EDUCACIÓN
SECRETARÍA DE EDUCACIÓN PÚBLICA

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

Research in Computing Science

Vol. 154 No. 5
May 2025

Research in Computing Science

***RESEARCH IN COMPUTING SCIENCE**, Año 25, Volumen 154, No. 5, Mayo de 2025, 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 Othón 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, otorgado por el Instituto Nacional del 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 Othón de Mendizábal, Col. Nueva Industrial Vallejo, C.P. 07738. Fecha de última modificación 05 de Mayo de 2025.*

***RESEARCH IN COMPUTING SCIENCE**, Year 25, Volume 154, No. 5, May, 2025, is a monthly publication edited by the National Polytechnic Institute through the Center for Computing Research. Av. Juan de Dios Bátiz S/N, Esq. Miguel Othón de Mendizábal, Nueva Industrial Vallejo, C.P. 07738, Mexico City, Tel. 57 29 60 00, ext. 56571. https://www.rcs.cic.ipn.mx. Editor in charge: Dr. Grigori Sidorov. Reservation of Exclusive Use Rights of Title No. 04-2019-082310242100-203. ISSN: pending, granted by the National Copyright Institute. Responsible for the latest update of this issue: the Computer Research Center, Dr. Grigori Sidorov, Av. Juan de Dios Bátiz S/N, Esq. Av. Miguel Othón de Mendizábal, Col. Nueva Industrial Vallejo, C.P. 07738. Last modified on May 5, 2025.*

Advances in Artificial Intelligence

Obdulia Pichardo-Lagunas (ed.)

Instituto Politécnico Nacional, Centro de Investigación en Computación
México 2025

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

Ciencia de datos y geointeligencia en la mejora de la experiencia del cliente

Grisel A. Porras, Alberto L. Munoz

Instituto Tecnológico de Estudios Superiores de Monterrey, Mexico

griselporras89@gmail.com, amunoz@tec.mx

Resumen Por medio de este artículo de investigación, se realizó una propuesta para la diversificación de servicios en gasolineras con la tecnología Swit en N.L, México, teniendo como eje central la aplicación de técnicas avanzadas de análisis de datos y geointeligencia. A nivel local, se identificaron patrones de consumo que proporcionaron información relevante para la toma de decisiones como la implementación de nuevos servicios dependiendo de las características demográficas y opiniones sobre el servicio.

Palabras clave: Geointeligencia, análisis geoespacial, ciencia de datos, geomarketing, gasolineras, diversificación, experiencia del cliente.

Leveraging Data Science and Geointelligence to Enhance Customer Experience

Resumen This research article presents a proposal for the diversification of services at gas stations using Swit technology in Nuevo León, Mexico, with a focus on the application of advanced data analysis and geointelligence techniques. At the local level, consumption patterns were identified, providing valuable insights for decision-making—such as the implementation of new services based on demographic characteristics and customer feedback.

Keywords: Geointelligence, geospatial analysis, data science, geomarketing, gas stations, diversification, customer experience.

1 Introducción

Hoy en día, la estrategia de diversificación juega un papel importante para el desarrollo financiero de una empresa. Con el avance de la tecnología y la necesidad de seguir a la vanguardia sobre la satisfacción del cliente, ha crecido la cantidad de empresas que sobrepasan los límites de las regiones, industrias y mercados, expandiendo así sus negocios.

Si bien es cierto que dicha estrategia juega un papel importante en el desarrollo prolífero de una empresa, es necesario mencionar que involucra incertidumbre, lo que genera que sea impredecible ante ciertas condiciones. Lo que se plantea es desarrollar conocimientos para contestar la pregunta inminente sobre la diversificación en gasolineras.

Esta estrategia puede realizarse cuando exista la naturaleza de implementar una mejor experiencia al mercado actual y que funcione como un diferenciador para los competidores. Entender estas fuentes de ventaja competitiva de las empresas ha derivado en la dirección del futuro económico, pues aprovechar estos activos permite responder a las oportunidades del entorno, evitando las debilidades y planteando la toma de decisiones consciente [7].

Las estaciones de servicio no han tenido grandes innovaciones en el mercado desde hace décadas, sin embargo, actualmente, el mercado se ha estado intentando abrir, de forma que ahora no solo pueden ofrecer un servicio, sino muchos. La diversificación de hoy en día puede incluir el mantenimiento del carro, servicios de lavado y tiendas de conveniencia.

Según información de PETROIntelligence, hasta junio 2023, se presentaron más de 13,400 estaciones de servicio en México, de las cuales, alrededor de 6,000 pertenecen a otra compañía diferente de Pemex, entre las más grandes se encuentran Oxxogas, Petro Seven y Corpogas. En comparación con años anteriores, la cantidad de estaciones incrementó aproximadamente 3 % cada una. En el caso de Nuevo León, se otorgaron concesiones por cada 88 km^2 de superficie. Además, en Nuevo León, se estima que la participación de marcas distintas a Pemex se ubica en un 72 % [10].

Conociendo un poco este contexto, abre la posibilidad de un nuevo mercado, en donde se aprovecha el área de oportunidad en la que los usuarios no reciben un servicio óptimo. De esta manera, se pueden capitalizar los tiempos muertos de espera durante el tiempo de abastecimiento.

Swit es un servicio que ha tomado forma desde hace unos años. Gracias a él, se han incorporado nuevas formas de adquirir y pagar servicios desde la gasolinera aprovechando los tiempos de espera, representando una estrategia de diversificación que mejora la experiencia del cliente.

Para mejorar estos esfuerzos, la ciencia de datos juega un papel fundamental para la toma de decisiones informadas. Por medio del análisis geoespacial, se busca implementar este servicio especializado, identificando áreas de enfoque prioritario para crear nuevas estrategias para la empresa. De este modo, se podrá implementar el servicio en una estación específica solo si se considera beneficioso para la zona en donde se evalúa.

Entran aquí factores importantes que se pueden o no relacionar entre sí y que subyacen dentro de la dinámica del sistema. Por ejemplo, la zona del país donde se encuentran, las restricciones de movilidad, la cercanía con otros puntos de interés así como el costo que cuesta el llenado. Estos son factores que, si son cuidadosamente considerados, pueden maximizar su éxito.

En relación con lo anterior, la presente investigación plantea analizar las características del cliente para obtener información más detallada de la industria

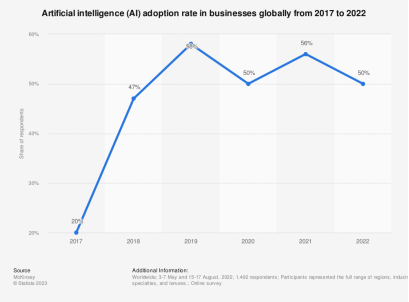


Fig. 1. Tasa de adopción de Inteligencia Artificial en empresas del 2017 al 2022.

y así tomar decisiones. Se propone realizar el desarrollo del análisis a partir del posicionamiento de diferentes variables que determinan en gran medida la rentabilidad del servicio que ofrece Swit.

2 Hipótesis

El conjunto de series de tiempo, clasificación, NLP y clustering de datos, sustentado por la información derivada de fuentes públicas para el análisis geoespacial, permitirá identificar tendencias de consumo en la industria energética, lo que se traducirá en una predicción más acertada de las necesidades del usuario y, por ende, en un incremento significativo en la experiencia del cliente dentro de las estaciones de servicio en evolución.

3 Estado del arte

Históricamente, la ciencia de datos en negocios siempre ha sido utilizada de forma empírica. Actualmente, según información de McKinsey & Company (Figura 1) comentan que la Inteligencia Artificial ha sido bien recibida como una herramienta que funciona como valor agregado en las empresas. Como bien se puede ver en la gráfica, desde el 2017 el valor de la tasa de adopción en los negocios se ha duplicado hasta el 2022. [5] Igualmente, aplicar estas técnicas permite generar estrategias que traen consigo mayor eficiencia y experiencias de cliente más positivas [1]. Por ejemplo, como se puede ver en la siguiente gráfica, las ganancias relacionadas al uso de inteligencia artificial aplicadas en negocios de todo el mundo ha traído consigo resultados satisfactorios para las mismas *(véase Figura 2)*.

Esta aplicación de nuevas estrategias con Inteligencia Artificial son beneficiosas en más de un departamento de una empresa. Como se observa en la siguiente gráfica (*Figura 3*) que se deriva del estudio realizado por McKinsey, se puede ver que el máximo incremento de las ganancias con el uso de programas con Inteligencia Artificial fue en el departamento de Marketing y ventas, así como en el Desarrollo de productos o servicios. [5]

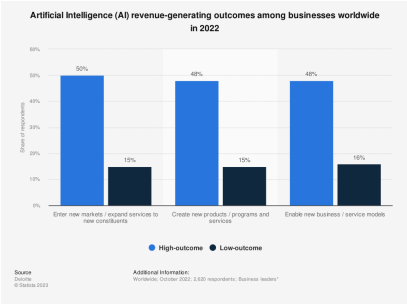


Fig. 2. Resultados de aplicar Inteligencia Artificial en empresas al rededor del mundo durante el 2022.

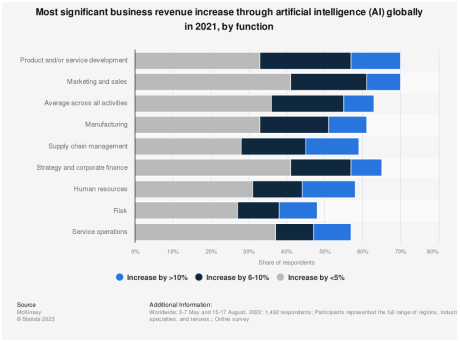


Fig. 3. Incremento de ganancias por el uso de Inteligencia Artificial en Negocios durante el 2021.

Aunque se espera que el mercado de Inteligencia Artificial tenga un crecimiento en los próximos años, aún queda un largo camino por recorrer, ya que muchas empresas todavía no ponen en práctica estos conocimientos. En la encuesta realizada por Forbes sobre el *'Deployment status of data science and machine learning in organizations worldwide as of 2019, by function'* se comenta que solo el 26.5 % dentro de las empresas encuestadas usan técnicas de ciencia de datos y machine learning. Por otro lado, solo el 32.5 % de las compañías encuestadas las utiliza en el departamento de Business intelligence competency center. [11]

En cuanto a las gasolineras, existe poca documentación relacionada con la diversificación de servicios, ya que se trata de un caso muy específico. No obstante, el artículo Çonstructing marketing decision support systems using data diffusion technology: A case study of gas station diversification"[8] expone los esfuerzos realizados en torno a esta transformación.

El estudio se centra en una empresa de Taiwán dedicada a la distribución de gasolina, la cual, debido a cambios en el mercado energético, la competencia creciente y la presión por la sostenibilidad, se ha visto obligada a diversificar sus

operaciones. Ante la disminución de márgenes en la venta de combustibles, la empresa ha tenido que explorar nuevas líneas de negocio dentro de sus estaciones de servicio, como tiendas de conveniencia, servicios de mantenimiento vehicular, lavado de autos, e incluso la instalación de cargadores eléctricos. Dado que se dispone de información limitada sobre el tema, el objetivo es evaluar la viabilidad de incorporar un nuevo servicio en alguna de sus estaciones. Para ello, se consideran distintas variables como:

- El precio es competitivo
- La cantidad de equipo es adecuado
- Los servicios son personalizados para la satisfacción del cliente
- El staff sirve a los consumidores de forma rápida
- El stafff sirve a los consumidores con entusiasmo

Además se miden otras variables como la lejanía a una carretera, al pueblo más cercano y a la ciudad más cercana. Apuntando a cada gasolinera con estas características para clasificar en clases A,B y C las gasolineras. Como los datos son pocos, la propuesta es utilizar una técnica de mega-trend-diffusion (MTD) para construir la toma de decisiones a partir de una red bayesiana. Según los resultados la precisión del modelo aumentó al utilizar este enfoque.

De igual forma, existen otros enfoques relacionados con geointeligencia. Dentro de ellos se encuentran el artículo *'Site selection for small gas stations using GIS'* [2] donde se realizó un modelo para posicionar de forma estratégica la ubicación óptima de una estación de combustible con GIS en Irán. Se compararon diferentes áreas, evaluando factores en cada una de ellas como:

- Distancias a la avenida principal
- Distancias a carreteras
- Distancias a estacionamientos
- Distancias a la ciudad más cercana
- Distancias a otras gasolineras
- Distancias a escuelas
- Distancia a hospitales

Cada factor, contiene un peso asociado en el que se busca atribuir a cada lugar posible una calificación. En este enfoque, se utilizaron diferentes técnicas y se probó cada una. Se encontró que los métodos como el Fuzzy analytical hierarchy process (FAHP) y analytical hierarchy process (ANP) eran los más flexibles.

De igual forma, también se ha realizado una investigación sobre estrategias de geomarketing aplicadas a un tipo de zona específica. Los artículos *'Creating a geodemographic classification model within geo-marketing: the case of Eskişehir province'* [4] y *'Building a Geo-Demographic Segmentation Model: the Case of Hanoi City, Vietnam'* [6] tienen un enfoque parecido. En ambos casos, se estudia una subdivisión de dos países, en el primer caso, en la ciudad de Eskişehir, Turkia y Hanoi, Vietnam.

En relación con lo anterior, se utilizaron técnicas como K-means Clustering, Principal Component Analysis (PCA), Hierarchical Clustering Analysis que dio como resultado la segmentación geodemográfica de los barrios de estas dos ciudades, para encontrar características de personas compartidas. Características como la edad, nivel de educación redujeron el costo y el tiempo en el proceso de ingresar o expandir los negocios en una nueva área.

Por último, se ha realizado una investigación sobre el uso de la información recabada por dispositivos IOT en el artículo *'Exploring the application of IoT in the service station business'* [9], donde se hizo uso de series de tiempo para entender el perfil del cliente. Por ejemplo, ver la relación que existía con el precio de la gasolina en la llegada de clientes, las horas pico y el tráfico en general. En este estudio se buscaba segmentar al mercado para poder ofrecer servicios más especializados que mejoraran la satisfacción hacia la estación, como la optimización del horario de empleados.

4 Descripción de las metodologías

La metodología propuesta para esta investigación combina diferentes direcciones cualitativas y cuantitativas. Gracias a ellas, será posible procesar los datos para analizar y predecir las tendencias de la industria energética en estaciones de servicio.

Tomando esto en consideración, se utilizará la metodología Cross Industry Standard Process for Data Mining, por sus siglas en inglés CRISP-DM, la cual, combina todas las etapas aplicables en un proyecto. El modelo contiene seis etapas dependientes entre sí, donde las secuencias no es completamente estricta [12].

1. Comprensión del negocio: Se establece el objetivo del proyecto para identificar las tendencias de consumo en la industria energética para mejorar la satisfacción del cliente. Asimismo, se realiza una evaluación para comprender el contexto de las necesidades actuales del cliente, así como los desafíos que implican proponer diferentes soluciones para las oportunidades de mejora.
2. Comprensión de los Datos: Primero, se realizará una recopilación de los datos por medio de fuentes públicas. Para fines de este proyecto se estará utilizando información del Instituto Nacional Estadística y Geografía (INEGI) para recopilar datos sobre gasolineras e información geoespacial que tengan que ver con el comportamiento del usuario. Posteriormente, se estarán evaluando los datos crudos para comprender un poco mejor las tendencias.
3. Preparación de los datos: El siguiente paso es la limpieza de los datos, incluyendo los faltantes, outliers y otros errores que pudieran existir. En adición a esto, se realizarán las transformaciones necesarias.
4. Modelos: Se estarán utilizando diferentes técnicas de Ciencia de Datos mencionadas anteriormente.

