormente, el objetivo de esta investigación es identificar grupos o patrones de información de los datos recolectados en la base de datos.

Por lo anterior el K-Means es una herramienta que permite generar un determinado número de grupos los cuales tienen características afines. Una de las ventajas de trabajar con K-Means es que los grupos se van ajustando de forma adecuada a través de un proceso iterativo, lo cual lleva a que converja el algoritmo. Otra ventaja es que cada uno de los grupos formados junto con los centroides pueden ser analizados.

Además de que parte de las bondades de este algoritmo es que podamos trabajar con diferentes métricas, como lo son la distancia euclidiana, Mahalanobis, entre otras [23]. Existen otros algoritmos de agrupamiento como MinMax distancia [32], Gaussian mixture [18], Principal component análisis [22], entre otros.

Es importante mencionar que debido a que los datos serán procesados en la nube se opto por elegir el K-Means como algoritmo de agrupamiento, ya que computacionalmente es más sencillo de procesar que los anteriores mencionados. Sin embargo, tiene sus correspondientes limitantes y se espera que se opte en una etapa posterior por un algoritmo más sofisticado.

4. Sistema de control automatizado mediante IoT

De acuerdo con diversas investigaciones presentadas[19], [28], los factores esenciales en la climatología y la ecología van a definir las características de

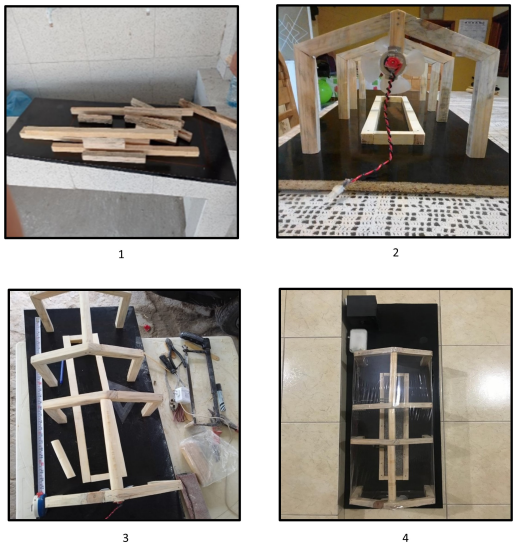


Fig. 2. Armado de invernadero a escala.

ciertas zonas y holgura para los cultivos en el invernadero de forma económica. El primer factor que es primordial para el desarrollo óptimo del invernadero, es la reducción de los cambios de temperatura ambiental.

El nivel térmico no es un problema trivial de resolver, ya que depende de las plantas y el clima del lugar. El crecimiento junto con el metabolismo en las plantas se le atribuyen generalmente a la temperatura a la que se encuentre dentro del invernadero, ya que no hay tejido o proceso fisiológico que no esté influenciado.

De acuerdo con Araque et al. [13] es importante considerar la humedad del substrato y de la atmosfera para la climatización de un invernadero. Los sistema de riego y la humidificación de la atmosfera se diseñan ad-hoc sobre los requerimientos de las especies cultivadas, así como también sobre las condiciones del clima del invernadero.

De acuerdo con [9] la ventilación natural es el proceso en donde se produce el intercambio entre el aire interior de una estructura y el aire exterior. Lo anterior debido a las diferencias de presión que determinan que el aire entre o salga y eso se provoca mediante dos fuerzas motoras que son la fuerza de gravedad y el efecto del viento exterior.

4.1. Diseño metodológico

Existen diferentes modelos en el estado del arte que han reportado resultados favorables para el cultivo de plantas [34][29]. En nuestra propuesta consta de dos componentes importantes hoy en día en la industria 4.0 que son el internet de las

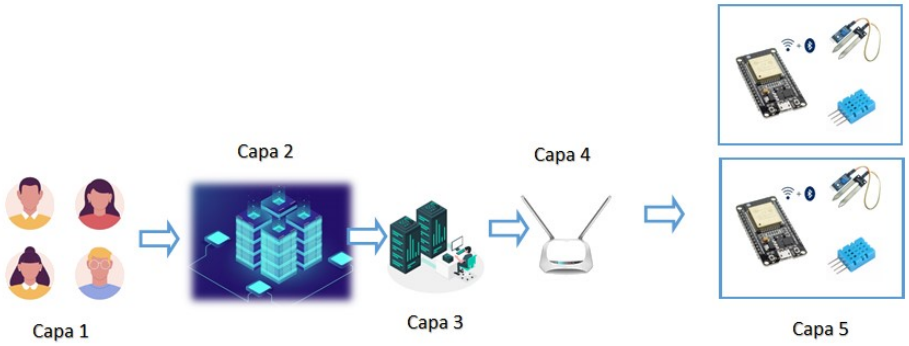


Fig. 3. Modelo del Invernadero de IoT.

cosas (IoT) y las técnicas de aprendizaje. De forma general nuestra metodología de diseño del sistema inteligente es el siguiente:

1. **Requerimientos**. En esta etapa se realizo el diseño de un prototipo a escala para probar el modelo de IoT y poder determinar que las variables de temperatura y humedad fueran controlados de forma adecuada.
2. **Diseño**. El diseño de nuestro sistema consta de cinco capas o elementos, dicho diseño se muestra en la figura 3 (fuente elaboración propia). La parte de interfaz de usuario es en la cual el usuario puede ver las variables de humedad y temperatura del invernadero a distancia. La segunda es la parte de inteligencia artificial u optimización, la cual se va ajustando a cada uno de los elementos o plantas dentro del invernadero de acuerdo con la información recolectada.
   La tercera y cuarta parte es el manejo de los dispositivos inalámbricos en el servidor en este caso será el Oracle y el gateway. Finalmente, cada uno de los dispositivos que denominaremos “agentes” controlarán un conjunto de sensores en red.
3. **Implementación**.El armado del prototipo se puede ver en la figura 2, el cual implico una fase de diseño a escala el cual es aplicable al invernadero en tamaño real.
4. **Verificación**. De acuerdo con la programación y al montaje de los componentes del circuito, se obtuvo como resultado el correcto funcionamiento del sistema dado que los sensores censan de manera correcta las variables (humedad, temperatura) permitiendo que los motores tanto el de la bomba de agua como el del ventilador realicen su función establecida mediante la programación.
5. **Plan de mantenimiento**. Este se realizará de acuerdo a la instalación real.

Fig. 4. Proceso de medición, procesamiento de datos.

5. Resultados

Para este trabajo primero se hizo un diseño de Sistema de riego el cual se desarrollo e implemento en un prototipo. La fase de elaboración del prototipo se hizo a escala como se puede observar en la figura 2. Este prototipo también se considero para ser implementado o colocado en ambientes cerrados, ya que en la actualidad es una tendencia que vemos en diversas investigaciones [27].

Los materiales que conforman el diseño de este invernadero son adaptables a espacios reducidos y con poca ventilación. Posteriormente, en otra etapa de diseño del prototipo se planteó una etapa de conexión a internet, la cual involucraría el monitorear y controlar diferentes variables como lo son la humedad del suelo, ambiente, nivel de CO_2 y temperatura. Para esta fase se utilizo un diseño y uso de los siguientes dispositivos: ESP32, DTH11, MQ135, YL69.

La información es tratada en el ESP32, posteriormente es mandada vía MQTT al servidor de Oracle el cual se encuentra en línea a una base de datos en PostgreSQL. Finalmente, la información que obtenga se caracterizará de acuerdo con periodos estacionales y se generaran al menos 3 clases. El proceso descrito anteriormente se puede observar en la figura 4.

6. Resutaldos preliminares con K-Means

Las mediciones del invernadero real se realizarán posteriormente a la instalación del invernadero en la escala real. Como parte preliminar se descargo la base de datos ROSES GREENHOUSE CULTIVATION DATABASE [7], para probar tanto el procesamiento de la información como retroalimentación y como la predicción de las condiciones adecuadas de los mismos. La base de datos Roses Greenhouse contiene un total de 4 clases que corresponden a:

- Ambiente sin agua.
- Ambiente correcto.
- Ambiente con mucho calor.
- Ambiente con mucho frio.

Se eligió el número de K=4, esto de acuerdo con cada uno de los escenarios planteados en la base de datos, y que son los que se plantean adecuar en nuestra experimentación. Se utilizo el software Weka [16], y se realizaron experimentaciones con 4 clusters lo cual es lo que muestra en la tabla 1.

En esta previa experimentación probamos que el modelo en conjunto con la parte de IoT es compatible y funcional. Estamos en etapa de armado y recolección de datos, por lo cual dicha experimentación esta en proceso.

Tabla 1. Resultados con k-Means.

		Cluster			
		0	1	2	3
		-0.32	-0.09	-0.16	-0.42
HS-Analog	Media	477.1401	369.0086	557.5339	773.6188
	DE	74.7067	195.4082	12.3686	22.7879
Luz	Media	13735.196	12355.2316	4157.4169	6498.48
	DE	14685.3859	16738.1693	2599.781	16288.6699
Temperatura	Media	27.8942	23.8652	21.7575	21.8978
	DE	5.627	1.9749	1.107	3.6495
CO_2	Media	120.7363	125.2128	163.7696	143.0185
	DE	47.7334	77.6604	9.8825	21.4481
HRAnalog	Media	66.7095	75.3229	88.0223	85.8679
	DE	16.026	3.1848	2.4419	11.4575
	Clase				
	1	1.0061	1.0037	1	125.9902
	2	1.1489	28.8511	1	1
	3	90.7825	1.2175	1	1
	4	8.5599	1.1863	50.2537	1
	total	101.4975	32.2586	53.2537	128.9902

7. Conclusiones

La integración de diferentes elementos como lo son el IoT, técnicas de Inteligencia Artificial y automatización nos acercan cade vez más a lo que es la Industria 4.0. Por lo tanto, la aplicación de estos elementos a un sector como lo son los invernaderos es de suma importancia, debido a su impacto en el medio ambiente y la economía. En cuanto a la parte de IoT y las técnicas de IA, en este trabajo se decidió optar por un algoritmo no supervisado ya que se requería solo ajustar a cuatro estaciones o tipos de contextos climáticos.

Es importante mencionar que como trabajo futuro se quiere ampliar a su uso o aplicación de otros algoritmos de aprendizaje como lo son redes neuronales o algoritmos de clasificación. Como trabajo futuro, está en ampliar la cantidad de variables a medir. Es importante señalar que estamos en una etapa de proceso de armado y recolección de datos, por lo cual dicha experimentación esta por complementarse.

Referencias

1. Ventajas y desventajas del riego por aspersión. Investigación Aplicada, vol. 1, no. 17 (2002)
2. Adla, S., Rai, N. K., Karumanchi, S. H., Tripathi, S., Disse, M., Pande, S.: Laboratory calibration and performance evaluation of low-cost capacitive and very low-cost resistive soil moisture sensors. Sensors, vol. 20, no. 2, pp. 363 (2020)
3. Akkaş, M. A., Sokullu, R.: An IoT-based greenhouse monitoring system with micaz motes. Procedia computer science, vol. 113, pp. 603–608 (2017)

4. Ardiansah, I., Bafdal, N., Suryadi, E., Bono, A.: Greenhouse monitoring and automation using Arduino: a review on precision farming and internet of things (IoT). International Journal on Advanced Science, Engineering and Information Technology, vol. 10, no. 2, pp. 703–709 (2020)
5. Ayaz, M., Ammad-Uddin, M., Sharif, Z., Mansour, A., Aggoune, E.-H. M.: Internet-of-things (IoT)-based smart agriculture: Toward making the fields talk. IEEE access, vol. 7, pp. 129551–129583 (2019)
6. Barraza Alvarez, F. V.: Importancia de un invernadero. Temas Agrarios, vol. 17, no. 2, pp. 18–29 (2012)
7. Champutiz, W., Rosero-Montalvo, P., Fuentes, E., Peluffo, D.: Roses greenhouse cultivation database repository (RosesGreenhDB) (2019)
8. Danita, M., Mathew, B., Shereen, N., Sharon, N., Paul, J. J.: IoT based automated greenhouse monitoring system. In: 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS). pp. 1933–1937. IEEE (2018)
9. Díaz-Sarmiento, H. O., Solano-Rojas, O. F.: Diseño y simulación del control climático para un invernadero y base de datos de registro, (2006)
10. El Khediri, S., Fakhet, W., Moulahi, T., Khan, R., Thaljaoui, A., Kachouri, A.: Improved node localization using K-Means clustering for wireless sensor networks. Computer Science Review, vol. 37, pp. 100284 (2020)
11. Farooq, M. S., Riaz, S., Abid, A., Umer, T., Zikria, Y. B.: Role of IoT technology in agriculture: A systematic literature review. Electronics, vol. 9, no. 2, pp. 319 (2020)
12. Food, of the United Nations, A. O.: (2022)
13. Galindo-Araque, D. S., Vargas-Sarmiento, M. C., Corredor-Gómez, J. P.: Caracterización de temperatura y humedad de suelos agricolas. Letras ConCiencia TecnoLógica, , no. 16, pp. 24–31 (2017) doi: 10.55411/26652544.129
14. Gay, W.: DHT11 sensor. Advanced Raspberry Pi, pp. 399–418 (2018)
15. Guijarro-Rodríguez, A., Cevallos-Torres, L., Preciado-Maila, D., Zambrano-Manzur, B. N.: Sistema de riego automatizado con arduino. Sistema, vol. 39, no. 37, pp. 27 (2018)
16. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I. H.: The WEKA data mining software: an update. ACM SIGKDD explorations newsletter, vol. 11, no. 1, pp. 10–18 (2009)
17. Hamerly, G., Elkan, C.: Learning the k in K-Means. Advances in neural information processing systems, vol. 16 (2003)
18. He, X., Cai, D., Shao, Y., Bao, H., Han, J.: Laplacian regularized gaussian mixture model for data clustering. IEEE Transactions on Knowledge and Data Engineering, vol. 23, no. 9, pp. 1406–1418 (2010)
19. Hernández-Sánchez, E. A.: Diseño e implementación de un invernadero inteligente a escala con dimensionamiento fotovoltaico para su posible sostenimiento eléctrico. Thesis, Benemérita Universidad Autónoma de Puebla (2021)
20. Ibrahim, H., Mostafa, N., Halawa, H., Elsalamouny, M., Daoud, R., Amer, H., Adel, Y., Shaarawi, A., Khattab, A., ElSayed, H.: A layered iot architecture for greenhouse monitoring and remote control. SN Applied Sciences, vol. 1, no. 3, pp. 1–12 (2019)
21. Jamil, F., Ibrahim, M., Ullah, I., Kim, S., Kahng, H. K., Kim, D. H.: Optimal smart contract for autonomous greenhouse environment based on IoT blockchain network in agriculture. Computers and Electronics in Agriculture, vol. 192, pp. 106573 (2022)

22. Kurita, T.: Principal component analysis (PCA). Computer Vision: A Reference Guide, pp. 1–4 (2019)
23. Likas, A., Vlassis, N., Verbeek, J. J.: The global K-Means clustering algorithm. Pattern recognition, vol. 36, no. 2, pp. 451–461 (2003)
24. Maraveas, C., Bartzanas, T.: Application of internet of things (IoT) for optimized greenhouse environments. AgriEngineering, vol. 3, no. 4, pp. 954–970 (2021)
25. Mohammed, M., Azmi, A., Zakaria, Z., Tajuddin, M., Isa, Z., Azmi, S.: IoT based monitoring and environment control system for indoor cultivation of oyster mushroom. In: Journal of Physics: Conference Series. vol. 1019, pp. 012053 (2018)
26. Nosirov, K., Begmatov, S., Arabboev, M., Kuchkorov, T., Chedjou, J., Kyamakya, K., De Silva, P., Abhiram, K.: The greenhouse control based-vision and sensors. In: Developments of Artificial Intelligence Technologies in Computation and Robotics: Proceedings of the 14th International FLINS Conference. pp. 1514–1523 (2020)
27. Paz, M., Fisher, P. R., Gómez, C.: Minimum light requirements for indoor gardening of lettuce. Urban Agriculture & Regional Food Systems, vol. 4, no. 1, pp. 1–10 (2019)
28. Pérez-Monsalve, J.: Un invernadero inteligente para optimizar los cultivos. Revista Universidad EAFIT, vol. 54, no. 173, pp. 136–139 (2019)
29. Raj, J. S., Ananthi, J. V.: Automation using IoT in greenhouse environment. Journal of Information Technology, vol. 1, no. 1, pp. 38–47 (2019)
30. Subahi, A. F., Bouazza, K. E.: An intelligent iot-based system design for controlling and monitoring greenhouse temperature. IEEE Access, vol. 8, pp. 125488–125500 (2020)
31. Tafa, Z., Ramadani, F., Cakolli, B.: The design of a ZigBee-based greenhouse monitoring system. In: 2018 7th Mediterranean Conference on Embedded Computing. pp. 1–4 (2018)
32. Tzortzis, G., Likas, A.: The minmax k-means clustering algorithm. Pattern recognition, vol. 47, no. 7, pp. 2505–2516 (2014)
33. Villalba-Hernández, C. E., Mocencahua-Mora, D., Sánchez-Gaspariano, L. A.: Tinkercad como alternativa para aprender conceptos básicos de electrónica. RD-ICUAP, vol. 7, no. 20, pp. 133–139 (2021)
34. Wang, J., Chen, M., Zhou, J., Li, P.: Data communication mechanism for greenhouse environment monitoring and control: An agent-based iot system. Information Processing in Agriculture, vol. 7, no. 3, pp. 444–455 (2020)

ISSN 1870-4069

Balanceo de clases mediante evolución diferencial

Rafael Muñoz-Cervantes, Efrén Mezura-Montes,
Héctor-Gabriel Acosta-Mesa

Universidad Veracruzana,
Instituto de Investigaciones en Inteligencia Artificial,
México

rafamucerv@gmail.com, {emezura, heacosta}@uv.mx

Resumen. El desbalanceo de clases en bases de datos suele ser muy común cuando nos enfrentamos a problemas de la vida real, por lo que se han desarrollado diversas técnicas para dar solución. En este trabajo se aborda el problema de balanceo de clases combinando técnicas de muestreo mediante Evolución Diferencial. La propuesta se compara contra técnicas clásicas de balanceo de clases y también se comparan variantes de Evolución Diferencial. Los resultados muestran que nuestra propuesta logra encontrar el método con sus hiperparámetros que mejor se desempeña para las bases de datos probadas.

Palabras clave: Datos desbalanceados, evolución diferencial, balanceo de clases.

Class Balancing Using Differential Evolution

Abstract. Class imbalance in databases is very common when we face real-life problems, then several techniques have been developed to provide a solution. In this work, we address the balancing class problem by combining sampling techniques by means of Differential Evolution. We compare the proposal against classical balancing techniques and Differential Evolution variants are also compared. The results show that our proposal manages to find the method with its hyperparameters that performs best for the tested databases.

Keywords: Imbalance data, differential evolution, class balancing.

1. Introducción

El aprendizaje automático ha logrado ocupar un lugar muy importante en áreas científicas e industriales debido a la flexibilidad y capacidad de contender con problemas relacionados a la vida real gracias a la existencia de distintas

técnicas capaces de hacer frente a la diversidad de situaciones con los que nos podemos encontrar. Uno de los problemas que aparece frecuentemente cuando se trabaja con bases de datos creadas a partir de problemas reales, es el desbalance.

Una base de datos se dice que está desbalanceada cuando el número de instancias pertenecientes a una o más clases es menor (*clase minoritaria*) en comparación con el resto de las clases disponibles (*clase mayoritaria*).

Este tipo de bases de datos se pueden encontrar en distintos campos de la computación como la Visión Computacional [4,5,9] y la Seguridad Informática [13], así como en tareas relacionadas con la detección de fraudes bancarios [20], clasificación de textos [15], predicción de mantenimiento [14] o detección de cáncer [21].

Las complicaciones que conlleva el aprendizaje a partir de una base de datos desbalanceada se ven reflejadas en el desempeño de los clasificadores, pues se crea un sesgo hacia la clase mayoritaria, por lo que muchos de los clasificadores obtienen un mal desempeño respecto a la clase minoritaria, la cual, suele ser la clase de interés en gran parte de las aplicaciones de la vida real. Para hacer frente a este problema, se han realizado contribuciones que logran contrarrestar los efectos de una base de datos desbalanceada al realizar una tarea de clasificación.

Estas aproximaciones se pueden dividir en 4 grupos: métodos de preprocesamiento, métodos de aprendizaje sensibles al costo, métodos centrados en algoritmos y métodos híbridos [8]. Dentro de los métodos de preprocesamiento se encuentran los métodos de selección de atributos y los métodos basados en muestreo, siendo estos últimos de nuestro interés debido a su popularidad [6], así como su efectividad para abordar el problema de desbalanceo de datos al realizar tareas de aprendizaje supervisado [3].

Los métodos basados en muestreo se dividen en tres clases: hybridsampling, oversampling y undersampling. Estas aproximaciones buscan aumentar el número de instancias de la clase minoritaria (oversampling), disminuir el número de instancias de la clase mayoritaria (undersampling) o realizar ambos procedimientos (hybridsampling).

Dentro de la literatura revisada se han encontrado más de 50 ejemplos de métodos basados en muestreo de las distintas aproximaciones mencionadas [6,17,19], sin embargo, de toda la literatura revisada, no logramos detectar alguna propuesta sobre la forma de elegir un método de muestreo para una base de datos dada, lo que encontramos fue que usualmente se seleccionan los métodos más populares y se fijan sus parámetros sin un criterio específico.