5. Evaluación: Evaluar el rendimiento de los modelos a partir de diferentes métricas. Ajustar hiperparámetros y comparar diferentes modelos para evitar underfitting y overfitting.
6. Despliege: Implementar los modelos realizados, actualizarlos y monitorearlos para garantizar su buen funcionamiento.

5 Diseño de experimentos

Los experimentos se estructurarán en tres fases bien definidas durante el proyecto para lograr los objetivos planteados al inicio del documento. En la primera fase, los datos que se estarán analizando provienen de diferentes fuentes como el Directorio Estadístico Nacional de Unidades Económicas (DENUE) y el Censo de Población y Vivienda 2020 (SCINCE 2020) provenientes del INEGI con el objetivo de recopilar datos sociodemográficos para llevar a cabo una clasificación exhaustiva de las zonas geográficas. Esto será útil para segmentar al mercado de forma más precisa y adaptar las estrategias según las características de cada área. Las variables que se estarán utilizando se pueden encontrar en la siguiente tabla (*Tabla* 1)

La selección de indicadores sobre las características sociodemográficas de la población y las viviendas del país vienen desagregados hasta el nivel de manzanas. Los indicadores se dividen en diferentes categorías, sin embargo, solo se trabajaron con los temas relacionados a Educación, Características Económicas, Hogares Censales y Vivienda. Esto con el objetivo de entender a mayor profundidad sobre la situación actual de cada vivienda, así como el estilo de vida que lleva cada zona en específico. Para facilitar el manejo y análisis de los datos, se construyeron nuevas variables a partir de los indicadores originales relacionados con educación y edad. Estas transformaciones permitieron condensar la información y agrupar a la población en categorías significativas para el análisis sociodemográfico. Se generaron las siguientes variables: Población Infantil, correspondiente a personas de 0 a 17 años; Adultos Jóvenes, que agrupa a la población de 18 a 29 años; Adultos en Edad Productiva, que incluye a personas de 30 a 49 años; y Adultos Mayores Activos, que representa a la población entre 50 y 64 años. En cuanto al nivel educativo, se definió la variable Escolaridad Baja, que considera a personas de 15 años o más con educación básica incompleta o al menos un grado cursado en educación media superior; y Escolaridad Superior, que incluye a personas de 25 años o más con al menos un grado aprobado de nivel superior.

Posteriormente, para un análisis más profundo, se estará evaluando la información de las reseñas de Google de veinte gasolineras con el fin de extraer información relevante y patrones de opinión de los usuarios relacionados con el tiempo de espera y el servicio al cliente.

En la segunda fase del proyecto, se procederá al entrenamiento de modelos de inteligencia artificial. Estos modelos servirán de punto de partida para la automatización relacionada con la toma de decisiones. Finalmente, en la tercera fase, se realizarán pruebas en entornos controlados para evaluar la eficacia,

Tabla. 1. Variables utilizadas en la segmentación de las zonas geográficas.

Variable	Significado
$ECO1_R$	Población de 12 años y más económicamente activa.
$VIV29_R$	Porcentaje de viviendas particulares habitadas que disponen de automóvil o camioneta.
$VIV33_R$	Porcentaje de viviendas particulares habitadas que disponen de televisor.
$VIV34_R$	Porcentaje de viviendas particulares habitadas que disponen de computadora, laptop o tablet.
$VIV35_R$	Porcentaje de viviendas particulares habitadas que disponen de línea telefónica fija.
$VIV36_R$	Porcentaje de viviendas particulares habitadas que disponen de teléfono celular.
$VIV37_R$	Porcentaje de viviendas particulares habitadas que disponen de Internet.
$VIV82_R$	Porcentaje de viviendas particulares habitadas que disponen de servicio de televisión de paga.
$VIV83_R$	Porcentaje de viviendas particulares habitadas que disponen de servicio de películas, música o videos de paga por Internet.
$EDU34_R$	Población de 15 años y más con educación básica incompleta.
$EDU37_R$	Población de 15 años y más con educación básica completa.
$EDU43_R$	Población de 18 años y más con al menos un grado aprobado en educación media superior.
$EDU46_R$	Población de 25 años y más con al menos un grado aprobado en educación superior.
$POB1$	Población total.
$POB8$	Población de 0 a 14 años.
$POB9$	Población de 15 a 17 años.
$POB13$	Población de 18 a 24 años.
$POB14$	Población de 30 a 49 años.
$POB15$	Población de 50 a 59 años.
$POB16$	Población de 60 a 64 años.
$POB30$	Población de 25 a 29 años.
$POB31$	Población de 30 a 34 años.
$POB32$	Población de 35 a 39 años.
$POB33$	Población de 40 a 44 años.
$POB34$	Población de 45 a 49 años.

precisión y adaptabilidad de los sistemas desarrollados. En primer lugar, el experimento se realizará en el estado de Nuevo León, se planea probarlo en dos entornos, Primero, en zonas conglomeradas con alto nivel poblacional (como el centro de Monterrey) donde se encuentran zonas específicas relacionadas con el nivel socioeconómico y otros aspectos relevantes como la edad.

Fig. 4. Cluster de manzanas utilizanado K-means, con k =5.

Posteriormente se estará evaluando en zonas con menor accesibilidad a servicios como centros de conveniencia y centros de acceso para el pago de servicios, así como menor nivel poblacional, por ejemplo, cerca de Galeana, al sur de Nuevo León.

Después de realizar la limpieza y transformación de datos, se propuso revisar la implementación de diferentes modelos de Machine Learning para encontrar las manzanas donde sería viable aplicar el servicio de Swit. Como se mencionó anteriormente, las variables utilizadas recopilan información sociodemográfica como las características de vivienda, la edad y la educación, así como la densidad poblacional. Adicional a eso, se estudia la zona, considerando tiendas de conveniencia como Oxxo o 7-Eleven, escuelas y la cantidad de cajeros CFEmáticos que se encuentran a una distancia cercana. Esto permite identificar las zonas de interés y la influencia de los competidores cercanos.

Primeramente, se realizó la segmentación de las zonas de vivienda en Nuevo León. Se puede observar el resultado usando k-means como se ve a continuación en la siguiente imagen (*Figura 4*). En este caso la predicción de 0 arroja las manzanas que cuentan con todos los servicios, por ejemplo, San Pedro. Por el otro lado, mientras más se aleja de este número, se acerca a zonas con mayor escasez de servicios no básicos, por ejemplo, la colonia Independencia.

En el siguiente mapa se puede visualizar las zonas con amplio potencial para colocar el servicio, pues son zonas retiradas de la ciudad metropolitana. Como se puede observar (*Figura 5*), las zonas que presentan mayor área de oportunidad en la accesibilidad al pago de servicios es la zona sur.

6 Resultados del Modelo

Con el objetivo de identificar zonas con alto potencial para la implementación del servicio, se entrenaron y evaluaron dos modelos de clasificación supervisada: **Random Forest** y **Regresión Logística**. Estos modelos fueron seleccionados por su solidez teórica, interpretabilidad y buen desempeño en tareas con múltiples variables categóricas y numéricas. [3]

Para la evaluación de ambos modelos, se utilizó el **área bajo la curva ROC (ROC AUC)**, una métrica estándar que mide la capacidad del modelo para

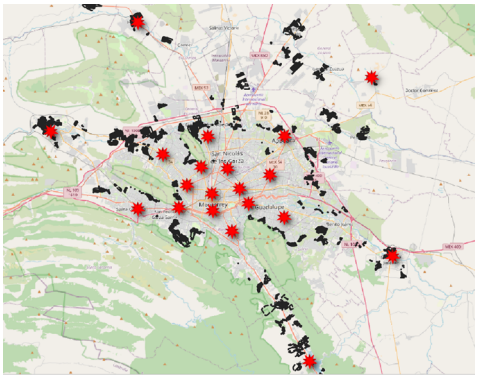


Fig. 5. Manzana con menos de 3 gasolineras y puntos referenciados de cajeros automáticos.

distinguir entre clases sin necesidad de fijar un umbral específico. Los resultados obtenidos fueron:

- **Random Forest:** ROC AUC = 0.922
- **Regresión Logística:** ROC AUC = 0.876

El modelo Random Forest demostró un desempeño superior, con una mayor capacidad discriminativa entre zonas óptimas y no óptimas. Para optimizar el rendimiento del modelo, se utilizó la técnica de Grid Search, la cual permite explorar de forma sistemática todas las combinaciones posibles de hiperparámetros definidos.

Los hiperparámetros ajustados fueron: número de árboles,profundidad máxima de cada árbol, número mínimo de muestras para dividir un nodo, número mínimo de muestras requeridas en una hoja. Una vez entrenado el modelo con los hiperparámetros óptimos, se procedió al análisis del cutoff a fin de determinar el umbral de decisión que ofrezca el mejor equilibrio entre precisión, recall y utilidad práctica.

Aunque la curva ROC permite evaluar la calidad general del modelo, para utilizarlo en la práctica es necesario establecer un cutoff, es decir, un umbral de probabilidad a partir del cual se clasifica una observación como “zona óptima”.

Se probaron distintos valores de cutoff y se observaron sus efectos en métricas clave como precisión, recall y F1-score. El análisis reveló que un cutoff de aproximadamente 0.42 proporciona el mejor equilibrio entre exactitud y cobertura, permitiendo un balance efectivo entre falsos positivos y falsos negativos. Con este umbral, el modelo Random Forest obtuvo los siguientes resultados:

- **Precisión:** 0.83
- **Recall:** 0.88
- **F1-score:** 0.85

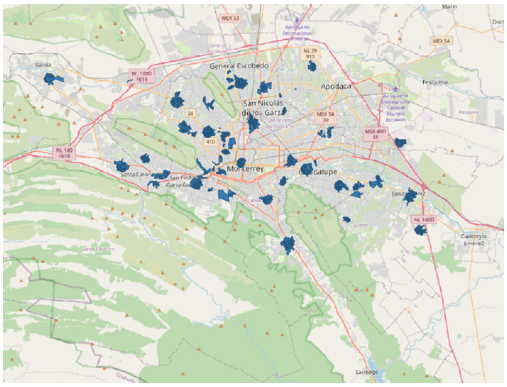


Fig. 6. Lugares potenciales para la aplicación del servicio.

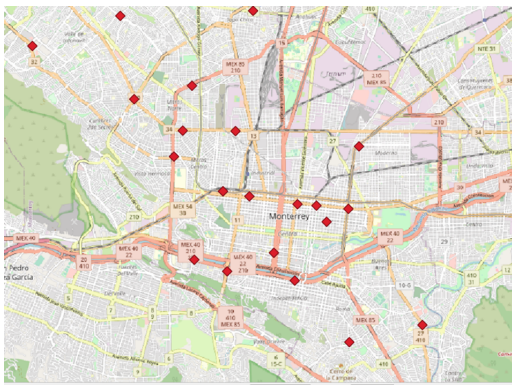


Fig. 7. Estudio de caso: 20 gasolineras.

– **Cost-matrix gain:** 0.48

Estos resultados validan el uso de Random Forest como herramienta central para apoyar decisiones estratégicas de localización, con una base sólida en datos geoespaciales y sociodemográficos.

Después de obtener la clasificación, en la Figura 6 se observa cuáles son las zonas donde se podría aplicar el servicio.

Se realiza entonces el siguiente procedimiento para determinar dónde colocar el servicio. Se estudia el caso de 20 gasolineras del centro de Monterrey elegidas de forma aleatoria. Estas se pueden observar a continuación en la Figura 7.

Para volver más completa la toma de decisiones, se obtuvieron alrededor de 400 comentarios sobre las reseñas de dichas gasolineras por medio de la API Outscrapper, la cual utiliza diferentes técnicas de web scrapping. El producto resultante fue una base resultante de las reseñas, la cual contenía información de la calificación asignada por el usuario y el comentario correspondiente. Como primer acercamiento, se quitaron las palabras comunes como artículos

Fig. 8. Simbología que contiene tres clases, basados en la calificación asignada a cada gasolinera.

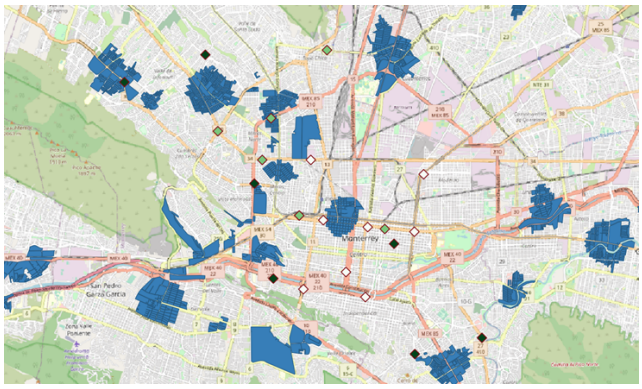


Fig. 9. Mapa que muestra las zonas potenciales junto con las gasolineras clasificadas.

que no servían para el análisis y se obtuvo un texto condensado. Sobresalen valores importantes como "Servicio", "Precio" y "Atención". Estos comentarios representan un sentimiento relacionado que puede estudiarse como Positivo, Negativo o Neutro, según sea el caso. De este modo, se clasificó cada comentario con dicha clasificación. Posteriormente, se evaluaron las manzanas que tenían alrededor de 1500 unidades de distancia una gasolinera cercana para que de este modo se pudiera obtener el tipo de mercado al que podía estar dirigido y la distancia a puntos de interés. A continuación, se muestra el resultado de la calificación asignada a cada gasolinera (*véase Figura 8*). Son tres clases que fueron agrupadas según el valor de la calificación. El tono verde oscuro representa las gasolineras que deberían considerarse para aplicar el servicio de Swit según el modelo.

Si comparamos el mapa propuesto de zonas potenciales (*Figura 9*), los lugares que arroja el resultado pueden ser bastante útiles, pues se encuentran en las zonas arrojadas y cuentan con una experiencia satisfactoria del cliente, que representa una parte importante del éxito de un negocio.

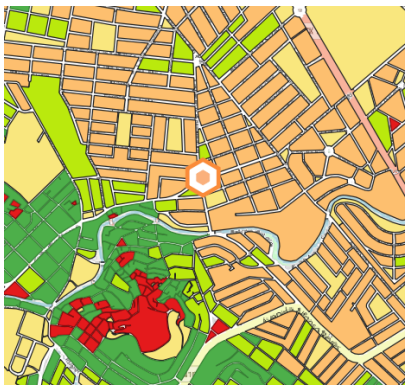


Fig. 10. Caso de estudio: Gasolinera Rio Aguanaval.

7 Descubrimientos

Actualmente cada día hay mayor atención a la toma de decisiones basadas en datos y surgen nombres como geomarketing, geointeligencia o ciencia de datos. Después de analizar el estado del arte, se propone el uso de variables similares para segmentar el mercado y proponer un servicio personalizado. La introducción de un servicio a un nuevo mercado fue el principal problema que fue identificado, ya que representa una gran incertidumbre, muchos de ellos proponían solo segmentar el mercado o solo utilizar puntos de interés sin considerar variables sociodemográficas. En este caso, se propuso un sistema de recomendación automatizado que utilizó diferentes técnicas de machine learning y geointeligencia que pudiera combinar las dos variables y sumarle la opinión del cliente, una variable muy importante a considerar. Este modelo implica un ahorro en el tiempo y una estrategia de ahorro de pérdida segura, pues disminuye el riesgo de que el servicio no funcione en el lugar escogido, no solo para las estaciones de servicio, si no para cualquier lugar en el que se pueda aplicar. Como se menciona anteriormente, se proporciona una metodología que tiene la capacidad de usarse en diferentes industrias y que representa un gran beneficio para sus usuarios.