Como resultado del intento por mejorar los métodos de balanceo basados en muestreo, así como también mejorar la selección de estos mismos para su aplicación, se ha realizado una combinación entre este campo del aprendizaje automático con técnicas alternativas de optimización como la computación evolutiva. La combinación de estas áreas ha demostrado tener un desempeño significativamente mejor, en comparación con las técnicas de aprendizaje por sí solas [18]. Un ejemplo de esto se puede ver en [10], donde se combina la optimización por enjambre de partículas para realizar tareas de optimización

de parámetros usados en técnicas de oversampling (SMOTE) y selección de instancias para realizar undersampling a una base de datos.

Otro ejemplo es el que se muestra en [22] donde se hace uso del algoritmo *Cross generational elitist selection Heterogeneous recombination Cataclysmic mutation algorithm* (CHC) para realizar una selección de instancias de la clase mayoritaria de tal forma que se realice un proceso de undersampling y las instancias seleccionadas sean lo más representativas posibles.

Tabla 1. Representación del individuo.

Método Oversampling	$\%\rho_O$	k_O	μ	**Método Undersampling**	$\%\rho_U$	k_U	**Orden**
$\{1,2,3\}$	$[0,1]$	$\{1,...,30\}$	$\{1,...,20\}$	$\{1,2,3\}$	$[0,1]$	{1,...,30}	{1,2,3,4}

También se ha usado el algoritmo de Evolución Diferencial para inspirar la creación de nuevas instancias en procesos de oversampling de la misma forma en la que ED explora el espacio de soluciones [7,**?**]. Por último, en [12] se propone el uso de la tercera versión del algoritmo *Non-Dominated Sorting Genetic Algorithm* (NSGA-III) para seleccionar un método de balanceo con sus respectivos hiperparámetros dada una base de datos desbalanceada, mediante una aproximación de mucho objetivos, donde se hace uso de cinco métricas relacionadas con distintas formas para medir el desempeño de los métodos.

Dentro de todas estas propuestas y aproximaciones se puede notar que a excepción de la última mencionada, no se realiza una selección entre distintos métodos de balanceo, simplemente se toma un método y se optimizan sus hiperparámetros o se usa el algoritmo bio-inspirado para seleccionar un número de instancias de forma que se realice undersampling al eliminar las instancias no seleccionadas. En la última propuesta mencionada ([12]) se realiza una selección del método de balanceo, sin embargo no se puede seleccionar la aproximación de hybridsampling, puesto que sólo se puede seleccionar un método de todas las opciones que se muestran.

Dado lo anterior, proponemos la selección de un método de balanceo con el cual se pueda introducir una base de datos desbalanceada y por medio de Evolución Diferencial, se realice la selección de una de las tres distintas aproximaciones basadas en muestreo (hybridsampling, oversampling y undersampling), así como la elección de sus respectivos hiperparámetros.

Para evaluar el método propuesto se seleccionan diez bases de datos desbalanceadas con un distinto nivel de desbalanceo medido por el radio de desbalanceo (IR) y se compara con todos los métodos usados de forma individual con los parámetros por defecto, de igual forma se comparan con dos algoritmos de balanceo que usan la aproximación hybridsampling con sus parámetros por defecto.

El resto de este trabajo se organiza como sigue. La Sección 2 presenta la propuesta de forma detallada tomando en cuenta la representación de soluciones, los métodos de balanceo seleccionados y la función de evaluación que se usa. La

Sección 3 muestra los experimentos y resultados, donde se habla de las bases de datos seleccionadas, los parámetros utilizados, así como la forma en que se desarrolló la experimentación. En la Sección 4 se realiza una discusión de los resultados obtenidos. Por último, en la Sección 5 se presentan las conclusiones de los resultados obtenidos y se habla del trabajo futuro.

2. Propuesta

El método propuesto para seleccionar la técnica de balanceo basada en muestreo para una base de datos dada, hace uso de Evolución Diferencial para explorar el espacio de soluciones, dicho espacio está compuesto por la combinación de las técnicas disponibles con sus respectivos parámetros. En esta sección se muestra un resumen del algoritmo evolutivo usado, así como dos puntos importantes dentro de esta propuesta que son la representación y la función de aptitud o desempeño.

2.1. Evolución diferencial

La evolución diferencial es un algoritmo de optimización propuesto por Storn y Price en 1995 [16]. Este algoritmo de optimización pertenece a la clase de algoritmos basados en poblaciones ya que se explora el espacio de búsqueda mediante la creación aleatoria de un conjunto de soluciones llamado población y mediante la aplicación de operadores de variación se generan nuevas soluciones para muestrear el espacio de búsqueda. Las operaciones que se aplican a los individuos de la población son: mutación, cruza y selección.

Las primeras dos operaciones manejan a los individuos representados mediante vectores y la última operación requiere de la evaluación de los individuos en una función que determinará el valor a optimizar, esta función se llama función de aptitud o desempeño. En el Algoritmo 1 se detalla el procedimiento utilizado por Evolución Diferencial. Dadas las necesidades de estas operaciones, se requiere una forma de representar a los individuos mediante un vector y establecer una función de desempeño para evaluar las soluciones, estos dos puntos se abordarán a continuación.

2.2. Representación

En el vector, las posiciones 1 a la 4 seleccionan el método de oversampling y sus respectivos parámetros, de la 5 a la 7 seleccionan el método de undersampling con sus parámetros, y por último, la posición 8 selecciona el orden en el que se aplicarán estos métodos. Dado que en el algoritmo de Evolución Diferencial los individuos se representan con vectores de números reales, para las variables que toman valores discretos se considera solamente la parte entera del valor seleccionado. La forma en que se representan los individuos está establecida por un vector con 8 elementos como se muestra en la Tabla 1. Para oversampling se

Algoritmo 1: Evolución Diferencial

NP: Tamaño de la población
NG: Número de generaciones
F: Factor de mutación
CR: Probabilidad de cruza
Se genera una población de tamaño NP de forma aleatoria con dimensión D dentro de los limites establecidos
$G = 0$
mientras $G < NG$ **hacer**
  para $i = 1$ *a* NP **hacer**
    Seleccionar tres individuos X_{r_1}, X_{r_2} y X_{r_3} distintos entre sí y que no sean el elemento i seleccionado
    Inicia Mutación:
    $V_i^G = X_{r_1}^G + F \cdot (X_{r_2}^G - X_{r_3}^G)$
    Inicia Cruza: $ea(1, D)$ es un entero aleatorio entre 1 y D, $a[0, 1)$ es un número aleatorio entre 0 y 1
    $j_{aleatorio} = ea(1, D)$
    para $j = 1$ *a* D **hacer**
      si $a[0, 1) \leq CR$ *o* $j = j_{aleatorio}$ **entonces**
        $U_{i,j}^G = V_{i,j}^G$
      en otro caso
        $U_{i,j}^G = X_{i,j}^G$
      fin
    fin
    Inicia Selección: $f(X)$ es la evaluación de X en la función de desempeño
    si $f(U_i^G) \leq f(X_i^G)$ **entonces**
      $X_i^{G+1} = U_i^G$
    en otro caso
      $X_i^{G+1} = X_i^G$
    fin
  fin
  $G = G + 1$
fin
devolver *Mejor individuo de la población*

seleccionaron los métodos 1: Random Oversampling, 2: SMOTE y 3: Borderline-SMOTE. Estos tres métodos de balanceo reciben como parámetro el porcentaje de radio de balanceo ($\%\rho_O$) que va a usar, lo que se traduce en el número de instancias por aumentar. El elemento k_O se usa en SMOTE y Borderline-SMOTE, donde representa el número de k vecinos a tomar en cuenta para aplicar el método.

Por último, Borderline-SMOTE usa el parámetro μ para seleccionar el número de puntos a tomar en cuenta para detectar si se encuentra un borde. Para undersampling, se seleccionaron los métodos 1: Random Undersampling, 2: NearMiss y 3: Cluster Centroids. De la misma forma que sucede en oversampling,

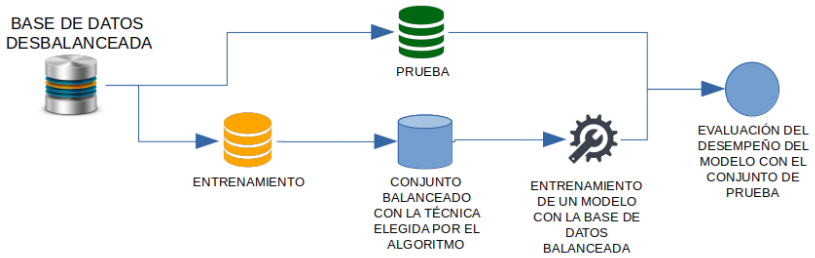


Fig. 1. Diagrama sobre la forma en que se evalúa el desempeño de una solución.

Tabla 2. Matriz de confusión.

	Clasificado Positivo	**Clasificado Negativo**
Positivo	VP	FN
Negativo	FP	VN

los tres métodos reciben el porcentaje de radio de balanceo ($\%\rho_U$) que se traduce en el número de instancias por disminuir. También se usa el parámetro k_U como el número de vecinos a tomar en cuenta para el método NearMiss.

Finalmente, el parámetro *Orden* toma cuatro valores posibles para indicar el orden de aplicación de los métodos de la siguiente forma: 1 = se aplica oversampling y 2 = se aplica undersampling. Para la aproximación de hybridsampling, se hace notar que el orden en que se aplican las técnicas de oversampling y undersampling pueden tener efectos distintos debido a las diferentes técnicas que se eligieron.

Un ejemplo de esto lo podemos notar si seleccionamos Borderline-SMOTE como técnica de oversampling, puesto que los límites encontrados entre la clase minoritaria y la clase mayoritaria original, pueden ser distintos a los límites entre la clase minoritaria y la clase mayoritaria luego de reducir el número de puntos al aplicar una técnica de undersampling.

Dado esto, se propone tomar los valores para hybridsampling de la siguiente forma: 3 = se aplica primero oversampling y después undersampling, 4 = se aplica primero undersampling y después oversampling. Con esto abordamos las tres aproximaciones posibles de los métodos de balanceo basados en muestreo (oversampling, undersampling, hybridsampling).

2.3. Función de desempeño

Debido a que, como todo algoritmo evolutivo, el algoritmo de Evolución Diferencial requiere una función para evaluar el desempeño de las soluciones elegidas, se hace uso de una medida para evaluar el desempeño de clasificación cuando se realiza el balanceo. Para poder realizar la evaluación de una solución,

se realizan distintos procedimientos. El primer paso es dividir la base de datos en dos conjuntos, un conjunto de entrenamiento y un conjunto de prueba, ambos conjuntos deben contener de forma proporcional ambas clases (minoritaria y mayoritaria).

Tabla 3. Bases de datos seleccionadas.