Se estudia el resultado propuesto que obtuvo la mayor calificación (*Figura 10*). Representa un punto estratégico, pues se encuentra en el centro de varias manzanas con diferente clasificación socioeconómica; de un lado se encuentra la zona cercana a la universidad Tecnológico de Monterrey y la colonia Roma, y en el otro extremo se encuentra cercano el Cerro de la Campana. Se encuentra cerca de escuelas como el Tecnológico de Monterrey, Instituto Excelsior, además de varias primarias y secundarias públicas. Todos estos factores son importantes dentro del modelo. Del mismo modo también cuenta con diferentes rangos de edad pues representa una zona de mucho movimiento para jóvenes, así como personas adultas que reconocen estas manzanas como lugares accesibles y seguros donde se puede rentar. De igual forma, tomaron peso las reseñas encontradas en

Google, pues de las veinte gasolineras estudiadas fue la tercera con el rating más alto (3.9), lo que insta a considerarla como una gasolinera con un buen servicio.

Si bien cuenta con un Oxxo muy cercano es importante considerar que una gran cantidad de gasolineras cuentan con este tipo de servicios de conveniencia por lo cual no afectó tanto la calificación de la gasolinera. Otro punto a tomar en cuenta es que el análisis de las reseñas puede estar sesgado a la falta de comentarios para cada gasolinera. En posibles y futuros trabajos, se espera trabajar con un mayor volumen de reseñas, para adoptar estrategias más personalizadas con las bases de datos que se pudieron analizar sobre la venta de gasolina.

Este es uno de los más grandes desafíos que debe ser abordado en el futuro para tener resultados más confiables y precisos. En el futuro se puede probar el modelo para verificar su eficacia en más ciudades y en otros contextos, no solo en gasolineras; pues la metodología permite adaptarlo y/o extenderlo según sea necesario.

En relación con lo anterior, se propone que el servicio ofrecido no tarde más que el tiempo promedio de espera de cada transacción para no afectar la opinión del usuario, así como ofrecer servicios según la zona. En esta gasolinera en específico se recomienda ofrecer las recargas, el pago de luz y la compra de boletos de cine para potenciar su ubicación estratégica. Del mismo modo, también se propone que se evite ofrecer el servicio en horas pico para no engrandecer el tiempo de espera en la fila y evitar pérdida de clientes.

Por último, tomando en cuenta las consideraciones éticas, es crucial considerar que la recopilación de datos sobre el comportamiento de los usuarios puede comprometer la privacidad de la empresa, pues son datos sensibles. La divulgación indebida de esta información podría tener consecuencias legales y dañar la confianza del cliente; es por esta razón que deben ser protegidos para prevenir el acceso no autorizado o el mal uso de estos.

8 Conclusiones

La diversificación representa una oportunidad a la que se enfrentan diferentes empresas. Gracias a la tecnología, se han desarrollado diferentes formas para facilitar la toma de decisiones conscientes que conlleven a la creación de estrategias flexibles y focalizadas a temas específicos para el desarrollo de una empresa. Por medio de diferentes herramientas de la ciencia de datos, fue posible crear un sistema de recomendación que, por medio del análisis de diferentes fuentes de información, lograra hacer un reconocimiento de las zonas potenciales donde aplicar un nuevo servicio. La ventaja de esta investigación es que tiene una metodología que puede aplicarse no solo a estaciones de servicio, sino a cualquier establecimiento que logre aprovechar los tiempos de espera largos, como bancos, hospitales, lava autos y oficinas de gobierno donde se realicen trámites o en conciertos.

Este reporte muestra la gran adaptabilidad que representa trabajar con Geointeligencia para las diferentes industrias, mostrando así una interactividad

con otros medios como la Ciencia de Datos. Esto abre nuevas posibilidades para la creación de propuestas que mejoren los resultados del modelo, haciéndolo más preciso y mucho más robusto, pues se pueden considerar más variables o ajustar otras dependiendo del caso.

Referencias

1. Ameen, N., Tarhini, A., Reppel, A., Anand, A.: Customer experiences in the age of artificial intelligence. Computers in Human Behavior, vol. 114, pp. 106548 (2021) doi: 10.1016/j.chb.2020.106548
2. Aslani, M., Alesheikh, A.: Site selection for small gas stations using gis. Scientific Research and Essays, vol. 6, pp. 1361–3171 (08 2011)
3. Couronné, R., Probst, P., Boulesteix, A.-L.: Random forest versus logistic regression: a large-scale benchmark experiment. BMC Bioinformatics, vol. 19, no. 1, pp. 270 (July 2018) doi: 10.1186/s12859-018-2264-5
4. Ergün, M., Uygucgil, H., atalık, O.: Creating a geodemographic classification model within geo-marketing: the case of eskişehir province. Bulletin of Geography. Socio-economic Series, pp. 45–61 (03 2020) doi: 10.2478/bog-2020-0003
5. Global, M.: The state of ai in 2022—and a half decade in review (2022), https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2022-and-a-half-decade-in-review
6. HANG, L., TUAN, B., MANH, V., CHI, N., BINH, B., DIEP, T.: Building a geo-demographic segmentation model: the case of hanoi city, vietnam (01 2021) doi: 10.2991/aebmr.k.211119.048
7. Le, H.: Literature review on diversification strategy, enterprise core competence and enterprise performance. American Journal of Industrial and Business Management, vol. 09, pp. 91–108 (01 2019) doi: 10.4236/ajibm.2019.91008
8. Li, D.-C., Lin, Y.-S., Huang, Y.-C.: Constructing marketing decision support systems using data diffusion technology: A case study of gas station diversification. Expert Syst. Appl., vol. 36, pp. 2525–2533 (03 2009) doi: 10.1016/j.eswa.2008.01.065
9. Marques, R., de Paula Ferreira, W., Nassif, G., Armellini, F., Dungen, J., de Santa-Eulalia, L. A.: Exploring the application of iot in the service station business. IFAC-PapersOnLine, vol. 54, no. 1, pp. 402–407 (2021) doi: 10.1016/j.ifacol.2021.08.163
10. PetroIntellingence: Fotografía del sector gasolinero en mexico (2021), https://petrointelligence.com/Contribuciones/fotografia-sector-feb2021.pdf
11. Statista: Deployment status of data science and machine learning in organizations worldwide as of 2019, by function. Graph (September 9 2019), https://www.statista.com/statistics/1053561/data- science-machine-learning-deployment-by-function/, retrieved September 06, 2023
12. Wirth, R., Hipp, J.: Crisp-dm: Towards a standard process model for data mining. Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining, (01 2000)

ISSN 1870-4069

Agentes autónomos descentralizados: Convergencia de grandes modelos de lenguaje y Blockchain para la toma de decisiones automatizadas, auditables y colaborativas

Jorge Polanco Roque

Instituto Tecnológico y de Estudios Superiores de Monterrey, México

jorge.polanco@tec.mx

Resumen. La conjunción de la inteligencia artificial (IA), especialmente a partir de modelos de lenguaje de gran tamaño (LLMs), y la tecnología de cadena de bloques (blockchain) está impulsando una transformación profunda de las estructuras de gobernanza y de los procesos de toma de decisiones en entornos distribuidos. En este contexto, surgen los Agentes Autónomos Descentralizados (AADs) como unidades de software capaces de razonar sobre datos registrados en contratos inteligentes, coordinarse con otros actores de la red y desencadenar acciones tanto on-chain como off-chain sin requerir intermediarios centralizados. Este trabajo presenta una propuesta conceptual para el desarrollo de AADs, articulando los componentes tecnológicos clave —infraestructura en la nube, interacción con redes blockchain, diseño de contratos inteligentes y oráculos— y su integración con modelos de IA que dotan a los agentes de capacidades cognitivas avanzadas. Se incluye un estudio de caso sobre votaciones descentralizadas, así como un marco preliminar de validación que aborda vulnerabilidades técnicas, riesgos de seguridad algorítmica y amenazas emergentes como ataques por inyección de prompts. Finalmente, se discuten los desafíos regulatorios y éticos, y se trazan líneas de investigación futura en torno a la gobernanza automatizada y transparente.

Palabras clave: Agentes autónomos descentralizados, blockchain, inteligencia artificial, modelos de lenguaje, smart contracts, validación de seguridad, auditoría algorítmica, prompt injection, gobernanza descentralizada, oráculos, web 3.0.

Decentralized Autonomous Agents: Convergence of Large Language Models and Blockchain for Automated, Auditable, and Collaborative Decision-Making

Abstract. The conjunction of artificial intelligence (AI), especially from large language models (LLMs), and blockchain technology is driving a profound transformation of governance structures and decision-making processes in distributed environments. In this context, decentralized autonomous agents (DAAs) emerge as software units capable of reasoning on data recorded in smart

contracts, coordinating with other network actors, and triggering both on-chain and off-chain actions without requiring centralized intermediaries. This paper presents a conceptual proposal for the development of DAAs, articulating the key technological components—cloud infrastructure, interaction with blockchain networks, smart contract and oracle design—and their integration with AI models that equip agents with advanced cognitive capabilities. A case study on decentralized voting is included, as well as a preliminary validation framework that addresses technical vulnerabilities, algorithmic security risks, and emerging threats such as prompt injection attacks. Finally, regulatory and ethical challenges are discussed, and future research directions for automated and transparent governance are outlined.

Keywords: Decentralized autonomous agents, blockchain, artificial intelligence, language models, smart contracts, security validation, algorithmic auditing, prompt injection, decentralized governance, oracles, web 3.0.

1. Introducción

La proliferación de tecnologías como la IA y blockchain ha posibilitado la construcción de ecosistemas digitales con un alto grado de autonomía y fiabilidad. La IA, impulsada por métodos de aprendizaje profundo y de aprendizaje automático tradicional, ha incrementado sus capacidades gracias a los denominados modelos de lenguaje de gran tamaño, que permiten extraer patrones y razonar de manera contextual a partir de un vasto conjunto de datos (Wang et al., 2019). Al mismo tiempo, la tecnología blockchain ofrece un entorno distribuido, inmutable y trazable para la ejecución de transacciones y el despliegue de aplicaciones descentralizadas, lo que reduce la dependencia de entidades centrales e incrementa la confianza y la transparencia (Nakamoto, 2008; Tapscott & Tapscott, 2016).

En la convergencia de estas dos corrientes tecnológicas se sitúan los Agentes Autónomos Descentralizados. Dichos agentes adquieren un rol activo en la red, recabando información on-chain y off-chain para procesarla mediante diversos tipos de algoritmos—particularmente, mediante LLMs— y emitir decisiones o recomendaciones que pueden plasmarse directamente en contratos inteligentes. Esto abre un abanico de oportunidades en la gobernanza distribuida, la gestión de cadenas de suministro, el arbitraje de disputas, la detección de fraude y la orquestación de recursos en organizaciones virtuales. Al mismo tiempo, surgen importantes preguntas acerca de la escalabilidad de estas soluciones, la responsabilidad legal en entornos completamente descentralizados y la forma en que se garantizará la privacidad de los datos personales (Yaga et al., 2018).

Este artículo se estructura en nueve secciones para analizar la convergencia entre inteligencia artificial y blockchain en la construcción de Agentes Autónomos Descentralizados. Tras esta introducción, la segunda sección expone los fundamentos de ambas tecnologías, con énfasis en los modelos de lenguaje de gran tamaño y su papel en la toma de decisiones en entornos distribuidos.

En la tercera sección, se presenta un marco conceptual para el desarrollo de AADs, detallando la infraestructura necesaria, la conexión con blockchain mediante proveedores como Infura, la gobernanza descentralizada y la integración con oráculos.

La cuarta sección introduce un estudio de caso sobre votaciones descentralizadas, donde un AAD con capacidades cognitivas avanzadas supervisa la emisión de votos y detecta irregularidades.

La quinta sección explora mejoras tecnológicas y propuestas algorítmicas para optimizar el desempeño de los AADs, incluyendo mecanismos de consenso basados en reputación, Zero-Knowledge Proofs y estrategias de escalabilidad mediante sharding temático. En la sexta sección, se analizan los desafíos normativos y éticos, con un enfoque en la responsabilidad legal de las decisiones automatizadas y su compatibilidad con marcos regulatorios.

La séptima sección introduce un marco preliminar de validación para Agentes Autónomos Descentralizados, donde se analizan vulnerabilidades técnicas asociadas a contratos inteligentes, riesgos derivados del comportamiento de modelos de lenguaje de gran tamaño, y posibles vectores de ataque como la manipulación de prompts. Se discuten estrategias de mitigación basadas en verificación formal, auditoría algorítmica, mecanismos híbridos con supervisión humana, y protección frente a inyecciones de instrucciones maliciosas en el flujo de interacción entre la IA y la blockchain. Finalmente, la octava sección presenta conclusiones y futuras líneas de investigación, abordando el impacto potencial de los AADs en la gobernanza descentralizada y la economía digital.

Este trabajo se inscribe dentro del enfoque de conceptualización técnica, con el objetivo de proponer un marco de referencia estructurado para el diseño e implementación de Agentes Autónomos Descentralizados. A diferencia de trabajos empíricos que validan implementaciones específicas, aquí se busca integrar elementos tecnológicos, arquitectónicos y normativos en una visión coherente que permita guiar desarrollos futuros. El objeto de estudio son los AADs entendidos como entidades de software capaces de ejercer agencia en entornos descentralizados, y el problema abordado es la ausencia de un marco articulado que contemple tanto sus capacidades técnicas como sus implicaciones en sistemas de gobernanza distribuida.

La propuesta se sustenta en antecedentes de investigación sobre modelos de lenguaje (Brown et al., 2020; Weidinger et al., 2022), infraestructuras blockchain (Wood, 2014; Wang et al., 2019) y estructuras de gobernanza algorítmica (Hassan & De Filippi, 2021), contribuyendo a articular una síntesis que permita mapear desafíos técnicos, éticosy regulatorios.

2. Fundamentos de Blockchain, votaciones descentralizadas e IA

La tecnología blockchain se fundamenta en un registro inmutable y distribuido, donde los nodos validan transacciones mediante protocolos de consenso criptográficos. Este paradigma, introducido con Bitcoin (Nakamoto, 2008), evolucionó con redes como Ethereum, que incorporan contratos inteligentes para la ejecución programática de transacciones (Buterin, 2014). Aunque el consenso inicial se basaba en prueba de trabajo (Proof-of-Work), alternativas como la prueba de participación (Proof-of-Stake) han mejorado la escalabilidad y reducido el consumo energético (Wood, 2014).

No obstante, los costos de transacción y la latencia han impulsado soluciones de segunda capa, como sidechains y rollups, que optimizan la eficiencia de la red principal (Wang et al., 2019).

Paralelamente, la inteligencia artificial ha avanzado desde modelos de aprendizaje automático tradicionales hasta arquitecturas neuronales profundas capaces de procesar grandes volúmenes de datos. En este contexto, los modelos de lenguaje de gran tamaño han demostrado un desempeño sobresaliente en comprensión y generación de texto, síntesis de información y razonamiento contextual (Zhang et al., 2019). Estas arquitecturas —como GPT-4— se entrenan con conjuntos masivos de datos en infraestructuras escalables y pueden ajustarse para tareas específicas de todo tipo.

La convergencia entre IA y blockchain radica en la capacidad de los modelos de lenguaje para automatizar y optimizar procesos en redes descentralizadas, mientras la blockchain garantiza transparencia, auditabilidad e integridad de los datos (Tapscott & Tapscott, 2016). Para lograr esta integración, parte del procesamiento de IA se delega a infraestructuras en la nube o entornos de computación distribuida, mientras la blockchain registra los resultados esenciales, asegurando inmutabilidad y consenso.

Un campo donde esta integración cobra especial relevancia es el de las votaciones descentralizadas. La inmutabilidad y trazabilidad de la blockchain permiten un registro confiable de cada voto y garantizan la transparencia en los resultados.
Sin embargo, la incorporación de agentes basados en modelos de lenguaje introduce una capa de interpretación y análisis que amplía las capacidades de supervisión y detección de irregularidades en el proceso electoral de una “DAO”, por ejemplo.

Las DAOs (Organizaciones Autónomas Descentralizadas) son estructuras de gobernanza basadas en contratos inteligentes, donde las decisiones se toman colectivamente mediante un sistema de votación con tokens. Una de las primeras experiencias de gobernanza algorítmica mediante DAOs fue documentada por DuPont (2019) en su estudio histórico sobre 'The DAO', ilustrando los riesgos técnicos y organizacionales que pueden surgir cuando sistemas autónomos gestionan recursos colectivos.

Su transparencia e inmutabilidad han permitido su adopción en sectores como finanzas descentralizadas (DeFi), inversión colectiva y gestión de comunidades, eliminando intermediarios y fortaleciendo la autonomía organizativa.

En este contexto, los LLM pueden actuar como supervisores inteligentes, analizando dinámicas de participación y detectando anomalías en tiempo real. Un modelo entrenado y/o contextualizado en gobernanza descentralizada puede evaluar si una propuesta contradice decisiones previas, identificar estrategias de manipulación por grupos minoritarios o detectar patrones de votación atípicos. Además, estos agentes pueden generar reportes en lenguaje natural, facilitando la toma de decisiones informadas por parte de la comunidad.

Más aún, los LLM pueden desempeñar un papel activo en la votación, operando como representantes algorítmicos dentro de una DAO. Un agente de este tipo podría analizar el historial de gobernanza, evaluar propuestas según criterios predefinidos y emitir votos en representación de usuarios que deleguen su decisión en la IA. Esto resulta particularmente útil en entornos donde la toma de decisiones requiere análisis técnico o económico complejo, permitiendo que un Agente Autónomo Descentralizado vote con base en información estructurada y en tiempo real.

Al combinar el razonamiento avanzado de los LLM con la seguridad y descentralización de la blockchain, es posible construir sistemas de gobernanza más eficientes y autónomos. En lugar de depender exclusivamente de reglas predefinidas o de la intervención manual de los votantes, las DAOs pueden evolucionar hacia modelos

de decisión más dinámicos, donde la IA no solo detecte irregularidades, sino que también proponga estrategias de mitigación, facilite la deliberación colectiva y participe activamente en la toma de decisiones.

Además, la participación de estos agentes refuerza la confianza en los procesos electorales, ya que su supervisión sigue reglas codificadas en contratos inteligentes, auditable por la comunidad y sin intervención de una autoridad central. Esto no solo mejora la resistencia de las votaciones frente a manipulaciones, sino que también permite la adaptación dinámica de las reglas de gobernanza conforme evoluciona la comunidad, consolidando así un modelo más robusto y participativo.

La integración de modelos de lenguaje de gran tamaño con blockchain redefine la gobernanza descentralizada al combinar automatización, transparencia y análisis avanzado en la toma de decisiones. Esta sinergia no solo fortalece la resiliencia de los sistemas de votación descentralizados, sino que también permite la evolución hacia modelos de gobernanza más dinámicos, eficientes y adaptativos, reduciendo riesgos de manipulación y mejorando la toma de decisiones en comunidades autónomas.

3. Arquitectura de los agentes autónomos descentralizados

El concepto de Agentes Autónomos Descentralizados surge como un intento de unificar la lógica de los agentes autónomos de IA y la ejecución segura en blockchain en un único marco de referencia. La necesidad de integrar inteligencia autónoma con contratos autoejecutables en registros distribuidos ha sido explorada por autores como Christidis y Devetsikiotis (2016), quienes sentaron las bases para la fusión de blockchain, contratos inteligentes y agentes autónomos en aplicaciones como IoT. Más recientemente, Kuznetsov et al. (2024) analizaron los retos vinculados a la integración de IA y tecnología blockchain, haciendo énfasis en los desafíos de seguridad, escalabilidad y confiabilidad en contextos descentralizados.

El surgimiento de Agentes Autónomos Descentralizados debe entenderse como un paso evolutivo posterior al desarrollo de Organizaciones Autónomas Descentralizadas y a los primeros experimentos de gobernanza algorítmica (DuPont, 2017; Wright & De Filippi, 2015). Mientras que las DAOs delegan la ejecución de acuerdos a contratos inteligentes, los AADs extienden esta capacidad mediante la incorporación explícita de mecanismos de razonamiento autónomo basados en IA, aproximándose a un modelo de agencia algorítmica distribuida. Esta visión se alinea con los lineamientos propuestos por Kuznetsov et al. (2024), quienes describen la arquitectura de sistemas híbridos donde la autonomía algorítmica y la inmutabilidad del registro distribuido permiten una toma de decisiones automatizada, segura y verificable.

Un AAD se concibe como una entidad de software que posee llaves criptográficas para firmar transacciones, extrae datos tanto de la cadena (on-chain) como de fuentes externas (off-chain), y razona sobre dichos datos mediante algoritmos de inteligencia artificial, con el objetivo de desplegar acciones automatizadas. En términos de arquitectura, la infraestructura para un AAD puede organizarse en capas que facilitan su diseño y despliegue:

- En la **Capa de Datos y Conectividad**, el AAD se vincula con proveedores de nodos, como Infura, para interactuar con la red principal sin necesidad de mantener un nodo completo, lo que reduce la complejidad de configuración y

mantenimiento. Además, esta capa integra servicios de almacenamiento como InterPlanetary File System (IPFS) o bases de datos en la nube para manejar datos masivos que no se almacenan directamente en la blockchain. El AAD también establece enlaces con oráculos, que suministran información externa, por ejemplo, datos de mercado o identidad digital, y permiten la verificación de eventos del mundo real.

- En la Capa de Inteligencia Artificial, se emplean modelos de lenguaje de gran tamaño para procesar y razonar sobre grandes volúmenes de información. Estos modelos se entrenan en infraestructuras de computación escalables (GPU, TPU) y luego se implementan en servidores o contenedores que puedan comunicarse con la capa on-chain mediante interfaces de programación. El LLM puede especializarse en la detección de fraude, el análisis de transacciones financieras, la clasificación de propuestas en la DAO o cualquier otra tarea requerida por la organización descentralizada.
- En la Capa de Blockchain y Contratos Inteligentes, se definen las reglas de operación de la red y las condiciones en las que el AAD puede tomar acciones específicas. Los contratos inteligentes, escritos en lenguajes como Solidity (en el caso de Ethereum), recogen la lógica necesaria para transferir activos, restringir comportamientos indebidos y gestionar eventos relevantes para el AAD. De esta manera, la validación de cada acción del agente se registra en la red, generando transparencia y facilitando el escrutinio público o privado.
- En la Capa de Gobernanza Descentralizada, la comunidad o los participantes que ostenten tokens de gobernanza votan las actualizaciones de políticas, los parámetros de IA y las potenciales sanciones a comportamientos maliciosos. El AAD puede tener derecho de voto si la comunidad así lo dispone, o puede asumir un papel de auditor a fin de identificar anomalías y proponer sanciones. Este mecanismo de participación abierta y verificable crea incentivos para la contribución honesta y dificulta la manipulación por parte de uno o pocos actores.
- Finalmente, en la Capa de Interfaz y Aplicaciones Híbridas, los usuarios finales, desarrolladores o empresas interactúan con el AAD y con la red. Esta capa puede incluir tableros de control, formularios de votación o servicios de suscripción a eventos. Combina tecnologías web 2.0 tradicionales (por ejemplo, servidores en AWS, front-ends JavaScript) con el acceso a la blockchain a travéss de librerías como web3.js o ethers.js, con la finalidad de presentar la información de la forma más accesible posible.

Esta arquitectura escalable y modular se sustenta en la sinergia entre la nube y la cadena de bloques, aprovechando la potencia de cómputo fuera de la cadena para entrenar y ejecutar la IA, mientras la blockchain garantiza la inmutabilidad, la trazabilidad y la gobernanza colectiva (Zyskind et al., 2015).

4. Caso de uso: Votaciones descentralizadas asistidas por LLMs

Un Agente Autónomo Descentralizado potenciado por un modelo de lenguaje de gran tamaño actúa como un sistema de monitoreo dinámico dentro de un proceso de

votación, contrastando las dinámicas actuales de participación con distribuciones históricas modeladas estadísticamente. A partir de este análisis, el AAD puede identificar patrones atípicos o desviaciones significativas y reportarlas formalmente mediante alertas registradas on-chain, desencadenando mecanismos de auditoría automática o intervención comunitaria. Este enfoque permite incorporar capacidades de supervisión avanzadas en estructuras de gobernanza descentralizada, siguiendo la lógica de resiliencia algorítmica delineada por Christidis y Devetsikiotis (2016).

La metodología para implementar este caso incluye la creación de un contrato inteligente que dicta el plazo de la votación y almacena los resultados de forma cifrada. El AAD recibe notificaciones en tiempo real de las transacciones vinculadas a la votación, gracias a la conexión con un nodo Ethereum vía Infura, y a un oráculo que extrae metadatos sobre los votantes. El LLM analiza la coherencia de cada voto respecto a la base histórica de comportamientos normales, enviando alertas al contrato inteligente cuando encuentra discrepancias o indicios de manipulación. En ese momento, la DAO puede suspender temporalmente la votación y requerir aclaraciones o auditorías, implicando nuevamente al LLM para evaluar evidencia adicional.

Los resultados de estas pruebas pudieran demostrar que la participación de un AAD supervisando la votación potencia la confianza de la comunidad, al brindar detección temprana de fraude y total transparencia de las acciones del agente, auditables en la cadena de bloques. No obstante, la experimentación pudiera señalar que el incremento del número de votos y de transacciones genera un aumento significativo de costos de gas, lo cual motiva la adopción de soluciones en sidechains o en capas de escalabilidad para gestionar la afluencia masiva de participantes de AAD.

5. Propuestas de mejoras tecnológicas y algorítmicas

La implementación práctica de Agentes Autónomos Descentralizados que emplean modelos de lenguaje de gran tamaño enfrenta retos en confiabilidad, privacidad y escalabilidad. Para abordarlos, existen diversas estrategias que permiten llevar estos agentes a entornos de producción de manera más segura y eficiente.

Un primer paso es ampliar los mecanismos de consenso más allá de la prueba de participación (Proof-of-Stake), incorporando métricas de reputación o confiabilidad generadas por la IA. Este enfoque requiere que cada nodo aporte un historial de comportamiento (p. ej., participación en votaciones, precisión en la detección de fraude o calidad de datos en procesos de aprendizaje federado). El LLM evalúa este historial y genera una puntuación de reputación dinámica, la cual se integra en el consenso para penalizar nodos maliciosos y recompensar contribuciones positivas.

En cuanto a la privacidad, el uso de cifrado homomórfico y Zero-Knowledge Proofs (ZKPs) (Zyskind et al., 2015) permite que los AADs trabajen con datos sensibles sin desencriptarlos. Por ejemplo, el agente podría validar la legitimidad de una transacción o marcarla como fraudulenta mediante pruebas criptográficas que demuestren la corrección de su análisis sin revelar su contenido. Aunque esta aproximación eleva la complejidad y el costo computacional, resulta esencial en aplicaciones donde la confidencialidad es prioritaria.

Para escalar estas operaciones, la fragmentación de la red en subcadenas especializadas o sharding temático (Wang et al., 2019) distribuye la carga de cómputo.

Una subcadena puede dedicarse al entrenamiento y la ejecución del modelo IA, mientras la cadena principal registra únicamente el hash de los resultados y las actualizaciones clave. Esto aligera la congestión y reduce los costos de transacción, pero exige un protocolo de interoperabilidad multicadena, que permita a los AADs leer y escribir en las distintas subcadenas y, a su vez, publicar en la red principal los resultados finales de su procesamiento.

Un ejemplo práctico de este flujo incluye la fase de entrenamiento y evaluación en una subcadena, seguida de la generación de una prueba criptográfica que describe los resultados, y finaliza con el registro en la cadena principal, donde un contrato inteligente solo acepta los resultados acompañados de dicha prueba. En el caso de votaciones, la misma lógica puede garantizar la validez de las decisiones tomadas por el LLM, exigiendo evidencias de su análisis on-chain para salvaguardar la transparencia del proceso.

La auditoría algorítmica constituye otro pilar fundamental. La DAO puede configurar contratos inteligentes de auditoría que obliguen al AAD a proporcionar explicaciones resumidas de las decisiones más relevantes, ya sea la suspensión de un voto masivo o la detección de comportamientos anómalos. Si la comunidad considera insuficientes las justificaciones, puede votar sanciones o la revocación de los privilegios del agente. Este mecanismo no solo promueve la responsabilidad del AAD, sino que también aporta claridad y confianza a los participantes.

La integración técnica entre IA y blockchain requiere librerías o APIs que faciliten la comunicación bidireccional. La IA debe consultar datos on-chain con baja latencia, lo que implica la adopción de servicios como Infura o nodos locales configurados de forma óptima. Por otra parte, los resultados de la IA tienen que transformarse en transacciones firmadas por el propio agente, para lo que se requiere un módulo criptográfico de alta seguridad, ya sea implementado en enclaves de hardware o en contenedores con medidas de protección reforzadas.

En conjunto, la incorporación de mecanismos de reputación IA, cifrado homomórfico, ZKPs y sharding temático permite construir AADs escalables y seguros. La auditoría algorítmica, sumada a la gobernanza descentralizada, refuerza la confianza en la toma de decisiones automatizada e impulsa nuevos modelos de colaboración y gobernanza en la economía digital.

6. Perspectiva normativa y ética

La creciente autonomía de los Agentes Autónomos Descentralizados conlleva desafíos significativos en el ámbito legal y ético, pues redefine la manera en que se asigna la responsabilidad, se salvaguarda la privacidad y se gestionan los sesgos inherentes a los modelos de inteligencia artificial.

Un primer punto de tensión surge en torno a la asignación de responsabilidad cuando las decisiones son tomadas por un AAD que, en sentido estricto, no se halla bajo el dominio de una autoridad central. Este fenómeno se enmarca dentro de lo que Wright y De Filippi (2015) conceptualizaron como 'Lex Cryptographia', es decir, sistemas donde el código y las reglas algorítmicas sustituyen —al menos parcialmente— a las formas tradicionales de regulación legal. Tal como expusieron Wright y De Filippi (2015) en su análisis sobre Lex Cryptographia, la programación de reglas

autoejecutables en entornos distribuidos redefine las nociones clásicas de contrato, autoridad y responsabilidad, trasladándolas al dominio del código. Los AADs, al operar bajo estos principios, intensifican los dilemas regulatorios al eliminar progresivamente los intermediarios humanos que tradicionalmente serían los sujetos de derecho.

Los marcos legales vigentes, concebidos para organizaciones con personalidad jurídica y jerarquías definidas, carecen de pautas claras para determinar quién responde en caso de que las acciones de un AAD ocasionen daños o vulneren derechos. Las DAO pueden implementar mecanismos internos de resolución de disputas y votaciones correctivas, pero el valor legal de tales resoluciones ante las jurisdicciones tradicionales es incierto (Tapscott & Tapscott, 2016). En la práctica, se abre la posibilidad de que los tribunales requieran identificar a los desarrolladores o promotores de la DAO, extendiendo la responsabilidad a quienes hayan facilitado la puesta en marcha del agente autónomo. Esto contrasta con la intención de descentralizar la gobernanza y reducir la dependencia de figuras centrales, generando un vacío regulatorio que exige la adaptación de los marcos normativos.

Un segundo punto delicado está relacionado con la protección de datos personales y la privacidad. La inmutabilidad de la cadena de bloques desafía la aplicación de leyes como el Reglamento General de Protección de Datos (RGPD) en Europa, que establece derechos como el de supresión o rectificación de la información (Yaga et al., 2018).