Nombre	#Atributos	#Instancias	#Ins. Minoritaria	#Ins. Mayoritaria	IR
glass1	9	214	76	138	1.82
pima	8	768	268	500	1.87
iris0	4	150	50	100	2
haberman	3	306	81	225	2.78
vehicle1	18	846	217	629	2.9
ecoli2	7	336	52	284	5.46
yeast0359vs78	8	506	50	456	9.12
ecoli0146vs5	6	280	20	260	13
glass4	9	214	13	201	15.47
yeast1458vs7	8	693	30	663	22.1

El siguiente paso es aplicar el método de balanceo seleccionado al conjunto de entrenamiento, ya sea la aplicación de undersampling, oversampling o ambas aproximaciones en los distintos ordenes disponibles con los respectivos hiperparámetros elegidos. Una vez que se balancea el conjunto de entrenamiento, se usa para entrenar un modelo de clasificación.Por último, se evalúa el desempeño del clasificador introduciendo el conjunto de prueba, al cual no se le aplicó ningún método de balanceo.

Este procedimiento se puede observar en la Figura 1. Una vez obtenidas las clases asignadas por el clasificador, se obtiene la matriz de confusión como la que se muestra en la Tabla 2 donde se incluye el número de elementos que tenían etiqueta positiva y el clasificador les asignó la etiqueta positiva (VP), el número de elementos que tenían etiqueta positiva, pero el clasificador les asignó la etiqueta negativa (FN), el número de elementos que tenían etiqueta negativa, pero el clasificador les asignó la etiqueta positiva (FN) y por último, el número de elementos que tenían etiqueta negativa y que el clasificador les asignó la etiqueta negativa (VN).

La construcción de esta matriz de confusión permite evaluar mediante distintas medidas de desempeño, en nuestro caso se hará uso de la G-Media (G-Mean), la cual combina medidas de sensitividad (1) y especificidad (2), por lo que existe un balance entre un buen desempeño de clasificación medido tanto para la clase mayoritaria como para la clase minoritaria.

Esta medida indica un buen desempeño en la clasificación mientras más grande sea, en cambio, mientras más pequeño sea el valor de la G-Mediana, el clasificador se desempeña de peor forma en la clasificación de ambas clases.

Esta medida es recomendada para evaluar el desempeño de clasificación en bases de datos desbalanceadas [1,17]:

$$\text{sensitividad} = \frac{VP}{VP + FN}, \tag{1}$$

$$\text{especificidad} = \frac{VN}{VN + FP}, \tag{2}$$

$$\text{G-Media} = \sqrt{\text{sensitividad} \times \text{especificidad}}. \tag{3}$$

3. Experimentos y resultados

En esta sección, se muestran los experimentos realizados con la propuesta presentada anteriormente, así como las configuraciones asignadas y las bases de datos usadas. Las bases de datos seleccionadas para los experimentos corresponden a la sección de bases de datos desbalanceadas pertenecientes al *KEEL-dataset repository* [2]. En la Tabla 3 se pueden apreciar las bases de datos que se seleccionaron, así como sus correspondientes números de instancias pertenecientes a las clases mayoritaria y minoritaria.

Como se puede notar, para realizar los experimentos hemos elegido un número variable de radio de balanceo (IR), por lo que podemos ver el desempeño de nuestra propuesta en diferentes grados de desbalanceo. Para realizar la selección de los valores de los parámetros necesarios para el algoritmo de Evolución Diferencial, se hizo uso del lenguaje de programación R donde se encuentra la paquetería *irace* [11] para la configuración automática del algoritmo.

Luego de la ejecución de *irace*, se obtuvieron los siguientes parámetros: Número de población = 44, Número de generaciones = 139, F = 1.46, CR = 0.06. Con estos parámetros mencionados, se realizaron 10 ejecuciones independientes. Todas estas pruebas fueron realizadas haciendo uso de Python y Jupyter Notebooks en una computadora portátil que cuenta con un procesador Intel Core i7-8750H con 16GB de RAM.

Para la comparación de resultados, se decidieron aplicar los métodos de balanceo individualmente con sus parámetros por defecto, así como agregar dos funciones que cuentan con la aproximación de hybridsampling (SMOTEENN y SMOTETomek), de igual forma, con sus parámetros por defecto. Los dos algoritmos de clasificación que se seleccionaron fueron Árbol de decisión (profundidad = 5) y Naive Bayes (parámetros por defecto). Con el fin de analizar el desempeño de la Evolución Diferencial como método de búsqueda, se hace uso de tres diferentes estrategias: DE/best/1/bin, DE/best/1/exp/ y DE/rand/1/exp. Esto con el fin de analizar los resultados obtenidos por la aplicación de la propuesta con distintas estrategias de búsqueda.

Los resultados obtenidos se presentan en las Tablas 4 y 5. Estos resultados se analizaron con una prueba de Shapiro-Wilk para verificar si provienen de una población que se distribuye de una forma normal, posteriormente se hace uso de la prueba Kruskal-Wallis para analizar si los individuos provienen de

la misma población y confirmar diferencias significativas. Con los resultados obtenidos de las pruebas anteriores, se procedió a realizar la prueba posthoc de Dunn-Sidak. Los resultados se expresan como la media de la G-Media más la diferencia a los cuartiles 25 y 75.

4. Discusión

Como podemos observar en la Tabla 4, existen diferencias de desempeño entre la población y se distinguen dos grupos, un grupo (a) formado por todos los métodos con los parámetros por defecto, incluidas las aproximaciones de hybridsampling y el grupo (b) donde se encuentra la propuesta que realizamos nosotros con las distintas estrategias de Evolución Diferencial.

Dados estos dos grupos, podemos darnos cuenta que existe una diferencia significativa entre el desempeño del grupo (a) y el desempeño del grupo (b). En esta tabla podemos observar que en todas las bases de datos, nuestra propuesta obtiene un mejor desempeño en términos de la G-Media.

En la Tabla 5 podemos observar que sucede lo mismo que en la Tabla 4, a excepción de los resultados obtenidos con la base de datos *ecoli0146vs5* donde se crean tres grupos, el grupo (a) que se desempeña peor que el grupo (b) y el grupo (ab) que se desempeña de forma similar a los dos grupos antes mencionados. En el resto de bases de datos, el grupo (b) que corresponde a nuestra propuesta, es superior en desempeño.

5. Conclusiones y trabajo futuro

Ya que no existe un criterio específico para la selección de un método de balanceo o sus hiperparámetros, es necesario encontrar las opciones que mejor se adapten a una base de datos dada, ya que no todas las bases de datos cumplen con las mismas condiciones. En este trabajo se presentó una propuesta que es capaz de seleccionar el mejor método de balanceo y sus hiperparámetros mediante Evolución Diferencial para una base de datos en partícular, de tal forma que se mejora el desempeño de una medida, en este caso se selecciona un método de balanceo que optimiza la G-Media, lo cual se traduce a un buen desempeño de clasificación tanto en la clase mayoritaria como en la minoritaria.

Aunque se aborda el problema de desbalanceo, sigue siendo un tema extenso, por lo que quedan temas para abordar en el futuro como el estudio en más bases de datos, el estudio de otras medidas distintas a la G-Media, el estudio en bases de datos con más de dos clases y el uso de otros algoritmos de optimización.

Agradecimientos. El primer autor agradece el apoyo de CONACyT mediante una beca para la realización de estudios de posgrado en el Instituto de Investigaciones en Inteligencia Artificial de la Universidad Veracruzana.

Tabla 4. Valores de la G-Media obtenidos haciendo uso del Árbol de decisión como clasificador. Los resultados obtenidos se expresan como el promedio del resultado de 10 evaluaciones. Los valores entre paréntesis representan la diferencia de la media al cuartil 25 de las 10 evaluaciones (lado izquierdo) y la diferencia de la media al cuartil 75 (lado derecho). En las últimas dos columnas se muestran las estadísticas relacionadas a la prueba Kruskal-Wallis. Las etiquetas corresponden a los grupos donde se observan diferencias siguiendo la prueba de Dunn-Sidak.

Nombre	RO	SMOTE	B-SMOTE	RU	NM	CC	SMOTEENN	SMOTETomek	DE/best/1/bin	DE/best/1/exp	DE/rand/1/bin	Estadísticas	
												H	p-value
glass1	0.7338^{a} (-0.0202,+0.0076)	0.7105^{a} (-0.022,+0.0114)	0.6742^{a} (-0.055,+0.0523)	0.7285^{a} (-0.0137,+0.0169)	0.5533^{a} (-0.0149,+0.0143)	0.6431^{a} (-0.0442,+0.0475)	0.7056^{a} (-0.0146,+0.0148)	0.6954^{a} (-0.0351,+0.0434)	0.9247^{b} (-0.0067,+0.0072)	0.9262^{b} (-0.0083,+0.0056)	$\mathbf{0.9274}^{b}$ (-0.0087,+0.005)	86.1079	3.155e-14
pima	0.7267^{a} (-0.0169,+0.0127)	0.7348^{a} (-0.0277,+0.0274)	0.7366^{a} (-0.0137,+0.019)	0.6977^{a} (-0.0113,+0.0032)	0.7117^{a} (-0.0045,+0.0019)	0.6873^{a} (-0.01,+0.014)	0.7311^{a} (-0.0064,+0.0101)	0.7239^{a} (-0.0259,+0.0253)	$\mathbf{0.8281}^{b}$ (-0.0054,+0.0069)	0.8274^{b} (-0.0053,+0.0044)	0.8275^{b} (-0.0046,+0.005)	77.9891	1.2426e-12
iris0	1.0(-0.0,+0.0)	1.0(-0.0,+0.0)	1.0(-0.0,+0.0)	1.0(-0.0,+0.0)	1.0(-0.0,+0.0)	1.0(-0.0,+0.0)	1.0(-0.0,+0.0)	1.0(-0.0,+0.0)	1.0(-0.0,+0.0)	1.0(-0.0,+0.0)	1.0(-0.0,+0.0)	NA	NA
haberman	0.5595^{a} (-0.0421,+0.0296)	0.5787^{a} (-0.0148,+0.0133)	0.5954^{a} (-0.0326,+0.0278)	$0.5943)^{a}$ (-0.0106,+0.0045)	0.5599^{a} (-0.0271,+0.0253)	0.573^{a} (-0.0168,+0.0038)	0.5592^{a} (-0.0234,+0.0185)	0.5668^{a} (-0.0195,+0.0183)	$\mathbf{0.8077}^{b}$ (-0.0062,+0.0057)	0.8034^{b} (-0.0105,+0.0092)	0.8018^{b} (-0.0031,+0.0036)	72.5177	1.4441e-11
vehicle1	0.7461^{a} (-0.0158,+0.0139)	0.75^{a} (-0.0128,+0.0113)	0.7432^{a} (-0.0221,+0.0148)	0.7292^{a} (-0.0125,+0.0172)	0.6738^{a} (-0.0233,+0.0245)	0.6815^{a} (-0.0229,+0.0226)	0.7345^{a} (-0.0042,+0.0079)	0.7481^{a} (-0.0076,+0.0083)	$\mathbf{0.8496}^{b}$ (-0.0073,+0.0061)	0.848^{b} (-0.0076,+0.0009)	0.8435^{b} (-0.0039,+0.0028)	88.3314	1.1465e-14
ecoli2	0.916^{a} (+0.0028,+0.0125)	0.9271^{a} (-0.0077,+0.0028)	0.8377^{a} (-0.0233,+0.0233)	0.886^{a} (-0.0238,+0.0187)	0.6964^{a} (-0.0266,+0.0286)	0.847^{a} (-0.0134,+0.0287)	0.9324^{a} (-0.0122,+0.0113)	0.931^{a} (-0.0109,+0.0142)	$\mathbf{1.0}^{b}$ (-0,+0)	$\mathbf{1.0}^{b}$ (-0,+0)	0.9988^{b} (+0.0011,+0.0011)	99.2312	7.7655e-17
yeast0359vs78	0.6268^{a} (-0.055,+0.04)	0.6636^{a} (-0.0238,+0.0185)	0.6446^{a} (-0.0092,+0.0006)	0.6517^{a} (-0.035,+0.0224)	0.7083^{a} (-0.0072,+0.0048)	0.5359^{a} (-0.0301,+0.0505)	0.7056^{a} (-0.0115,+0.0196)	0.6697^{a} (-0.0414,+0.0383)	$\mathbf{0.8919}^{b}$ (-0.0096,+0.0083)	0.8883^{b} (-0.0106,+0.0044)	0.8836^{b} (-0.006,+0.0054)	87.6765	1.5452e-14
ecoli0146vs5	0.8155^{a} (-0.0095,-0.0095)	0.9511^{a} (-0.0334,+0.0295)	0.9539^{a} (-0.0514,+0.0332)	0.8082^{a} (-0.0292,+0.0457)	0.7679^{a} (-0.0,+0.0)	0.8359^{a} (-0.0038,-0.0038)	0.9409^{a} (-0.0443,+0.0445)	0.9105^{a} (-0.0213,+0.0094)	$\mathbf{1.0}^{b}$ (-0,+0)	$\mathbf{1.0}^{b}$ (-0,+0)	$\mathbf{1.0}^{b}$ (-0,+0)	98.9884	8.6841e-17
glass4	0.6989^{a} (-0.0035,+0.0024)	0.6983^{a} (-0.0029,+0.0029)	0.6995^{a} (-0.0026,+0.0018)	0.761^{a} (-0.0775,+0.0576)	0.866^{a} (-0.0,+0.0)	0.5901^{a} (+0.0122,+0.0372)	0.6942^{a} (-0.0047,+0.0012)	0.696^{a} (-0.0006,+0.0038)	$\mathbf{1.0}^{b}$ (-0,+0)	$\mathbf{1.0}^{b}$ (-0,+0)	$\mathbf{1.0}^{b}$ (-0,+0)	92.5082	1.7014e-15
yeast1458vs7	0.4075^{a} (-0.0887,+0.1354)	0.5225^{a} (-0.0313,+0.0238)	0.6068^{a} (-0.0658,+0.0353)	0.4726^{a} (-0.0816,+0.0973)	0.5353^{a} (-0.0437,+0.0396)	0.5255^{a} (-0.0181,+0.0105)	0.5471^{a} (-0.0388,+0.0451)	0.5823^{a} (-0.0373,+0.0141)	$\mathbf{0.8862}^{b}$ (-0.0109,+0.0119)	0.8723^{b} (-0.0132,+0.0099)	0.8751^{b} (-0.0115,+0.0059)	75.23	4.2916e-12