Si bien la cadena de bloques garantiza la trazabilidad y la verificación de registros, esta característica choca con la posibilidad de eliminar completamente un dato o revertirlo de manera forense. Para abordar estas exigencias, se investigan técnicas criptográficas de "borrado selectivo" o chameleon hashes, que permiten enmendar transacciones bajo consenso de la red, aunque su adopción masiva conlleva profundos cambios en el diseño de los protocolos y puede diluir, en cierta medida, el principio de inmutabilidad. Otra línea de desarrollo se centra en la disociación y el cifrado homomórfico, para que el contenido sensible nunca se exponga en texto plano, aunque estos métodos suelen incrementar la complejidad de la red y los costos de cómputo.

En cuanto a la ética de la IA, la naturaleza de los modelos de lenguaje de gran tamaño implica el riesgo de reproducir sesgos de los datos de entrenamiento, lo que puede derivar en discriminaciones o falsos positivos y perjudicar a ciertos sectores de la población. Los AADs que incorporen IA en su lógica decisoria enfrentan la necesidad de asegurar que dichos modelos cumplan criterios de justicia y equidad. Una manera de mitigar estos riesgos es la auditoría algorítmica, en la que se somete el comportamiento del modelo a evaluaciones continuas por parte de la comunidad o de equipos especializados. Sin embargo, la responsabilidad de corregir dichos sesgos recae en un colectivo difuso: los propietarios de tokens de gobernanza, los desarrolladores que implementan las actualizaciones y, en última instancia, la comunidad de nodos.

Si el AAD asume funciones críticas —por ejemplo, supervisar transacciones financieras o decidir políticas de un fondo de inversión—, estos sesgos podrían tener consecuencias sociales y económicas de gran alcance.

La transparencia de las operaciones constituye otro factor clave en la legitimidad de los AADs. La capacidad de un agente para firmar transacciones y desencadenar acciones en la blockchain sin una supervisión humana explícita puede generar preocupación si su lógica interna se percibe como una "caja negra". Para enfrentar esta inquietud, las DAO pueden exigir que el agente ofrezca descripciones mínimamente explicables de su razonamiento en los casos con mayor impacto. Herramientas de

explicabilidad (por ejemplo, mecanismos de extracción de reglas o resúmenes de atención en modelos de lenguaje) permiten a los participantes comprender las motivaciones del AAD y, de ser necesario, responder con votaciones que limiten su autonomía o exijan reentrenamientos. No obstante, esta aspiración de transparencia debe equilibrarse con la protección de datos y la propiedad intelectual, pues exponer todos los detalles del modelo podría revelar información confidencial o estratégica.

Otro eje de tensión concierne a la compatibilidad con los sistemas legales existentes. Muchas jurisdicciones no reconocen la figura de un agente autónomo sin personalidad jurídica, lo que complica su participación en contratos legalmente vinculantes. La figura del "contrato inteligente" en sí misma plantea cuestiones acerca de la ejecutabilidad legal de los acuerdos, especialmente si las cláusulas se ejecutan de manera automática sin posibilidad de reclamación ante un órgano jurisdiccional. Algunas iniciativas exploran la creación de organizaciones híbridas, donde la DAO se registra bajo formas legales tradicionales, lo que habilitaría una responsabilidad más clara en caso de disputas o demandas.

Finalmente, la responsabilidad en la actualización y el mantenimiento de estos agentes abre interrogantes relacionados con la continuidad de su desarrollo y la legitimidad de los cambios introducidos. Un AAD basado en IA no es estático: sus modelos pueden requerir reentrenamientos y/o contextualizaciones nuevas, ajustes de parámetros o cambios en el set de datos que se considera confiable. La DAO debe definir procesos transparentes para aprobar cada actualización y asegurar que los equipos técnicos cumplan con requisitos de calidad y validación. Si los procedimientos resultan ambiguos o demasiado restrictivos, se arriesga la obsolescencia del agente; en cambio, si se otorga un margen de maniobra excesivo, la gobernanza colectiva puede perder control en la gobernanza.

En conclusión, la autonomía de los AADs y su integración con modelos de inteligencia artificial generan un conjunto de desafíos que no se limitan al plano técnico, sino que se extienden al terreno legal, ético y socioeconómico. La posibilidad de que la DAO sirva como espacio de resolución de disputas y vigilancia colectiva ofrece oportunidades para repensar la responsabilidad y la equidad, pero también introduce incertidumbre sobre la eficacia de tales mecanismos fuera del ecosistema blockchain. En la medida en que las iniciativas tecnológicas avancen, resultará crucial articular diálogos con los reguladores y los actores de la sociedad civil, a fin de diseñar marcos normativos flexibles que reconozcan la naturaleza distribuida de estos agentes y, al mismo tiempo, protejan los derechos fundamentales y promuevan la innovación responsable.

7. Validación de la propuesta: Análisis de vulnerabilidades y consideraciones de seguridad

La validación rigurosa de los Agentes Autónomos Descentralizados resulta indispensable para garantizar su viabilidad en entornos de gobernanza descentralizada. Esta validación debe abordar tanto los riesgos inherentes a la infraestructura blockchain como las vulnerabilidades asociadas al uso de modelos de lenguaje de gran tamaño.

En primer lugar, la resiliencia de los contratos inteligentes que sustentan las operaciones de los AADs debe ser verificada mediante técnicas de verificación formal

y análisis estático de seguridad, utilizando herramientas como MythX o Slither (Feist, Grieco, & Groce, 2019). Estas técnicas permiten detectar vulnerabilidades comunes como reentrancy attacks, integer overflows o accesos indebidos, contribuyendo a mitigar riesgos sistémicos antes del despliegue.

En segundo término, el comportamiento del modelo de lenguaje debe someterse a evaluaciones de robustez adversarial (Goodfellow, Shlens, & Szegedy, 2015). La generación de ejemplos adversarios y la implementación de mecanismos de defensa como el adversarial training resultan esenciales para identificar debilidades en la inferencia del modelo, especialmente frente a inputs maliciosos diseñados para inducir comportamientos anómalos o sesgados.

Particular atención requiere la mitigación de ataques basados en ingeniería de prompts (Prompt Injection Attacks), donde un agente malicioso podría manipular las instrucciones internas del modelo para alterar su comportamiento previsto (Perez, Ranta, & Raffel, 2022; Carlini et al., 2021). Para reducir este riesgo, es necesario diseñar prompts robustos, implementar filtros de sanitización de entradas, emplear técnicas de prompt verification, y limitar las acciones críticas a aquellos outputs validados explícitamente mediante políticas de control (Weidinger et al., 2022).

Adicionalmente, se propone la incorporación de mecanismos de validación híbrida (Human-in-the-Loop, HITL) (Amershi et al., 2014), en los cuales decisiones críticas tomadas por el AAD sean auditadas por validadores humanos antes de su ejecución definitiva. Este enfoque híbrido refuerza la confiabilidad general del sistema, equilibrando eficiencia automatizada con supervisión consciente.

La consistencia en las dinámicas de gobernanza también debe ser validada mediante simulaciones de ataque de gobernanza (governance attack simulations), que modelen escenarios de manipulación de votaciones, colusión de nodos o concentración de poder mediante la acumulación de tokens (Hassan & De Filippi, 2021).

Finalmente, se enfatiza la necesidad de establecer un esquema de auditoría algorítmica continua (Brundage et al., 2020), en el cual los AADs deban proporcionar explicaciones auditables de sus acciones relevantes, registradas on-chain, permitiendo así una trazabilidad robusta y un control comunitario efectivo.

En conjunto, estos mecanismos conforman un marco de validación preliminar que refuerza la robustez, confiabilidad y legitimidad de los Agentes Autónomos Descentralizados en escenarios productivos, sentando las bases para su adopción escalable y su integración segura en ecosistemas de gobernanza distribuida.

8. Conclusiones y futuras líneas de investigación

La construcción de Agentes Autónomos Descentralizados con modelos de lenguaje de gran tamaño en redes blockchain configura un nuevo paradigma de gobernanza y automatización que trasciende la lógica habitual de sistemas centralizados. Al permitir la participación directa de la IA en la ejecución de contratos inteligentes y la toma de decisiones colectivas, se abren vías innovadoras para la gestión de recursos, la supervisión de procesos y la detección temprana de anomalías. Esta convergencia de IA y blockchain, no obstante, exige equilibrios técnicos, normativos y éticos que se han discutido a lo largo del presente documento.

Por un lado, la optimización técnica demanda tanto soluciones de escalabilidad (sharding temático, rollups, sidechains) como la adopción de técnicas criptográficas avanzadas (cifrado homomórfico, Zero-Knowledge Proofs) para proteger la privacidad y respaldar el procesamiento seguro de datos sensibles. A su vez, la construcción de módulos de reputación y mecanismos de auditoría algorítmica fortalece la confiabilidad y la transparencia de los AADs, evitando la dependencia de una autoridad central y promoviendo la participación de la comunidad en la validación de acciones y resultados.

En el plano legal y ético, la autonomía creciente de los AADs invita a redefinir los esquemas de responsabilidad y protección de derechos. Es esencial aclarar los alcances de la "personalidad virtual", la vinculación legal de contratos inteligentes y el tratamiento de datos almacenados en una cadena inmutable. Asimismo, mitigar sesgos en la IA y asegurar la explicabilidad de las decisiones se vuelve crítico para la confianza y la aceptación social de estos sistemas, sobre todo cuando desempeñan funciones sensibles o de alto impacto.

De cara al futuro, se vislumbran múltiples líneas de investigación y desarrollo:

1. **Arquitecturas Multicadena y Aprendizaje Federado**: Profundizar en protocolos que conecten diversas redes orientadas a tareas específicas (por ejemplo, entrenamiento de IA, almacenamiento masivo, ejecución de contratos de gobernanza), para escalar las operaciones y reducir costos.

2. **Modelos de Gobernanza Reputacional**: Diseñar mecanismos de consenso dinámicos basados en indicadores de contribución y fiabilidad, calculados por la IA a partir de historiales de comportamiento de los nodos.

3. **Explicabilidad y Auditoría Algorítmica**: Consolidar metodologías para que los AADs ofrezcan descripciones claras de sus procesos de inferencia y razonamiento, en especial ante decisiones conflictivas, y posibilitar la participación activa de la comunidad en la corrección de sesgos.

4. **Identidad Digital y Protección de Datos**: Integrar soluciones como Self-Sovereign Identity (SSI) y técnicas de borrado selectivo para conciliar la inmutabilidad de la cadena con las demandas normativas y el derecho al olvido.

5. **Interacción con la Economía Real**: Evaluar el rol de los AADs en sectores como logística, finanzas tradicionales, seguros o administración pública, y estudiar cómo se adaptan los procedimientos legales para acoger a estos agentes en la práctica.

No obstante, el desarrollo de Agentes Autónomos Descentralizados enfrenta limitaciones críticas que deben ser abordadas en investigaciones futuras. Entre ellas destacan la validación exhaustiva de su robustez frente a ataques adversarios (Perez et al., 2022; Carlini et al., 2021), el diseño de mecanismos de gobernanza adaptativa capaces de resistir dinámicas de colusión, y la necesidad de protocolos de interoperabilidad multicadena que aseguren su escalabilidad práctica. La consolidación de estos agentes requerirá asimismo marcos regulatorios flexibles que equilibren la innovación técnica con la protección de derechos fundamentales, en línea con

propuestas recientes sobre gobernanza algorítmica responsable (Hassan & De Filippi, 2021).

En síntesis, los AADs combinados con LLMs constituyen un avance significativo en la convergencia de la inteligencia artificial y la tecnología blockchain, al posibilitar la toma de decisiones autónoma bajo mecanismos de gobernanza colectiva y alta auditabilidad. La naturaleza descentralizada de estos entornos, unida a la creciente sofisticación de los agentes cognitivos, plantea retos y oportunidades que invitan a un debate multidisciplinario, donde confluyan la ingeniería, el derecho, la economía y la ética. Conforme se consoliden estas soluciones y se establezcan los marcos regulatorios apropiados, es probable que los AADs desempeñen un papel creciente en la configuración de la próxima generación de sistemas de gestión y organización social.

Referencias

1. Amershi, S., Cakmak, M., Knox, W.B.: Power to the People: The Role of Humans in Interactive Machine Learning. AI Magazine, 35(4), pp. 105–120 (2014) doi: 10.1609/aimag.v35i4.2513.
2. Bonneau, J., Miller, A., Clark, J.: SoK: Research Perspectives and Challenges for Bitcoin and Cryptocurrencies. In: IEEE Symposium on Security and Privacy, pp. 104–121 (2015) doi: 10.1109/SP.2015.14.
3. Brown, T.B., Mann, B., Ryder, N.: Language Models are Few-shot Learners. Advances in Neural Information Processing Systems, 33, pp. 1877–1901 (2020) doi: 10.48550/arXiv.2005.14165.
4. Brundage, M., Avin, S., Clark, J.: Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims (2020) doi: 10.48550/arXiv. 2004.07213.
5. Buterin, V.: A Next-generation Smart Contract and Decentralized Application Platform. Ethereum White Paper, 3, pp. 1–36 (2013)
6. Carlini, N., Tramer, F., Wallace, E.: Extracting Training Data from Large Language Models. In: USENIX Security Symposium, pp. 1–19 (2021) doi: 10.48550/arXiv. 2012.07805.
7. Christidis, K., Devetsikiotis, M.: Blockchains and Smart Contracts for the Internet of Things. IEEE Access, 4, pp. 2292–2303 (2016) doi:10.1109/ACCESS. 2016.2566339.
8. ConsenSys Diligence. MythX: Smart Contract Security Analysis Service. https://mythx.io. (2020)
9. DuPont, Q.: Experiments in Algorithmic Governance: A History and Ethnography of 'The DAO,' a Failed Decentralized Autonomous Organization. En M. Campbell-Verduyn (Ed.), Bitcoin and Beyond: Cryptocurrencies, Blockchains, and Global Governance pp. 157–177 (2017) doi: 10.4324/9781315211909-8.
10. Feist, J., Grieco, G., Groce, A.: Slither: A Static Analysis Framework for Smart Contracts. In: Proceedings of the 2nd International Workshop on Emerging Trends in Software Engineering for Blockchain, pp. 8–15 (2019) doi: 10.1109/ WETSEB.2019.00008.
11. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and Harnessing Adversarial Examples. pp. 1–15 (2015) doi: 10.48550/arXiv.1412.6572.

12. Hassan, S., De Filippi, P.: Decentralized Autonomous Organizations: Beyond the Hype. Internet Policy Review, 10(2) (2021) doi: 10.14763/2021.2.1556.
13. Kuznetsov, O., Sernani, P., Romeo, L.: On the Integration of Artificial Intelligence and Blockchain Technology: A Perspective About Security. In: IEEE Access, 12, pp. 3881–3897 (2024) doi: 10.1109/ACCESS.2023.3349019.
14. Nakamoto, S.: Bitcoin: A Peer-to-Peer Electronic Cash System. pp. 1–9 https://bitcoin.org/bitcoin.pdf (2008)
15. Perez, F., Ribeiro, I.: Ignore Previous Prompt: Attack Techniques for Language Models. pp. 1–21 (2022) doi: 10.48550/arXiv.2211.09527.
16. Tapscott, D., Tapscott, A.: Blockchain Revolution: How the Technology Behind Bitcoin and Other Cryptocurrencies is Changing the World. Penguin (2016)
17. Wang, S., Ouyang, L., Yuan, Y.: Blockchain-Enabled Smart Contracts: Architecture, Applications, and Future Trends. In: IEEE Transactions on Systems, Man, and Cybernetics: Systems, 49(11), pp. 2266–2277 (2019) doi: 10.1109/TSMC.2019.2895123.
18. Weidinger, L., Mellor, J., Rauh, M.: Ethical and Social Risks or Harm from Language Models (2021) doi: 10.48550/arXiv.2112.04359.
19. Wood, G.: Ethereum: A Secure Decentralised Generalised Transaction Ledger. Ethereum Project Yellow Paper, 151, pp. 1–32 (2014) https://ethereum.org/en/whitepaper/.
20. Wright, A., De Filippi, P.: Decentralized Blockchain Technology and the Rise of Lex Cryptographia. SSRN Electronic Journal, 34, pp. 41–52 (2015) doi: 10.2139/ssrn.2580664.
21. Yaga, D., Mell, P., Roby, N.: Blockchain Technology Overview. National Institute of Standards and Technology (2018) doi: 10.6028/NIST.IR.8202.
22. Zhang, Y., Kasahara, S., Shen, Y.: Smart Contract-Based Access Control for the Internet of Things. In: IEEE Internet of Things Journal, 6(2), pp. 1594–1605 (2019) doi: 10.1109/JIOT.2018.2847705.
23. Zyskind, G., Nathan, O., Pentland, A.: Decentralizing Privacy: Using Blockchain to Protect Personal Data. In: IEEE Security and Privacy Workshops, pp. 180–184 (2015) doi: 10.1109/SPW.2015.27.

ISSN 1870-4069

Low-Resource Hardware Performance Analysis for Real-Time Facial Recognition

Edgar Abidán Padilla Luis[1], Fernando Perez-Tellez[2], David Pinto[1]

Benemérita Universidad Autónoma de Puebla,
Faculty of Computer science,
Mexico

Technological University Dublin,
Faculty of Computing Digital and Data,
Ireland

edgar.abidan.pl@gmail.com, david.pinto@correo.buap.mx,
Fernando.PerezTellez@tudublin.ie

Abstract. Real-time facial recognition technology faces challenges on low-resource hardware due to processing and memory limitations. This study analyzes the performance of the Raspberry Pi 4 and 5 compared to a standard desktop computer, evaluating CPU usage, RAM, temperature, and latency across five different models. The results highlight the limitations and feasibility of these devices, providing a guide for selecting appropriate hardware based on performance and resources.

Keywords: Face recognition, real-time processing, low-resource hardware.

1 Introduction

Facial recognition is a biometric technology that identifies or verifies a person's identity by analyzing facial features. This process involves face detection in images, extraction of distinctive features, and comparison with a database [15]. However, it faces ethical challenges related to privacy and bias [15]. Additionally, technical issues such as variability in lighting conditions and facial expressions exist [14]. Its applications range from security and commerce to government and healthcare [8]. Currently, facial recognition is used as a real-time tool [5], [1] and [10]. Nevertheless, choosing a device with limited computational resources to integrate this technology is always challenging. Therefore, this article compares the performance of several facial recognition models between two widely used hardware platforms in real-time applications (Raspberry Pi 4 and Raspberry Pi 5) and a conventional desktop computer.

2 Related Work

Facial recognition is a widely studied task in the field of computer vision. There are numerous state-of-the-art articles that present facial recognition systems, which are implemented on various types of hardware. For example, in [19,6,10,25,7] we can find the implementation and analysis of facial recognition tasks on Jetson platforms. Another commonly used hardware is the Raspberry Pi, as shown in [13,1,19,14], this type of board is frequently used in facial recognition tasks due to its versatility and efficiency. In this paper, we used the Raspberry Pi 4 and Raspberry Pi 5 boards to conduct the experiments.

Dewantoro *et al* [6] conduct a performance comparison of a line-following robot using the Raspberry Pi and Jetson Nano as CPU controllers and find that the accuracy in the task of recognizing line paths is 96% when using the Raspberry Pi, while it is 98% when using the Jetson Nano.

Manni *et al* [11] performed a study at the "Smart Living Technologies" laboratory of the Institute for Microelectronics and Microsystems (IMM) in Lecce, Italy, to validate a proposed approach for real-time heart rate monitoring using various hardware platforms, including Raspberry Pi 4, Odroid N2+, and Jetson Nano. The experiments revealed that Raspberry Pi 4 exhibited the highest CPU consumption due to the lack of a GPU, impacting the execution time of the face detection block using the Dlib library. Additionally, it was observed that Raspberry Pi 4 showed slightly higher memory usage compared to the other platforms, although all displayed similar behaviors in terms of memory usage. Concerning power consumption, Raspberry Pi 4 recorded the lowest consumption, likely due to the absence of a cooling unit. In terms of pipeline accuracy, high accuracy was achieved with a resolution of 640 $\times$ 480 and a distance of 0.5 m across all evaluated platforms. However, Raspberry Pi 4 stood out for its inferior performance compared to Odroid N2+ and Jetson Nano, which demonstrated execution times comparable to a standard PC and slightly lower than a laptop.

Biglaru and Tang [4] evaluated the performance of various machine learning packages when running trained models on different edge hardware platforms. Latency, memory footprint, and energy consumption were compared across the AlexNet and SqueezeNet neural network models. MXNet performed well on MacBook, while TensorFlow showed good performance on FogNode. On Jetson TX2, PyTorch exhibited shorter inference times compared to TensorFlow on AlexNet. Caffe2 demonstrated efficiency in executing SqueezeNet on Raspberry Pi, albeit with memory limitations. It was identified that model loading takes more time than inference in some packages, indicating opportunities for edge optimization. Memory and energy consumption showed a varied trade-off among packages, with MXNet noted for energy efficiency and PyTorch for lower memory usage on Jetson TX2.

Baobaid *et al* [2] conducted a study on facial detection and recognition systems, comparing neural network-based and non-neural network-based algorithms. They found that neural network algorithms, like FaceNet, outperform non-neural ones in terms of accuracy. The performance was tested

on different hardware accelerators such as Raspberry Pi, Jetson Nano GPU, and GTX1060 GPU. Raspberry Pi showed limited performance, with a processing rate of less than 1 FPS for detection and an average of 1.5 FPS for recognition using neural network algorithms. However, its performance significantly improved with the use of hybrid accelerators like Intel Movidius. While GPUs proved superior in accuracy and processing time, FPGAs were noted for their attractiveness in power consumption and execution time, suggesting a future approach of heterogeneous systems combining these technologies.

Khan *et al* [9] carry out a comparative study of three facial recognition algorithms on multicore systems, evaluating their speed and accuracy by taking 13 samples per person. The system's compatibility across different machines, including low-spec configurations, was demonstrated, emphasizing the use of Raspberry Pi to eliminate platform dependencies. Results presented in tables and graphs indicated that the LBPH algorithm was the most accurate, achieving up to 90% accuracy but with higher time consumption due to its binary pattern computation. In contrast, the Fisher Face algorithm was the fastest but least accurate, reaching a maximum of 85% accuracy.

Although Jetson Nano models typically outperform Raspberry Pi in performance, they require more computational resources and are more expensive. Motivated by these differences, this study focuses exclusively on the popular Raspberry Pi 4 and Raspberry Pi 5 models.

3 Experiment Description

In the reviewed literature, facial recognition systems conduct experiments in highly controlled environments, using cameras that focus on faces at very short distances, as shown in these papers [19,2,1,26,13,9,22,12], Fig. 1 shows an example of a typical image used in these analyses. This setup allows facial recognition models to achieve precise results. However, in a real-world environment, this design is inadequate, as cameras are not always positioned at the same distance, and people's faces are not close to the lens. For this reason, we designed our experiment to emulate a real-world setting.

The experiment was conducted in an office environment. Fig. 2 shows a floor plan of the location where the experiment took place. The gray areas represent desk sections, while the white areas indicate open spaces where people can walk. The room has only one entrance and exit door. For the experiment, a camera was placed next to Desk Section A, pointing toward the door. Four individuals participated, each performing the following steps: entering the office through the door, walking from the door to the camera, passing in front of the camera while still within its range, moving toward Desk Section B, exiting the camera's range, and moving toward Desk Section C. Fig. 2 illustrates the path taken by the participants.

Among the participants, only two were included in the database with images taken at various distances within the camera's range. Fig. 3 shows some images from the database.

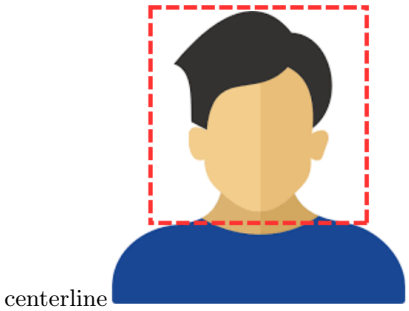


Fig. 1. Example of a typical image analyzed in state-of-the-art papers.

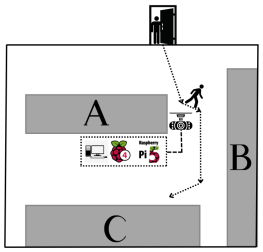
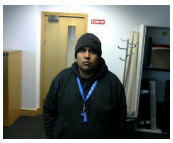
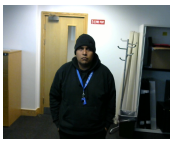


Fig. 2. In this experiment a person opens the door, walks in front of the webcam connected to one of the three devices used.

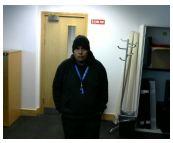
(a) Image 1 of the person 1.



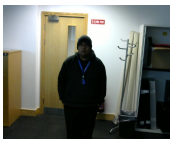
(b) Image 2 of the person 1.



(c) Image 3 of the person 1.



(d) Image 4 of the person 1.

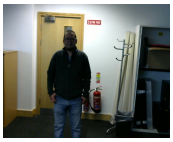


(e) Image 5 of the person 1.

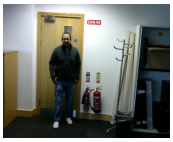
(f) Image 1 of the person 2.

(g) Image 2 of the person 2.

(h) Image 3 of the person 2.



(i) Image 4 of the person 2.



(j) Image 5 of the person 2.

Fig. 3. Images in the Database.

3.1 Methodology

We developed a system based on the methodology shown in Fig. 4. The system flow is as follows: a camera records in real time, capturing images with a size of

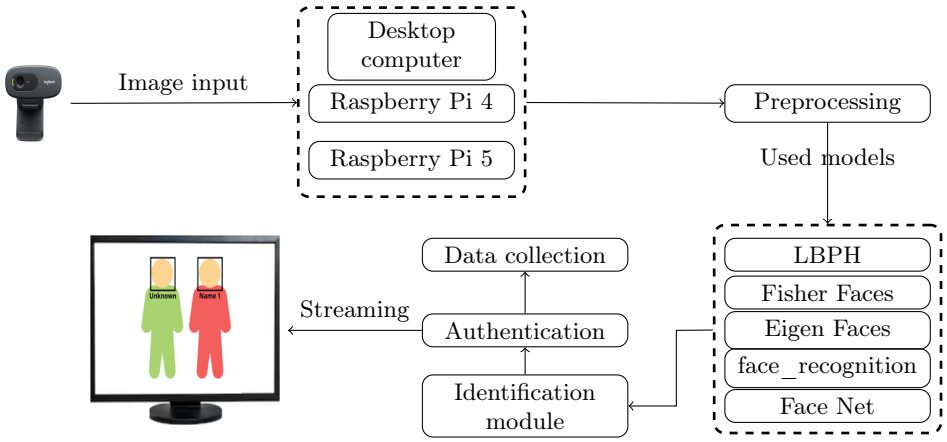


Fig. 4. Methodology used to evaluate performance on hardware with limited resources.

1280px × 720px. Subsequently, one of three devices is used: a desktop computer, a Raspberry Pi 4, or a Raspberry Pi 5. On the selected device, each image goes through a preprocessing module where it is converted to grayscale and resized to 20% of its original size. This resizing is done to reduce latency and enable real-time facial recognition.

Each of the devices runs five facial recognition models: Local Binary Patterns Histogram (LBPH) [12], Fisher Faces (FF) [3] , Eigen Faces (EF) [16], Face_recognition (FR) [7], and Face Net (FN) [13]. Each model is run separately, meaning the five models are not executed simultaneously. The models were previously trained to recognize two people. Each image is analyzed by the models in the identification module. Then, in the authentication module, if a person from the database matches an analyzed face, the name is assigned; otherwise, they are labeled as 'unknown'. Additionally, a rectangle is drawn around the detected face. Finally, the image with the results from the authentication module is displayed on a monitor. At the same time, another module stores the data for later analysis.

Hardware overview Below, we present the characteristics of the devices used to conduct this study.

Webcam The Logitech C270 HD Webcam is a budget-friendly webcam that offers clear HD 720p video calls, noise-canceling microphone, automatic light correction, and a wide field of view. It's easy to set up and use, making it a great choice for basic video calls and streaming.

(a) Configuration of the Raspberry Pi 4 with accessories.

(b) Single-board view of the Raspberry Pi 4.

Fig. 5. Overview of the Raspberry Pi 4 hardware configuration and its single-board computer.

Desktop Computer The workstation is a Dell OptiPlex 5050 with 32GB of RAM, an Intel Core i5-6500 processor, Mesa Intel HD Graphics 530, and 512GB of storage. It runs Ubuntu 22.04.4 LTS.

Raspberry Pi 4 The Raspberry Pi 4, a powerful single-board computer is shown in Fig.5b, running on the latest Bookworm OS, boasts a Broadcom BCM2711 SoC, quad-core Cortex-A72 CPU at 1.8GHz, and LPDDR4-3200 SDRAM memory of 8GB. For enhanced protection and improved thermal performance, the Raspberry Pi 4 can be equipped with a protective case and heatsinks. These additions safeguard the board from physical damage and ensure efficient heat dissipation during operation, extending its lifespan and maintaining stable performance as illustrated in Fig. 5a, especially when running demanding tasks.

Raspberry Pi 5 The Raspberry Pi 5 with Raspberry Pi Bookworm operating system is the latest and most powerful single-board computer from the Raspberry Pi Foundation this board is shown in Fig. 6b. The Raspberry Pi 5 boasts a powerful Broadcom BCM2712 quad-core ARM Cortex-A76 processor at 2.4 GHz, 8GB of LPDDR4X-4267 RAM, and a VideoCore VII GPU for superior performance. For further expansion, it includes a PCIe 2.0 x1 interface. The board is powered by a 5V/5A USB-C port and features a real-time clock. Finally, the Raspberry Pi 5 incorporates a power button for more intuitive device control. This device has been enhanced with a protective case, heat sinks, and a fan for better performance as illustrated in Fig. 6a.

4 Experimental Results

This section presents the results of the experiments, focusing on four key resources that devices consume and are critical for software implementation: CPU usage, RAM consumption during task execution, device temperature, and latency in real-time model execution.

(a) Configuration of the Raspberry Pi 5 with accessories.

(b) Single-board view of the Raspberry Pi 5.

Fig. 6. Overview of the Raspberry Pi 5 hardware configuration and its single-board computer.

Tab. 1. Latency on Desktop, Raspberry Pi 4 and Raspberry Pi 5.

Model	Desktop				Raspberry Pi 4				Raspberry Pi 5			
	min	**max**	**mean**	**std**	**min**	**max**	**mean**	**std**	**min**	**max**	**mean**	**std**
FF	0.01	0.16	**0.03**	0.01	0.04	0.35	0.05	0.02	0.02	0.12	0.03	0.01
EF	0.01	0.16	0.03	0.01	0.04	0.35	0.05	0.02	0.02	0.12	0.03	0.01
LBPH	0.01	0.17	0.03	0.01	0.04	**3.61**	0.08	**0.31**	0.02	1.95	0.04	0.16
FR	0.02	0.24	0.06	0.06	0.10	0.50	**0.18**	0.15	0.05	0.26	0.08	0.07
FN	0.02	0.24	**0.07**	0.04	0.12	1.09	**0.45**	0.32	0.05	0.92	0.18	0.15

In order to analyze the performance of the devices, the average and standard deviation were obtained from calculated metrics. Additionally, the minimum and maximum range achieved by the data was also determined.