Tabla 5. Valores de la G-Media obtenidos haciendo uso del algoritmo Naive Bayes como clasificador. Los resultados obtenidos se expresan como el promedio del resultado de 10 evaluaciones. Los valores entre paréntesis representan la diferencia de la media al cuartil 25 de las 10 evaluaciones (lado izquierdo) y la diferencia de la media al cuartil 75 (lado derecho). En las últimas dos columnas se muestran las estadísticas relacionadas a la prueba Kruskal-Wallis. Las etiquetas corresponden a los grupos donde se observan diferencias siguiendo la prueba de Dunn-Sidak.

Nombre	RO	SMOTE	B-SMOTE	RU	NM	CC	SMOTEENN	SMOTETomek	DE/best/1/bin	DE/best/1/exp	DE/rand/1/bin	Estadísticas	
												H	p-value
glass1	0.5941^{a} (-0.0268,+0.0318)	0.6113^{a} (-0.0528,+0.054)	0.6129^{a} (-0.0192,+0.0343)	0.6001^{a} (-0.0425,+0.0365)	0.6132^{a} (-0.0037,+0.0422)	0.6129^{a} (-0.0522,+0.0091)	0.6165^{a} (-0.0222,+0.0405)	0.6397^{a} (-0.046,+0.0512)	$\mathbf{0.8402}^{b}$ (-0.0097,+0.0050)	0.8392^{b} (-0.0101,+0.0122)	0.8272^{b} (-0.0063,+0.0049)	67.5398	1.3205e-10
pima	0.6981^{a} (-0.0132,+0.0157)	0.7111^{a} (-0.0169,+0.0144)	0.6833^{a} (-0.0161,+0.0209)	0.7131^{a} (-0.0225,+0.0223)	0.6918^{a} (-0.0243,+0.0206)	0.697^{a} (-0.0287,+0.0216)	0.73^{a} (-0.025,+0.0309)	0.7082^{a} (-0.0077,+0.0123)	0.8260^{b} (-0.0081,+0.0070)	0.8270^{b} (-0.0036,+0.0026)	$\mathbf{0.8282}^{b}$ (-0.004,+0.0021)	70.1248	4.1943e-11
iris0	1.0(-0.0,+0.0)	1.0(-0.0,+0.0)	1.0(-0.0,+0.0)	1.0(-0.0,+0.0)	1.0(-0.0,+0.0)	1.0(-0.0,+0.0)	1.0(-0.0,+0.0)	1.0(-0.0,+0.0)	1.0(-0.0,+0.0)	1.0(-0.0,+0.0)	1.0(-0.0,+0.0)	NA	NA
haberman	0.4388^{a} (-0.0443,+0.051)	0.3779^{a} (-0.0802,+0.0684)	0.4616^{a} (-0.0188,+0.0272)	0.4186^{a} (-0.0564,+0.0519)	0.432^{a} (-0.036,+0.0393)	0.3762^{a} (-0.0776,+0.0572)	0.404^{a} (-0.0407,+0.0607)	0.4494^{a} (-0.0379,+0.0639)	0.7083^{b} (-0.0029,+0.0030)	0.7016^{b} (-0.0045,+0.0053)	$\mathbf{0.7092}^{b}$ (-0.0121,+0.0014)	69.7754	4.8991e-11
vehicle1	0.6844^{a} (-0.048,+0.0428)	0.687^{a} (-0.0224,+0.0178)	0.6906^{a} (-0.0236,+0.015)	0.6847^{a} (-0.0047,+0.0092)	0.6848^{a} (-0.0148,+0.0193)	0.6738^{a} (-0.0124,+0.0097)	0.6601^{a} (-0.0284,+0.043)	0.6867^{a} (-0.0074,+0.0251)	0.7914^{b} (-0.0107,+0.0079)	0.8118^{b} (-0.0029,+0.0048)	$\mathbf{0.8129}^{b}$ (-0.0061,+0.0044)	67.8260	1.1633e-10
ecoli2	0.5163^{a} (-0.1415,+0.1504)	0.4756^{a} (-0.195,+0.1915)	0.4858^{a} (-0.1346,+0.1557)	0.5311^{a} (-0.1117,+0.1577)	0.4583^{a} (-0.1153,+0.1089)	0.5705^{a} (-0.1238,+0.1102)	0.4076^{a} (-0.0646,+0.0014)	0.4894^{a} (-0.1219,+0.0833)	0.8737^{b} (-0.0039,+0.0049)	$\mathbf{0.8838}^{b}$ (-0.0164,-0.0066)	0.8779^{b} (-0.0051,+0.0006)	67.9555	1.0984e-10
yeast0359vs78	0.351^{a} (-0.0447,+0.0518)	0.342^{a} (-0.0373,+0.0179)	0.3403^{a} (-0.0384,+0.0389)	0.3135^{a} (-0.0368,+0.0362)	0.3018^{a} (-0.0307,+0.04)	0.3518^{a} (-0.042,+0.0605)	0.3338^{a} (-0.0141,+0.0262)	0.3251^{a} (-0.0536,+0.0245)	0.6989^{b} (-0.0233,+0.0233)	0.7129^{b} (+0.0093,+0.0093)	$\mathbf{0.7328}^{b}$ (-0.0106,-0.0025)	69.1378	6.5030e-11
ecoli0146vs5	0.8974^{a} (-0.0627,+0.0745)	0.8257^{ab} (-0.0197,+0.0664)	0.847^{a} (-0.0358,+0.0266)	0.8504^{ab} (-0.0378,+0.0551)	0.824^{a} (-0.0638,+0.0463)	0.8362^{ab} (-0.038,+0.0392)	0.7926^{a} (-0.0345,+0.0434)	0.8138^{a} (-0.0806,+0.0591)	$\mathbf{1.0}^{b}$ (-0.0,+0.0)	$\mathbf{1.0}^{b}$ (-0.0,+0.0)	$\mathbf{1.0}^{b}$ (-0.0,+0.0)	68.3224	9.3368e-11
glass4	0.4834^{a} (-0.3605,+0.3083)	0.3393^{a} (-0.3393,+0.2589)	0.3069^{a} (-0.3069,+0.1848)	0.5123^{a} (-0.028,+-0.0123)	0.4919^{a} (-0.37,+0.3153)	0.3434^{a} (-0.3434,+0.234)	0.3762^{a} (-0.2597,+0.1165)	0.5434^{a} (-0.0831,+0.1458)	$\mathbf{0.9859}^{b}$ (-0.0024,+0.0037)	0.9851^{b} (-0.0016,-0.0016)	0.9851^{b} (-0.0016,-0.0016)	70.2971	3.8848e-11
yeast1458vs7	0.292^{a} (-0.0392,+0.0309)	0.2753^{a} (-0.0249,+0.0169)	0.3039^{a} (-0.0204,+0.0111)	0.3249^{a} (-0.0242,+0.0187)	0.2914^{a} (-0.0079,+0.0194)	0.293^{a} (-0.0254,+0.0078)	0.294^{a} (-0.0172,+0.021)	0.2968^{a} (-0.0292,+0.0081)	0.5051^{b} (-0.0039,+0.0031)	$\mathbf{0.5082}^{b}$ (-0.0069,+0.0029)	0.5055^{b} (-0.0042,+0.0044)	69.5501	5.4149e-11

Referencias

1. Akosa, J. S.: Predictive accuracy: A misleading performance measure for highly imbalanced data (2017)
2. Alcalá-Fdez, J., Fernández, A., Luengo, J., Derrac, J., García, S.: KEEL data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework. J. Multiple Valued Log. Soft Comput., vol. 17, no. 2-3, pp. 255–287 (2011)
3. Cao, L., Zhai, Y.: Imbalanced data classification based on a hybrid resampling SVM method, pp. 1533–1536 (2015) doi: 10.1109/UIC-ATC-ScalCom-CBDCom-IoP.2015.275
4. Gao, Z., Zhang, L. F., Chen, M. Y., Hauptmann, A., Zhang, H., Cai, A. N.: Enhanced and hierarchical structure algorithm for data imbalance problem in semantic extraction under massive video dataset. Multimedia Tools Appl., vol. 68, no. 3, pp. 641–657 (2014) doi: 10.1007/s11042-012-1071-7
5. Hodge, V., Austin, J.: A survey of outlier detection methodologies. Artificial Intelligence Review, vol. 22, pp. 85–126 (2004) doi: 10.1023/B:AIRE.0000045502.10941.a9
6. Kaur, H., Pannu, H. S., Malhi, A. K.: A systematic review on imbalanced data challenges in machine learning: Applications and solutions. ACM Comput. Surv., vol. 52, no. 4 (2019) doi: 10.1145/3343440
7. Kaya, E., Korkmaz, S., Sahman, M. A., Cinar, A. C.: DEBOHID: A differential evolution based oversampling approach for highly imbalanced datasets. Expert Systems with Applications, vol. 169, pp. 114482 (2021) doi: 10.1016/j.eswa.2020.114482
8. Korkmaz, S., Şahman, M. A., Cinar, A. C., Kaya, E.: Boosting the oversampling methods based on differential evolution strategies for imbalanced learning. Applied Soft Computing, vol. 112, pp. 107787 (2021) doi: 10.1016/j.asoc.2021.107787
9. Kubat, M., Holte, R., Matwin, S.: A survey of outlier detection methodologies. Machine Learning, vol. 30, pp. 195–215 (1998) doi: 10.1023/A:1007452223027
10. Li, J., Fong, S., Wong, R. K., Chu, V. W.: Adaptive multi-objective swarm fusion for imbalanced data classification. Information Fusion, vol. 39, pp. 1–24 (2018) doi: 10.1016/j.inffus.2017.03.007
11. López-Ibáñez, M., Dubois-Lacoste, J., Cáceres, L. P., Stützle, T., Birattari, M.: The irace package: Iterated racing for automatic algorithm configuration. Operations Research Perspectives, vol. 3, pp. 43–58 (2016) doi: 10.1016/j.orp.2016.09.002
12. Miranda, P. B., Morais, R. F., Silva, R. M.: Using a many-objective optimization algorithm to select sampling approaches for imbalanced datasets. In: 2018 IEEE Congress on Evolutionary Computation (CEC). pp. 1–7 (2018) doi: 10.1109/CEC.2018.8477988
13. Nepal, S., Pathan, M.: Security, privacy and trust in cloud systems. Springer Publishing Company, Incorporated (2013)
14. Ramentol, E., Gondres, I., Lajes, S., Bello, R., Caballero, Y., Cornelis, C., Herrera, F.: Fuzzy-rough imbalanced learning for the diagnosis of high voltage circuit breaker maintenance: The SMOTE-FRST-2T algorithm. Engineering Applications of Artificial Intelligence, vol. 48, pp. 134–139 (2016) doi: 10.1016/j.engappai.2015.10.009
15. Sahin, Y., Bulkan, S., Duman, E.: A cost-sensitive decision tree approach for fraud detection. Expert Systems with Applications, vol. 40, no. 15, pp. 5916–5923 (2013) doi: 10.1016/j.eswa.2013.05.021

16. Storn, R., Price, K.: Differential evolution: A simple and efficient adaptive scheme for global optimization over continuous spaces. Journal of Global Optimization, vol. 23 (1995)
17. Susan, S., Amitesh: The balancing trick: Optimized sampling of imbalanced datasets—a brief survey of the recent state of the art. Engineering Reports, vol. 3 (2021) doi: 10.1002/eng2.12298
18. Telikani, A., Tahmassebi, A., Banzhaf, W., Gandomi, A. H.: Evolutionary machine learning: A survey. ACM Comput. Surv., vol. 54, no. 8 (2021) doi: 10.1145/3467477
19. Upadhyay, K., Kaur, P., Prasad, S.: State of the art on data level methods to address class imbalance problem in binary classification, vol. 8, pp. 975–903 (2021)
20. Wei, W., Li, J., Cao, L., Ou, Y., Chen, J.: Effective detection of sophisticated online banking fraud on extremely imbalanced data. World Wide Web, vol. 16 (2013) doi: 10.1007/s11280-012-0178-0
21. Woods, K. S., Doss, C. C., Bowyer, K. W., Solka, J. L., Priebe, C. E., Kegelmeyer, W. P.: Comparative evaluation of pattern recognition techniques for detection of microcalcifications in mammography. International Journal of Pattern Recognition and Artificial Intelligence, vol. 7, no. 6, pp. 1417–1436 (1993) doi: 10.1142/S0218001493000698
22. Zhu, Y., Yan, Y., Zhang, Y., Zhang, Y.: EHSO: Evolutionary hybrid sampling in overlapping scenarios for imbalanced learning. Neurocomputing, vol. 417, pp. 333–346 (2020) doi: 10.1016/j.neucom.2020.08.060

ISSN 1870-4069

Edadismo e inteligencia artificial

José Martín Castro-Manzano

Universidad Popular Autónoma del Estado de Puebla,
Facultad de Filosofía,
México

josemartin.castro@upaep.mx

Resumen. Dentro de las investigaciones sobre ética y filosofía de la inteligencia artificial podemos encontrar estudios sobre discriminación que se han centrado en el problema de los sesgos raciales y de género; sin embargo, poca atención se ha prestado a un sesgo igual o más importante que los anteriores: el sesgo relacionado con la edad. Así, debido a que este sesgo ha sido escasamente explorado en el contexto de la inteligencia artificial, en este trabajo pretendemos alcanzar dos metas: primero, exponer el problema de la discriminación por edad con especial énfasis en la situación de las personas adultas mayores; y segundo, argumentar que la inteligencia artificial no está libre de este sesgo.

Palabras clave: Discriminación, edad, ética de la inteligencia artificial.

Ageism and Artificial Intelligence

Abstract. Within the ethics and philosophy of artificial intelligence we can find studies on discrimination that have focused on the problem of racial and gender biases; however, little attention has been paid to a bias that is equally or more important than the above: age-related bias. Thus, because this bias has been scarcely explored in the context of artificial intelligence, in this work we intend to achieve two goals: first, to expose the problem of age discrimination with special emphasis on the situation of older adults; and second, to argue that artificial intelligence is not free from this bias.

Keywords: Discrimination, age, ethics of artificial intelligence.

1. Introducción

Dentro de las investigaciones sobre ética y filosofía de la inteligencia artificial podemos encontrar estudios sobre discriminación que se han centrado en el problema de los sesgos raciales y de género [24]; sin embargo, poca atención se ha prestado a un sesgo igual o más importante que los anteriores: el sesgo relacionado con la edad [7]. Este sesgo se conoce como discriminación negativa por edad o edadismo y es un fenómeno social multifacético que la Organización Mundial de la Salud ha definido como el conjunto de estereotipos

(lo que pensamos), prejuicios (lo que sentimos) y acciones (lo que hacemos) dirigidos hacia las personas en función de su edad. Bajo esta definición, esta forma de discriminación tiene tres dimensiones (pensamientos, sentimientos y comportamientos), tres manifestaciones (personal, interpersonal e institucional) y dos formas de expresión (consciente e inconsciente) [23].

Ahora bien, debido a que este sesgo ha sido escasamente explorado en el contexto de la inteligencia artificial, en este trabajo pretendemos alcanzar dos metas: primero, exponer el problema de la discriminación por edad con especial énfasis en la situación de las personas adultas mayores; y segundo, argumentar que la inteligencia artificial no está libre de este sesgo.

Para alcanzar estas metas hemos diseñado nuestra contribución de la siguiente manera: primero exponemos algunos antecedentes relevantes para enmarcar el problema, después argumentamos que la inteligencia artificial no está libre de este sesgo y, al final, cerramos con una breve discusión.

2. Antecedentes

Hoy parece una perogrullada afirmar que ciertas aplicaciones de inteligencia artificial (IA, en adelante) no están libres de prejuicios raciales o de género [6,14], pero aún así vale la pena recordar algunos ejemplos paradigmáticos a manera de antecedente. Por ejemplo, se ha mostrado que ciertos algoritmos subestiman los riesgos para la salud de personas negras en comparación con personas blancas [17].

El problema de estos algoritmos es que toman en cuenta los costos de atención médica de las personas, pero no consideran la causa principal del menor gasto en atención médica de las personas negras, a saber, el acceso reducido a la atención médica debido al racismo sistémico. Otros casos de sesgo racial incluyen sistemas de IA que asignan sentencias de cárcel más largas a personas negras [1]. En cuanto al género, también se han identificado sesgos contra las mujeres, como la menor probabilidad de que reciban anuncios de búsqueda de empleo para puestos bien remunerados [9] y discriminación laboral [10].

Este sesgo se puede atribuir a que los algoritmos aprenden no solo de datos cuantitativos sino también de textos que codifican asociaciones semánticas histórico-culturales, como asociaciones entre nombres masculinos y el concepto de trabajo, y por el contrario, nombres femeninos y el concepto de familia [5]. Ciertamente, alguien podría pensar que casos como estos son aislados, pero esta sería una opinión acrítica. En primer lugar, estos sesgos han permitido una discriminación negativa de facto a ciertas personas de manera injustificada y sistemática porque se les han negado ciertos derechos o bienes por razones irrelevantes [11]; y en segundo lugar, y todavía más importante, estos sesgos podrían seguir causando daños innecesarios, injustificados e indeseables, como los daños de asignación y los daños de representación [16].

Los daños de asignación se refieren a la distribución de derechos y oportunidades (como cuándo ser liberada bajo fianza o recibir notificaciones sobre posibles perspectivas laborales); los daños de representación se refieren a cómo la sociedad representa y percibe diferentes grupos o identidades [8]. Haremos

referencia a estos antecedentes más adelante, pero lo que debería quedar claro en este punto es que ciertas aplicaciones de IA no están libres de sesgos que favorecen la discriminación negativa y, como el edadismo no ha sido revisado con el mismo interés que otras formas de discriminación, a continuación exponemos el problema del edadismo con especial énfasis en la situación de las personas adultas mayores.

3. Edadismo e inteligencia artificial

En enero de 1969 la agencia de vivienda pública del Distrito de Columbia, la National Capital Housing Authority, celebró audiencias sobre su propuesta de compra de Regency House, un edificio de departamentos en Chevy Chase, para personas ancianas en estado de pobreza. La ciudadanía blanca de clase media y mediana edad de Chevy Chase compareció en las audiencias y protestó, por diversos motivos, contra la propuesta.

Entre las protestas estaban, además de las preocupaciones económicas, las siguientes: "Le abrirías la puerta a gente que no sabe cómo vivir", "La vivienda pública tiene que llegar en algún momento, pero no en este momento ni en este lugar", o el típico "No estoy en contra de los viejos, créanme", pero "¿Quién quiere a todos esos viejos alrededor?"