4.1 Latency

Table 1 shows data obtained on latency in seconds across three hardware platforms (Desktop, Raspberry Pi 4, and Raspberry Pi 5). Each row represents one of the five different models that were implemented.

The Desktop consistently shows the lowest latency for all models, with averages ranging from 0.03 to 0.07 seconds, highlighting its superior processing capabilities. Comparing the Raspberry Pi devices, the Raspberry Pi 5 demonstrates better performance than the Raspberry Pi 4, with lower average and maximum latencies, as well as more consistent results. Regarding the models, FF and EF yield very similar results and maintain low latency across all hardware, while LBPH performs well on Desktop and Raspberry Pi 5 but exhibits significantly higher latency on Raspberry Pi 4 (maximum of 3.61 seconds and a standard deviation of 0.31). On the other hand, FR and FN are the most demanding models, especially on Raspberry Pi 4, where they reach averages of 0.18 and 0.45 seconds, respectively, although they perform better on Desktop and Raspberry Pi 5. This indicates that simpler models like FF and EF are suitable even for low-power devices like the Raspberry Pi, while more complex models such as FR and FN may not be ideal for Raspberry Pi 4

Tab. 2. RAM consumed on Desktop, Raspberry Pi 4 and Raspberry Pi 5.

Model	Desktop				Raspberry Pi 4				Raspberry Pi 5			
	min	max	mean	std	min	max	mean	std	min	max	mean	std
FF	2	1068	949	248	18.88	944	721	252	1	954	786	245
EF	2	1068	947	249	1	933	698	280	1	954	787	236
LBPH	1	1064	941	246	1	950	672	279	1	932	718	277
FR	1	1064	948	246	70	933	**755**	243	65	926	**768**	**240**
FN	1	1070	950	249	1	934	**644**	254	1	984	769	253

due to high latency and variability, though they are more feasible on Raspberry Pi 5. Overall, for real-time or latency-sensitive applications, Desktop remains the best option, while Raspberry Pi 5 could be a viable alternative for simpler models or optimized complex models, and Raspberry Pi 4 may be insufficient for demanding applications.

4.2 RAM Memory Consumed

Table 2 shows data obtained on RAM usage when running the models on the three types of hardware. These data are measured in MB.

On Desktop, all models show average values close to 950 MB with consistent standard deviations between 246 and 249 MB, indicating lower variability compared to other platforms. On Raspberry Pi 4, the average RAM consumption is lower than on Desktop, ranging from 644 MB (FN) to 755 MB (FR), but with higher variability, reaching standard deviations of up to 280 MB, suggesting less stable performance; additionally, some models, such as EF, have minimum RAM values of 1. On the other hand, Raspberry Pi 5 demonstrates better performance in terms of RAM consumption compared to Raspberry Pi 4, with higher but more consistent average values and lower standard deviations, indicating more stable behavior; for instance, the FR model has an average consumption of 768 MB with a low standard deviation of 240 MB, while LBPH shows the lowest average consumption across all platforms, standing out for its RAM efficiency. Comparing the platforms, Desktop uses more RAM on average, likely due to its higher processing capacity, while Raspberry Pi 5 proves to be more efficient than Raspberry Pi 4 by achieving a better balance between average consumption and stability. Regarding the models, FF shows high average consumption on Desktop and Raspberry Pi 5, while LBPH stands out as the most efficient in terms of RAM across all platforms, suggesting that the choice of model and hardware will depend on system priorities, such as stability, lower RAM consumption, or overall efficiency, with Raspberry Pi 5 being a notable improvement over Raspberry Pi 4.

4.3 CPU Usage

Table 3 present the CPU usage percentages for the models across the platforms.

Tab. 3. Usage CPU on Desktop, Raspberry Pi 4 and Raspberry Pi 5 with the five models.

CPU	Desktop				Raspberry Pi 4				Raspberry Pi 5			
	min	max	mean	std	min	max	mean	std	min	max	mean	std
FF model												
CPU 1	0	100	17.31	10.29	0	94.9	17.14	13.04	0	60	8.57	7.73
CPU 2	7.10	66.7	20.39	9.29	0	70	17.65	10.13	0	40	8.80	6.32
CPU 3	0	100	21.33	15.23	0	100	41.62	17.17	0	82.5	17.45	11.54
CPU 4	4.20	91.7	18.14	10.19	0	75	18.98	13.70	0	50	22.45	9.14
EF model												
CPU 1	0.00	100.00	17.57	10.03	7.10	86.70	22.79	14.76	0.00	75.00	29.09	16.98
CPU 2	4.00	76.90	22.19	10.84	0.00	58.30	19.06	10.69	0.00	57.10	18.19	11.26
CPU 3	0.00	100.00	18.53	10.82	7.10	100.00	29.62	19.39	0.00	40.00	11.90	10.82
CPU 4	0.00	100.00	17.37	12.82	0.00	75.00	29.09	16.98	0.00	89.50	13.33	11.64
LBPH model												
CPU 1	0.00	50.00	16.64	7.33	0.00	70.00	22.79	15.62	0.00	100.00	26.42	18.85
CPU 2	0.00	100.00	22.49	12.76	0.00	61.10	26.84	15.27	0.00	60.00	10.59	9.75
CPU 3	7.70	66.70	18.78	8.84	0.00	68.80	17.01	9.04	0.00	40.00	7.93	7.59
CPU 4	7.70	100.00	18.49	12.84	7.10	100.00	31.42	21.81	0.00	50.00	12.29	10.12
FR model												
CPU 1	0	100	15.10	16.93	0	100	15.10	16.93	0	35.70	7.42	7.20
CPU 2	0	100	24.74	29.36	0	100	24.74	29.36	0	72.70	7.13	12.58
CPU 3	0	83.30	10.71	13.34	0	83.30	10.71	13.34	0	100	21.77	29.17
CPU 4	0	100	18.26	23.64	0	100	18.26	23.64	0	100	39.84	32.72
FN model												
CPU 1	19.00	100.00	61.73	19.39	0.00	100.00	18.26	23.64	41.90	100.00	99.27	5.81
CPU 2	20.10	100.00	62.44	19.26	20.10	100.00	62.44	19.26	37.10	100.00	99.39	6.17
CPU 3	30.80	100.00	61.10	20.30	30.80	100.00	61.10	20.30	36.60	100.00	89.35	19.41
CPU 4	37.50	100.00	64.86	18.19	37.50	100.00	64.86	18.19	36.90	100.00	98.78	8.64

When analyzing the data from the tables showing CPU usage for the face recognition models, it was observed that CPU usage on the Desktop is generally higher compared to the Raspberry Pi 4 and Raspberry Pi 5, reflecting the greater processing power of the desktop hardware.

The Raspberry Pi 4 and Raspberry Pi 5 have lower CPU usage peaks and greater variations, indicating less stable and more fluctuating processing capabilities, with a higher standard deviation, suggesting that their workload is less consistent than on the Desktop.

Although the Raspberry Pi 5 shows slight performance improvements compared to the Raspberry Pi 4, both still exhibit inferior performance compared to the Desktop, with the Raspberry Pi 5 reaching an average performance close to the Desktop in some models, but with greater variability and higher maximum loads.

In terms of CPU demand, the FN model (which is more resource-intensive) shows the highest CPU usage values, especially on the Desktop, indicating that this model is more demanding in terms of processing. In contrast, the LBPH and FR models present a more moderate CPU load, making them more suitable for resource-limited platforms like the Raspberry Pi.

We can conclude that the Raspberry Pi 4 and 5, although more affordable and energy-efficient, exhibit more variable and less powerful performance compared to the Desktop. More demanding models like FN are better suited for the Desktop, while models like FF, EF, and LBPH can be run efficiently on the Raspberry Pi, although with somewhat less stable performance.

4.4 Temperature

The temperature data, measured in degrees Celsius, is presented in Table 4.

Tab. 4. Temperature on Desktop, Raspberry Pi 4 and Raspberry Pi 5.

Model	Desktop				Raspberry Pi 4				Raspberry Pi 5			
	min	**max**	**mean**	**std**	**min**	**max**	**mean**	**std**	**min**	**max**	**mean**	**std**
FF	28	28	28	0	53	57	54	0.88	53	61	55	1.61
EF	28	28	28	0	56	60	57	0.92	54	62	56	1.60
LBPH	28	28	28	0	**48**	51	50	0.57	**52**	58	54	1.06
FR	28	28	28	0	60	65	62	0.90	57	62	59	0.95
FN	28	28	28	0	66	**72**	**69**	1.90	64	**73**	70	2.18

Based on the temperature data from the table, it can be observed that the temperatures on the Desktop remain constant at 28°C across all models, suggesting that the Desktop hardware maintains a stable and low temperature during operation, likely due to a superior cooling system. On the other hand, the temperatures on Raspberry Pi 4 are significantly higher, ranging from 48°C to 72°C, particularly for the FN model, which is more demanding. The average temperatures for this model are between 49°C and 69°C, with a standard deviation ranging from 0.57 to 1.90. Raspberry Pi 5, in contrast, shows slightly lower temperatures than Raspberry Pi 4, with minimum temperatures around 52°C and maximum temperatures reaching up to 73°C. The average temperatures range from 54°C to 69°C, and the standard deviation varies between 0.95 and 2.18, reflecting a level of thermal variability similar to that of Raspberry Pi 4. The FN model generates the highest temperatures on both Raspberry Pi 4 and Raspberry Pi 5, surpassing 70°C, while the FF and LBPH models maintain lower and more stable temperatures across all platforms. The FR and EF models exhibit intermediate temperatures, with Raspberry Pi 4 reaching higher temperatures compared to Raspberry Pi 5. In general, we can observe that while the Desktop maintains low and stable temperatures, the Raspberry Pi 4 and Raspberry Pi 5 exhibit higher and more variable temperatures, particularly during intensive processing, with Raspberry Pi 5 providing slightly better thermal performance than Raspberry Pi 4.

5 Conclusions and Future Work

Our study demonstrates that desktop computers outperform Raspberry Pi platforms in terms of latency, CPU usage, and RAM consumption, as anticipated. Desktop systems consistently maintain low and stable latencies across all evaluated models, while Raspberry Pi devices, particularly the Raspberry Pi 4 exhibit higher latency and greater variability. This suggests that Raspberry Pi platforms are less suitable for high-demand or time-sensitive applications. Although the Raspberry Pi 5 offers a more stable performance compared to the Raspberry Pi 4, it still falls short of the desktop computer in both latency and CPU usage.

In terms of RAM consumption, the desktop computer records higher average usage but with lower variability, while the Raspberry Pi 5 demonstrates

improved efficiency relative to the Raspberry Pi 4, which shows more inconsistent performance and greater fluctuations. Regarding thermal behavior, the desktop computer maintains low and stable operating temperatures, whereas Raspberry Pi devices experience higher and more variable temperatures, particularly when running more demanding models such as FN.

Overall, while the Raspberry Pi 4 and 5 present more affordable and energy-efficient alternatives, their performance remains more variable and generally inferior compared to desktop systems. Consequently, more demanding models like FN are better suited for execution on desktop computers, whereas simpler models such as FF, EF, and LBPH can be deployed effectively on Raspberry Pi devices, albeit with slightly less stable performance.

From these findings, it can be inferred that security systems, which require low latency and high stability, are better suited to desktop platforms. Conversely, simpler facial recognition models can be effectively implemented on Raspberry Pi devices, accepting some degree of reduced stability. Similarly, home automation applications may benefit from the use of Raspberry Pi platforms, particularly when utilizing less demanding models, thereby taking advantage of their energy efficiency and low cost.

Future work will focus on expanding the dataset by incorporating a larger number of identities and images to improve the generalizability of the results. Advanced facial recognition models will be evaluated to enhance performance, particularly on low-resource devices. Additionally, the hardware comparison will be broadened to include popular edge platforms such as the Jetson Nano and Odroid N2+, providing a more comprehensive understanding of system performance across different environments. Optimization techniques will be explored, along with the integration of hybrid hardware accelerators, to improve the efficiency of more complex models. Furthermore, a deeper thermal analysis will be conducted, and energy consumption metrics will be incorporated, recognizing their critical role in the evaluation of embedded systems.

References

1. Anusudha, K., et al.: Real time face recognition system based on yolo and insightface. Multimedia Tools and Applications, vol. 83, no. 11, pp. 31893–31910 (2024)
2. Baobaid, A., Meribout, M., Tiwari, V. K., Pena, J. P.: Hardware accelerators for real-time face recognition: A survey. IEEE Access, vol. 10, pp. 83723–83739 (2022)
3. Belhumeur, P. N., Hespanha, J. P., Kriegman, D. J.: Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. In: Computer Vision—ECCV'96: 4th European Conference on Computer Vision Cambridge, UK, April 15–18, 1996 Proceedings, Volume I 4. pp. 43–58. Springer (1996)
4. Biglari, A., Tang, W.: A review of embedded machine learning based on hardware, application, and sensing scheme. Sensors, vol. 23, no. 4, pp. 2131 (2023)
5. Deeba, F., Memon, H., Dharejo, F. A., Ahmed, A., Ghaffar, A.: Lbph-based enhanced real-time face recognition. International Journal of Advanced Computer Science and Applications, vol. 10, no. 5 (2019)

6. Dewantoro, G., Mansuri, J., Setiaji, F. D.: Comparative study of computer vision based line followers using raspberry pi and jetson nano. Jurnal Rekayasa Elektrika, vol. 17, no. 4 (2021)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
8. Kaur, P., Krishan, K., Sharma, S. K., Kanchan, T.: Facial-recognition algorithms: A literature review. Medicine, Science and the Law, vol. 60, no. 2, pp. 131–139 (2020)
9. Khan, M. A., Shaikh, M. K., bin Mazhar, S. A., Mehboob, K., et al.: Comparative analysis for a real time face recognition system using raspberry pi. In: 2017 IEEE 4th International Conference on Smart Instrumentation, Measurement and Application (ICSIMA). pp. 1–4. IEEE (2017)
10. Kocacinar, B., Tas, B., Akbulut, F. P., Catal, C., Mishra, D.: A real-time cnn-based lightweight mobile masked face recognition system. Ieee Access, vol. 10, pp. 63496–63507 (2022)
11. Manni, A., Caroppo, A., Rescio, G., Siciliano, P., Leone, A.: Benchmarking of contactless heart rate measurement systems in arm-based embedded platforms. Sensors, vol. 23, no. 7, pp. 3507 (2023)
12. Rodriguez, Y., Marcel, S.: Face authentication using adapted local binary pattern histograms. In: Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006, Proceedings, Part IV 9. pp. 321–332. Springer (2006)
13. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 815–823 (2015)
14. Singh, S., Prasad, S.: Techniques and challenges of face recognition: A critical review. Procedia computer science, vol. 143, pp. 536–543 (2018)
15. Smith, M., Miller, S.: The ethical application of biometric facial recognition technology. Ai & Society, vol. 37, no. 1, pp. 167–175 (2022)
16. Turk, M., Pentland, A.: Eigenfaces for recognition. Journal of cognitive neuroscience, vol. 3, no. 1, pp. 71–86 (1991)

ISSN 1870-4069

NLP Applied to Musical Harmonic Structures for Music Emotion Recognition

Leonardo Daniel Villanueva Medina, Efrén Gorrostieta Hurtado

Universidad Autónoma de Queretaro,
Mexico

lvillanueva@uaq.mx

Abstract. Music emotion recognition (M.E.R.) is a multidisciplinary field that integrates computer science, affective computing, and neuroscience elements to analyze musical features to detect emotions. Most research in this field has focused on low and mid-level features, often ignoring theoretical and harmonic aspects of music. In this work, we propose using regression-based machine learning models applied to word embeddings in harmonic structures (chords). The results indicate an RMSE of 0.0252 and an R^2 score of 0.9751 for the valence dimension, in comparison with the arousal, an RMSE of 0.1319 and an R^2 score of 0.4676. These findings indicate that incorporating theoretical and harmonic concepts enhances the performance of M.E.R models, particularly in the valence dimension, reflecting improved detection of the positivity of emotions.