La anécdota descrita previamente le permitió a Robert N. Butler exponer un problema que, desde aquel entonces, necesitaba más atención. Así, en 1969 salió a la luz su Age-ism: Another form of bigotry en The Gerontologist [4]. Allí, Butler ofreció una definición del edadismo como el prejuicio de un grupo de edad contra otro grupo de edad y un argumento básico que podemos especificar de la siguiente manera:

1. El racismo y el clasismo no tienen justificación moral.
2. El edadismo es relevantemente similar al racismo y al clasismo.
3. Luego, el edadismo no tiene justificación moral.

La verdad de la primera premisa ya era fácil de justificar en aquel entonces y es más fácil de verificar ahora. El racismo y el clasismo no tienen justificación moral porque son formas de discriminación ilegítima, esto es, son maneras de particionar la estructura de las relaciones sociales con base en jerarquías accesorias o accidentales que, por tanto, no tienen un vínculo relevante con la justicia.

La verdad de la segunda premisa, sin embargo, requiere más explicación. Lo que hizo Butler en aquel trabajo fue mostrar que el edadismo también particiona la estructura de las relaciones sociales de manera ilegítima, como lo hacen el racismo y el clasismo, pero para ilustrar este punto con más claridad y, al mismo tiempo, mostrar que el edadismo es un problema real, consideremos algunas de sus causas y revisemos si, como justificaciones, son suficientes y legítimas: ya podemos adelantar que no lo son.

Pues bien, para mostrar que el edadismo es un problema real podemos investigar sus causas y efectos a través de tres niveles de teorías que corresponden

a las manifestaciones personal, interpersonal e institucional: tenemos, entonces, las teorías micronivel, mesonivel y macronivel.

Las primeras enfatizan las causas del edadismo en las personas; las segundas, en las relaciones personales; y las terceras, en las instituciones y la cultura. Para explicitar estas teorías, a continuación reproducimos la exposición de [2]. Entre las teorías micronivel podemos encontrar a la teoría de la gestión del terror, la teoría de la identidad social y la teoría del contenido estereotípico.

La primera sugiere que, dado que las personas adultas mayores son como un recordatorio constante de nuestra mortalidad y vulnerabilidad, preferimos sostener discursos y visiones del mundo que valoran la juventud como mecanismo de defensa ante la ansiedad que produce la presencia de personas adultas mayores [12].

La teoría de la identidad social sostiene que las personas actúan motu proprio pero también como miembros de sus grupos de referencia. La pertenencia a un grupo es la base de la identidad individual y determina las relaciones con personas de otros grupos. Esta teoría postula, así, que las personas buscan una identidad positiva aplicando sesgos que crean distinciones entre su grupo y otros, y como la edad puede ser un criterio para la identificación social, esta teoría explica la discriminación por edad [21]. La teoría del contenido estereotípico sugiere que los grupos de personas se clasifican comúnmente por diferentes niveles de calidez y competencia.

Las personas adultas mayores, por ejemplo, son comúnmente percibidas como cálidas pero incompetentes. Estas percepciones producen sentimientos de compasión y simpatía, y menos sentimientos de envidia. Según esta teoría, la discriminación por edad tiene su origen en la infancia y se desarrolla a lo largo de la vida. Por ejemplo, desde la infancia podemos percibir a las personas adultas mayores negativamente con respecto a las dimensiones de actividad y potencia, y positivamente con respecto a la bondad social [22].

En suma, de acuerdo con estas teorías micronivel, las causas del edadismo son personales, por lo que alguien podría justificar que tiene un sesgo contra las personas adultas mayores pero que no es edadista apelando a alguna de las siguientes razones: a que la presencia de personas adultas mayores produce ansiedad, a que buscamos una mejor identidad positiva, o a que desde la infancia hemos desarrollado un prejuicio contra las personas adultas mayores; sin embargo, ninguna de estas apelaciones podría contar como una justificación suficiente.

Basta sustituir, en cada apelación, al rasgo de la edad por un rasgo alternativo como la raza, la clase social o el sexo, por ejemplo. Sin embargo, la discriminación por edad no siempre comienza a nivel individual, algunas de sus causas se pueden rastrear en las relaciones interpersonales.

En este grupo de teorías mesonivel encontramos, por ejemplo, a la teoría evolutiva, la teoría de la segregación y la teoría de la amenaza intergrupal. La primera sostiene que las personas están programadas filogenéticamente para ser parte de un grupo y aprenden que su propio bienestar es interdependiente del de otros miembros del mismo grupo.

En esta teoría, la edad, la riqueza, la reputación y la salud juegan un papel fundamental en la determinación de si se brindará asistencia o no, porque es más probable que se ayude a las personas que se perciben con un mayor potencial reproductivo, y cuando la vida está en peligro, es más probable que las personas ayuden a sus familiares y a aquellas que son más jóvenes [3].

La teoría de la segregación por edad afirma que en la mayoría de las sociedades occidentales modernas existe una clara segregación entre personas jóvenes y ancianas basada en guiones de vida (life scripts) planificados previamente, y que incluyen preconcepciones sobre cómo debe ser la educación, la familia, el trabajo y la jubilación [18].

Cuando las generaciones más jóvenes y las mayores no se involucran socialmente, es cuando florece la discriminación por edad [13]. Por otro lado, la teoría de la amenaza intergrupal sugiere que las personas reaccionan de manera hostil hacia grupos externos, particularmente cuando se perciben como potencialmente dañinos.

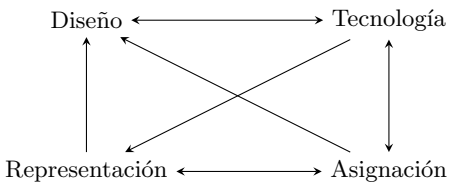


Fig. 1. Modelo de ciclos de injusticia adaptado de [7].

Esta teoría identifica dos tipos de amenazas, las reales y las simbólicas, que sirven para aumentar la hostilidad y el conflicto entre grupos. Las amenazas reales se refieren a las amenazas al poder, los recursos y el bienestar del grupo; las amenazas simbólicas afectan el sistema de creencias y valores del grupo [20].

Así, de acuerdo con estas teorías mesonivel, las causas del edadismo son interpersonales, por lo que alguien podría justificar que tiene un sesgo contra las personas adultas mayores pero que no es edadista, apelando a que tenemos una predeterminación genética para discriminar, a que la sociedad occidental ya tiene guiones de vida preconcebidos, o a que las personas adultas mayores son una amenaza a los valores del bienestar.

Con todo, como en el caso de las teorías micronivel, estas causas explican, pero no justifican. Por último, tenemos teorías macronivel según las cuales los valores culturales son la causa del desprecio a las personas adultas mayores. La teoría de la modernización postula, por ejemplo, que a través del proceso de modernización social, que incluye avances en tecnología y medicina, las personas adultas mayores han perdido su estatus social.

Paradójicamente, los avances en tecnología y medicina han resultado en un mayor número de personas adultas mayores, pero no en una adecuada valoración de las mismas. La vejez ya no representa supervivencia exitosa, sino fragilidad y discapacidad; y ya no necesariamente representa sabiduría, sino obsolencia. Y

aunque esta teoría se ocupa principalmente de la disminución del estatus de las personas adultas mayores, también predice un aumento en el poder y el estatus de las generaciones más jóvenes [15].

Por tanto, de acuerdo con esta teoría macronivel, las causas del edadismo son institucionales y culturales, por lo que alguien podría justificar que tiene un sesgo contra las personas adultas mayores pero que no es edadista apelando a que nuestra condición socio-cultural o nuestro momento histórico nos da licencia para discriminar; sin embargo, como ocurre con las explicaciones anteriores, todos estos aspectos descriptivos dan cuenta de las múltiples causas del edadismo pero no lo justifican.

Y así, si estos argumentos y teorías son de alguna utilidad, debería ser para mostrar que el edadismo es un problema real, con causas y efectos reales nuestro primer objetivo; no obstante, otra cosa es mostrar que la IA no está libre de este sesgo. Por ello, para argumentar que sí puede estarlo nuestro segundo objetivo, consideremos que tanto el desarrollo como el uso de las tecnologías de la información han excluido a las personas adultas mayores creando una brecha digital y su exclusión social en el desarrollo y uso tecnologías digitales muestra un sesgo relacionado con la edad en la IA [19]. Para visualizar esto, consideremos un par de modelos.

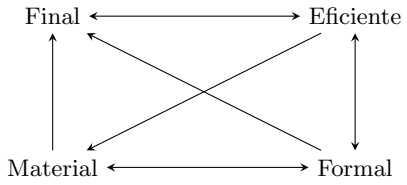


Fig. 2. Modelo causal.

El primero es de [7] y nos permite explicar cómo los ciclos de injusticia en las tecnologías digitales dan como resultado un edadismo en las aplicaciones de IA (Figura 1). De acuerdo con este modelo de [7] de ciclos de injusticia, las aplicaciones de IA pueden producir y reforzar los sesgos de edad a través de múltiples vías. Los estereotipos existentes sobre las personas adultas mayores como tecnológicamente incompetentes (Representación) afectan los prejuicios, lo que puede conducir a la exclusión de estas personas de los procesos de investigación y diseño de tecnologías (Diseño).

Por ejemplo, los estereotipos de edad se ven reforzados por el hecho de que las nuevas tecnologías de la información para personas de edad se centran principalmente en resolver problemas de salud y envejecimiento (Tecnología). Así, la brecha digital (Asignación), junto con los patrones en las aplicaciones existentes, da como resultado conjuntos de datos que representan de manera imprecisa al resto de personas de edad (Tecnología).

Estos conjuntos de datos sesgados incentivan un mayor desarrollo tecnológico que se centra principalmente en las necesidades de atención médica (Diseño),

pero la disponibilidad limitada de tecnologías digitales que atienden otras necesidades, intereses y aspiraciones de las personas de edad termina aumentando aún más la brecha digital (Asignación).

De esta manera, los nuevos sistemas refuerzan la desigualdad y magnifican la exclusión social de subsectores de la población que se consideran como una "clase inferior digital" compuesta principalmente por grupos de personas adultas mayores, pobres y marginadas [7]. Si este modelo no basta para ilustrar nuestro segundo objetivo que el edadismo puede estar presente en la IA, podemos ofrecer otro modelo basado en la distinción entre causas materiales, formales, eficientes y finales de Aristóteles (Física II, 3 y Metafísica V, 2).

Según este modelo, una causa material corresponde con aquello de lo cual esta compuesto un ítem, por ejemplo, el mármol de una estatua; una causa formal es aquella que da cuenta de la estructura de dicho ítem, como la forma de la estatua. La causa eficiente, por otro lado, es la fuente del ítem, como la persona que produce la estatua. Y por último, una causa final se define como aquello para lo cual se produce un ítem, como que conmemorar es el fin de la estatua.

En consecuencia, de acuerdo con este modelo, una aplicación o un sistema de IA sería pluricausal: las causas materiales de una aplicación o sistema de IA serían los datos; las causas formales, los algoritmos; las causas eficientes, las personas o instituciones que diseñan los algoritmos y obtienen los datos; las finales, las personas para las cuales se diseñan las aplicaciones.

Dicho de otro modo, un sistema de IA o una aplicación de IA no es el algoritmo, pero no es lo que es sin algoritmo; no son los datos, pero no puede ser sin datos; no es quien diseña, pero no puede ser sin diseño; y no es una usuaria, pero no tiene razón de ser sin usuarias (Figura 2).

De acuerdo con este modelo causal, las aplicaciones o sistemas de IA pueden producir edadismo cuando los datos (Materia) no son representativos, lo que puede causar que quienes las diseñan (Eficiencia) no tomen en cuenta ciertos elementos importantes en sus algoritmos (Forma), por lo que no estarían tomando en cuenta a las usuarias finales (Final).

De esta forma, las aplicaciones de IA pueden reproducir o crear relaciones de inequidad y exclusión social. En particular, cuando los datos y quienes diseñan algoritmos reproducen sesgos edadistas, con especial énfasis en personas adultas mayores, las personas para las cuales se diseñan las aplicaciones pueden encontrarse en relaciones ilegítimas que no tienen un vínculo relevante con la justicia social.

Y así, hasta este punto, habríamos logrado nuestros dos objetivos; sin embargo, algunas personas podrían no estar convencidas de que el edadismo es un problema real o de que la IA no está libre de este sesgo. Algunas de las posibles objeciones que se podrían ofrecer para justificar semejante creencia podrían ser las siguientes.

Objeción 1. El error categorial. La inteligencia artificial no puede ser edadista, porque solo las personas pueden serlo, y los sistemas de inteligencia artificial no son personas. Respuesta: es verdad que la IA no es edadista, pero ese no es el punto de esta contribución. Seguramente, tampoco diríamos que la ciencia

es racista solo porque han existido ciertos programas de investigación racistas. El punto de esta contribución es mostrar, más bien, que ciertas aplicaciones de IA podrían estar introduciendo y reproduciendo sesgos edadistas, como que es verdad que han existido programas científicos con políticas racistas.

Objeción 2. La exageración. El problema del edadismo es una exageración, no es tan grave. Respuesta: en primer lugar, reconocer que un problema es una exageración no implica que no sea un problema; y en segundo lugar, una vez que se concede que sí es un problema, la determinación de su gravedad puede parecer una cuestión de carácter subjetivo; sin embargo, el hecho de que la ONU y la OMS hayan propuesto a la década 2021-2030 como la Década del Envejecimiento Saludable es un indicador de que el problema del edadismo está en tendencia.

Objeción 3. La minimización. El problema del edadismo ya está considerado en la inteligencia artificial, pues hay toda una rama de la misma dedicada al mejoramiento de la vida de las personas que están envejeciendo. Respuesta: el problema de esta objeción es que pierde de vista una diferencia importante, a saber, que edadismo (ageism) y envejecimiento (ageing) no son coextensivos. Sin duda, está muy bien que haya toda una rama de la IA dedicada a favorecer la vida durante el envejecimiento, pero eso no implica que el edadismo no es un problema más. Sería contraintuitivo argumentar que como hay una rama del derecho que se dedica a resolver problemas de género, entonces no existe el problema del sexismo, por poner un contraejemplo.

Objeción 4. La normalización. El problema del edadismo se irá resolviendo poco a poco y de manera orgánica, conforme avance la disciplina, por lo que no es necesario hacer esfuerzos adicionales por resolver un problema que se irá resolviendo por sí solo.

Respuesta: hay muchos problemas sociales que pueden resolverse de manera orgánica, pero ello no implica que no debamos llevar a cabo ciertas acciones concretas para resolverlo. Por ejemplo, también hay problemas de salud que, seguramente, se pueden ir resolviendo poco a poco, pero ello no implica que no tengamos que hacer esfuerzos adicionales.

Objeción 5. La responsabilidad. El problema del edadismo no es problema de la IA, es problema de las mismas personas adultas mayores que no se saben adaptar a los rápidos y continuos cambios tecnológicos. Respuesta: aun si es verdad que algunas personas adultas mayores no se pueden adaptar a los cambios tecnológicos, eso no implica que tengan que padecer las inequidades resultantes de su falta de adaptación. El problema más grave de esta objeción es que no hace visible las intersecciones con la edad y es tan débil como aquella que pretende revictimizar a las víctimas.

4. Conclusiones

Dado que el sesgo de edad ha sido escasamente explorado en el contexto de la IA, en este trabajo hemos intentado alcanzar dos metas: primero, exponer el problema de la discriminación por edad con especial énfasis en la situación de

las personas adultas mayores; y segundo, argumentar que la IA no está libre de este sesgo.

Para lograr estos objetivos hemos presentado, primero, el concepto de edadismo, sus causas y algunas razones por las cuales no tiene justificación; y segundo, hemos mostrado un par de modelos que ilustran cómo la IA no necesariamente está libre de este sesgo. Si estas consideraciones y estos objetivos tienen algún sentido, esperamos que sean suficientemente interesantes como para atraer más atención a un problema crítico que estará en tendencia.

Quienes participamos de las disciplinas asociadas a la IA deberíamos estar en primera fila para enfrentarlo. Quienes compartimos este presente deberíamos reconsiderar nuestro futuro, porque allí pasaremos el resto de nuestras vidas. Por último, para cerrar, nos gustaría comentar cuatro temas a considerar en trabajos futuros:

a) Es necesario establecer un marco ético-jurídico para enfrentar los desafíos del edadismo en la IA.
b) Igualmente, es preciso discutir un marco ético-político para normar los mecanismos de creación y distribución de aplicaciones de IA.
c) Es deseable, además, diseñar programas educativos y de divulgación para preparar a las personas que forman parte del ciclo de creación y consumo de sistemas de IA.
d) Para avanzar con paso firme en todas estas líneas, es justo ofrecer evidencia empírica de la presencia de este sesgo, especialmente en México.

Agradecimientos. Nos gustaría agradecer a los revisores por sus valiosas observaciones y precisas correcciones. Este trabajo fue financiado por un Proyecto de Investigación UPAEP y por el fondo común del Instituto Promotor del Bien Común.

Referencias

1. Angwin, J., Kirchner, L., Larson, J., Mattu, S.: Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. ProPublica, (2016)
2. Ayalon, L., Tesch-Römer, C.: Introduction to the section: Ageism - concept and origins, pp. 1–10. Springer International Publishing (2018)
3. Burnstein, E., Crandall, C. S., Kitayama, S.: Some neo-darwinian decision rules for altruism: Weighing cues for inclusive fitness as a function of the biological importance of the decision. Journal of Personality and Social Psychology, vol. 67, pp. 773–789 (1994)
4. Butler, R. N.: Age-ism: Another form of bigotry. The Gerontologist, vol. 9, no. 4 Part 1, pp. 243–246 (1969) doi: 10.1093/geront/9.4_Part_1.243
5. Caliskan, A., Bryson, J. J., Narayanan, A.: Semantics derived automatically from language corpora contain human-like biases. Science, vol. 356, no. 6334, pp. 183–186 (2017) doi: 10.1126/science.aal4230

6. Chen, I. Y., Szolovits, P., Ghassemi, M.: Can AI help reduce disparities in general medical and mental health care? AMA journal of ethics, vol. 21, no. 2, pp. E167—179 (2019) doi: 10.1001/amajethics.2019.167
7. Chu, C. H., Nyrup, R., Leslie, K., Shi, J., Bianchi, A., Lyn, A., McNicholl, M., Khan, S., Rahimi, S., Grenier, A.: Digital ageism: Challenges and opportunities in artificial intelligence for older adults. The Gerontologist, (2022) doi: 10.1093/geront/gnab167
8. Danks, D., London, A. J.: Algorithmic bias in autonomous systems. In: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17. pp. 4691–4697 (2017) doi: 10.24963/ijcai.2017/654
9. Dastin, J.: Amazon scraps secret AI recruiting tool that showed bias against women. Reuters, (2018)
10. Datta, A., Tschantz, M. C., Datta, A.: Automated experiments on ad privacy settings. In: Proceedings on Privacy Enhancing Technologies. vol. 2015, pp. 92–112 (2015) doi: doi:10.1515/popets-2015-0007
11. Friedman, B., Nissenbaum, H.: Bias in computer systems. ACM Trans. Inf. Syst., vol. 14, no. 3, pp. 330–347 (1996) doi: 10.1145/230538.230561
12. Greenberg, J., Pyszczynski, T., Solomon, S.: The causes and consequences of a need for self-esteem: A terror management theory (1986)
13. Hagestad, G. O., Uhlenberg, P.: The social separation of old and young: A root of ageism. Journal of Social Issues, vol. 61, no. 2, pp. 343–360 (2005) doi: https://doi.org/10.1111/j.1540-4560.2005.00409.x
14. Howard, A. M., Borenstein, J.: The ugly truth about ourselves and our robot creations: The problem of bias and social inequity. Science and Engineering Ethics, vol. 24, pp. 1521–1536 (2018)
15. Johnson, M., Curran, J., Cowgill, D., Holmes, L.: Aging and modernization (1972)
16. K, C.: The trouble with bias—NIPS 2017 keynote. The Artificial Intelligence Channel, (2017)
17. Obermeyer, Z., Powers, B., Vogeli, C., Mullainathan, S.: Dissecting racial bias in an algorithm used to manage the health of populations. Science, vol. 366, no. 6464, pp. 447–453 (2019) doi: 10.1126/science.aax2342
18. Riley, M. W., Kahn, R. L., Foner, A., Mack, K. A.: Age and structural lag : society's failure to provide meaningful opportunities in work, family, and leisure. Contemporary Sociology, vol. 25, pp. 382 (1996)
19. Rosales, A., Fernández-Ardèvol, M.: Structural ageism in big data approaches. Nordicom Review, vol. 40, no. s1, pp. 51–64 (2019) doi: doi:10.2478/nor-2019-0013
20. Stephan, W. G., Ybarra, O., Morrison, K. R.: Intergroup threat theory (2011)
21. Tajfel, H., Turner, J. C.: An integrative theory of intergroup conflict. (1979)
22. Vauclair, C. M., Rodrigues, R. B., Marques, S., Esteves, C. S., Cunha, F., Gerardo, F.: Doddering but dear ... even in the eyes of young children? age stereotyping and prejudice in childhood and adolescence. International Journal of Psychology, vol. 53, pp. 63–70 (2018)
23. WHO: Global Report on Ageism. UN (2021)
24. Zhavoronkov, A., Mamoshina, P., Vanhaelen, Q., Scheibye-Knudsen, M., Moskalev, A., Aliper, A.: Artificial intelligence for aging and longevity research: Recent advances and perspectives. Ageing Research Reviews, vol. 49, pp. 49–66 (2019) doi: 10.1016/j.arr.2018.11.003

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

http://rcs.cic.ipn.mx

Centro de Investigación
en Computación