Keywords: mer, word embeddings, machine learning, musical features.

1 Introduction

Music has remarkably impacted social, cultural, and political aspects. For this reason, it has been the target of many studies, one of them being the relationship between emotions and music [13] since music is a means of expression capable of evoking emotions [6].

Music Emotion Recognition (M.E.R.) has incorporated knowledge from several fields, such as computer science, affective computing, and neuroscience. It aims to analyze musical features extracted from audio signals (low and mid-level) and abstract features such as song lyrics (high-level) [13, 9, 15, 7].

Within M.E.R.'s works, two approaches for linking emotions and songs predominate. The first one attaches a general emotion to the whole work (song-level), a static approach. The second, dynamic approach, focuses on detecting the music emotion variations that occur through the song, namely MEDV (music emotion variation) [9, 6].

Emotional perception is complex because it involves multiple variables, such as the song or external information, such as the listener's social, cultural, and emotional context [17, 7].

Selecting the appropriate taxonomy is crucial for clearly delineating the problem as either a multi-class classification or a regression task [9]. In this

regard, there are two main approaches: categorical taxonomies, which represent emotions through adjectives (such as Hevner's model [11]), and dimensional taxonomies, which represent emotions employing numerical values (such as Russell and Thayer's models [18, 16, 20]). The dimensional representation is organized based on two emotional axes: valence and arousal. Valence represents how pleasant an emotion is, and arousal represents excitement [7].

Traditional M.E.R. works are commonly based on the analysis of low-mid-level features. Therefore, as several works have proposed, information and features directly linked to emotions are needed [15, 17, 5, 22, 8].

Important musical concepts, such as theory and harmony, must be understood in the music-emotion relationship [11, 19, 15]. In this context, there are two essential elements: scales and chords. A scale is a succession of notes that follow a pattern. At the same time, a chord is a sequence of more than two notes. Each note forms a chord within a scale, which serves a function, such as determining the scale mode (major or minor) or indicating a transition or end/rest of a segment [10].

Similarly, according to Steinbeis' experiments [19], the listener expects a sense of closure by resting chords at the end of a harmonic progression. Replacing these with transitional chords can alter that emotional experience and change the expectation of the work's end.

Additionally, a certain similarity between the representation of chords and natural language has been pointed out, thus enabling the application of Natural Language Processing (NLP) techniques [12, 8].

This work introduces a technique combining harmony ideas and musical theory to improve music emotion recognition. We present adapted word embeddings to song chords based on natural language processing (NLP) methods. This method chooses suitable machine learning models to interpret the chord embeddings, facilitating a simpler, detailed analysis. Finally, the main goal of this work is to predict valence and arousal values to ensure that we can precisely detect the emotions perceived in music.

2 Background

The traditional methodology of M.E.R works is based on analyzing low-mid-level features through machine learning models. Nevertheless, multimodal strategies that assemble deep learning, NLP, and traditional techniques have been adopted, generating robust models.

In this regard, Panda et al.[15] underline the need for design elements that capture the music-emotion relation. This work is based on a novel set of features and employs a support vector machine model for the multi-classification problem, reaching a 76.4% value for an f1-score metric.

On the other hand, Yang's work [21] carries out emotion recognition through a Back Propagation algorithm (BP). The model's input was a set of six different low-level features. Enhance the BP algorithm with metaheuristic techniques, specifically an artificial bee colony algorithm. The results indicate an MAE of

0.8872, RMSE of 0.1066, and an R^2 of 0.4606 for the valence dimension. For the arousal dimension, an MAE of 0.9156, RMSE of 0.1322, and R^2 of 0.6687.

Some multi-modal approaches merge low-mid-level features with high-level information through deep learning and NLP models. In [17], the work addressed emotion recognition by analyzing two types of features. Researchers used CNN models for the low-level features (in spectrogram form) and applied several NLP methods to the song lyrics, reaching the best results with BERT embeddings.

On the other hand, in work [3], the focus is on the MER task using source separation from the PMEmo dataset, where the audio is split into four tracks (vocals, bass, drums, and other), from which spectrograms were extracted.

At the same time, efforts have been made to detect emotions using chords and harmony progressions. Cho's work [5] performs emotion detection based on MIDI and audio files. This work uses a chord matrix (coding the chord position) in combination with low-level features to predict the valence and arousal values through an SVR model. This results in an MSE of 0.67 for the valence dimension with the MIDI files and an MSE of 0.65 for the arousal using the audio files. In contrast, Zhang's work [22] tackles the M.E.R problem through statistical methods. A database that bonds emotion to a set of chords and identifies the chords by the analysis of low-level features.

NLP methods are not limited to song lyrics analysis. To demonstrate that embeddings can describe chords like music theory does, Lahnala's work [12], for example, uses Word2Vec predictive embeddings to capture the association between chords. On the other hand, Greer [8] improves the precision of music emotion classification by approaching the problem as a multi-classification problem by combining lyrics and chords into shared vectors.

3 Methods

Figure 1 illustrates the overall methodology diagram used in this work. The following sections provide a detailed explanation of each step.

3.1 Dataset

This study uses two well-known datasets in the M.E.R field: PMEMO [23] and MEDV [1]. Both datasets include static dimensional annotations composed of valence and arousal axes whose values are normalized from 0 to 1. The MEDV dataset includes 1802 MP3 audio files for 45 seconds. The PMEMO dataset has 767 audio MP3 files whose duration may vary. The sample frequency is the same in both datasets, 44.1 kHz.

3.2 Audio File Conversion

This work employs the Python libraries Librosa [14], Soundfile, and FFmpeg, which have functions that boost reading, conversion between different types of audio files, and writing of new files. The audio file format was converted from MP3 to new WAV files.

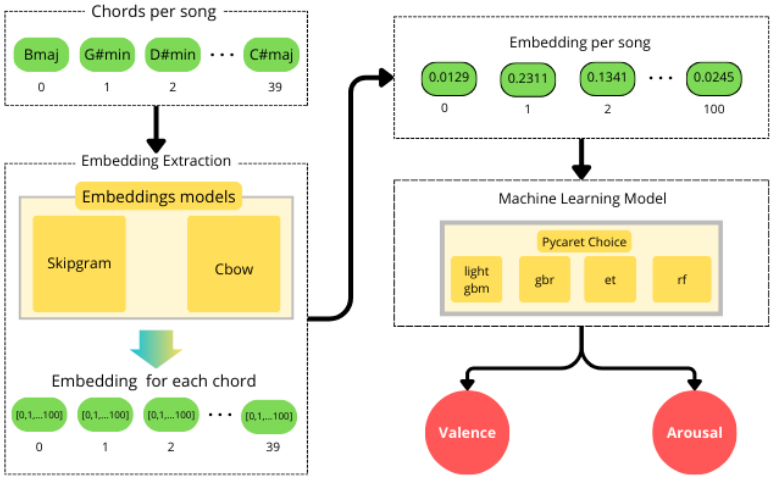


Fig. 1. Methodology implemented for the music emotion recognition (dimensional taxonomy).

3.3 Chord Detection and Data Augmentation

Chord detection was performed using the MADMOM [4] library (MADMOM only works with WAV files). In this way, chords were retrieved from the 2569 audio files. However, MADMOM only recognizes major and minor chords.

Chord transposition was used for data augmentation, a musical technique that increases or decreases the notes of a chord. Thus, transposition was applied in half-tone, whole-tone intervals (ascending and descending). The table 1 shows an example of this technique. With the increased data, the corpus was extended to 12845 progressions. Even with the limitation of MADMOM and the fact that there are only 12 sounds in Western music, the final dictionary is restricted to 24 chords.

3.4 Embeddings

Cooccurrence-based embeddings, such as word2vec, capture words' semantic and syntactic information and represent it in a vector space. In this way, each word forms a vector $\mathbb{R}^N$ where the dimension $N \geq 100$. Thus, based on their proximity in the plane, it can be known which words share context [17]. Skip-gram and CBOW are methods of such embeddings. Skip-gram predicts the context using the core word, while CBOW identifies the core word from the surrounding context [12, 8].

Based on Lanhala's work [12], CBOW and skip-gram were used as embeddings. The size of the embeddings was set to 100 and 200, while the

Table 1. Data augmentation for the song "I Have Questions" by "Camila Cabello" from the PMEMO dataset (just the first four chords).

Chords	Arousal	Valence	Type
G#min Emaj F#maj G#min	0.7375	0.7375	whole-tone Down
Amin Fmaj Gmaj Amin	0.7375	0.7375	half-tone Down
A#min F#maj G#maj A#min	**0.7375**	**0.7375**	**Source**
Bmin Gmaj Amaj Bmin	0.7375	0.7375	half-tone Up
Cmin G#maj A#maj Cmin	0.7375	0.7375	whole-tone Up

windows were 5, 10, and 20. Each song was set to a maximum length of 40 chords, truncating with zeros if necessary, and the embeddings for each chord were averaged to obtain a unique embedding.

Subsequently, a PCA dimensionality reduction algorithm was applied to visualize the relationship between chords in a two-dimensional plane. Figure 2 shows a circular arrangement similar to the "circle of fifths." This structure indicates the relationship and similarity between chords [12, 10]. Despite the limited dictionary, the representation of the relationship between major and minor chords is inadequate in the plan.

3.5 Machine Learning Model

A machine learning model has been implemented for emotion recognition. PyCaret library [2] was employed to automate the model selection and validation process for a regression task. The experimental setup used a train size of 80%, with the remaining 20% reserved for validation. The split was performed randomly to ensure a representative data distribution across both subsets. Various regression algorithms were trained and evaluated during the model comparison phase. Model performance was assessed based on the mean absolute error (MAE), root mean squared error (RMSE), and the coefficient of determination (R^2). Models were ranked according to these metrics, and the best-performing model was selected based on the primary evaluation criteria. Finally, a 10-fold cross-validation was performed to validate the best model chosen for valence and arousal.

The models that consistently performed best were the **Gradient Boosting Regressor (GBR)**, **Light Gradient Boosting Machine (LIGHTBGM)**, **Random Forest (RF)**, and **Extra Trees Classifier (ET)**. The model takes the unique embeddings as input and produces the valence and arousal values as output.

The RMSE, MAE, and R^2 metrics are the most commonly used in regression problems for emotion recognition [21].

The root mean square error (RMSE) measures the dispersion of errors and penalizes extreme values (equation 1, [21]). The mean absolute error (MAE) calculates the average difference between the actual and predicted values (equation 2, [21]). The coefficient of determination R^2 evaluates how well the

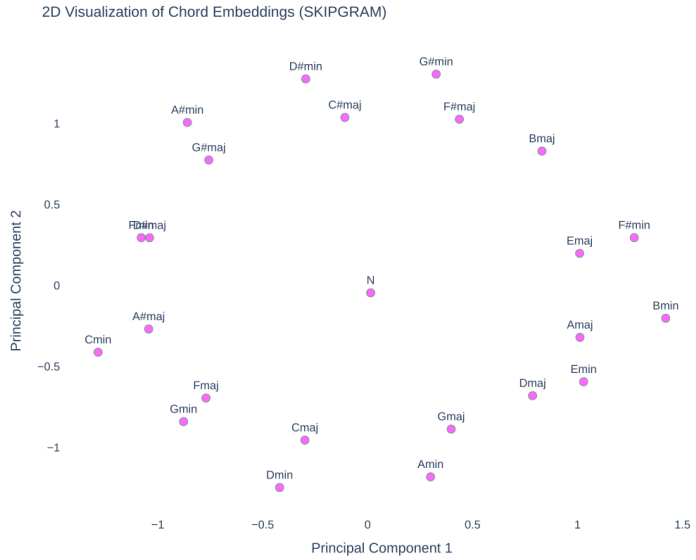


Fig. 2. Vector representation of the relationships captured by the single chord embeddings. The minor chords form the outer circle, and the major chords form the inner circle.

predicted values match the actual values; an R^2 close to 1 indicates better accuracy (equation 3, [21]).

Where y_i are the observed values, $\hat{y}_i$ is the model predictions, and $\bar{y}$ represents the mean of the observed values:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \tag{1}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right|, \tag{2}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}. \tag{3}$$

4 Results

4.1 Embeddings

The Skip-gram and CBOW models retrieved the embedding of each dictionary element. Both methods capture the relationship between chords just as music theory does. However, when evaluating the cosine similarity of the five most similar chords (figure 3), CBOW showed lower values, so Skip-gram embeddings were used exclusively in the subsequent experiments.

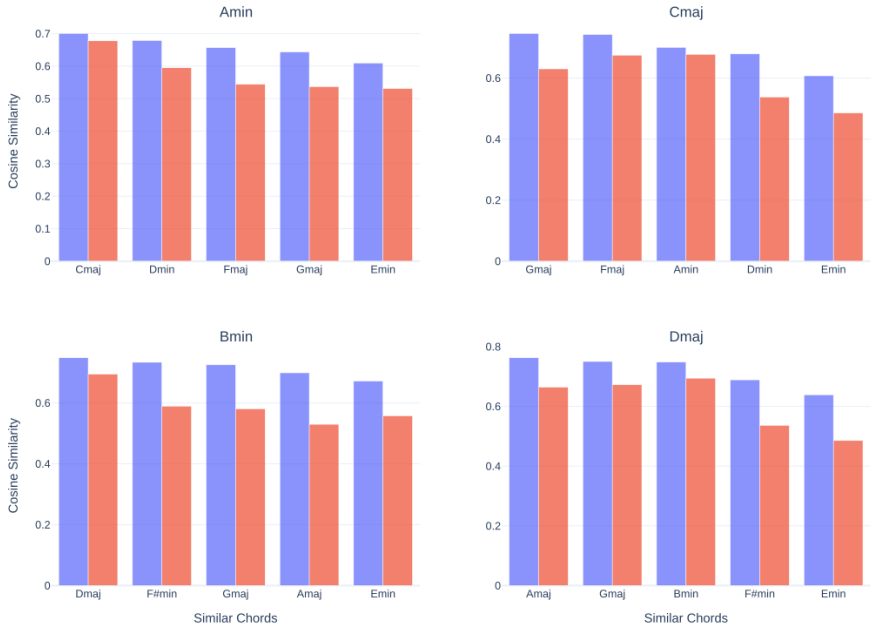


Fig. 3. Blue: skip-gram. Red: Cbow. It shows the five most similar chords: A minor, C major, B minor, and D major.

4.2 Emotion Recognition

The search for the best machine learning models with PyCaret yields metrics from Tables **??** and 2, which compare the models grouped by valence and arousal.

In the arousal dimension, the GBR and LIGHTGBM models demonstrate comparable results. The difference between errors is slight and does not exceed 0.0006 points for the MAE and 0.0019 for the RMSE. The models differ in the R^2, with the highest value, 0.4620, and the lowest, 0.4573.

According to Table 2, in the dimension of arousal, the error is comparatively reduced when working with embeddings of size 100, and the value of R^2 is higher than with embeddings of size 200. The best result is obtained with embeddings of size 100 and a window of 10, employing the lightgbm model.

In the valence dimension, the LIGHTGMB model achieves fewer errors and better scores on the R^2. In general, for this axis, better results are obtained with embeddings of size 200. However, the best result is achieved with 100 embeddings and a 5-window, with an MAE of 0.0175, an RMSE of 0.0260, and an R^2 of 0.9733.

Table **??** shows how, in the embedding vectors with larger size (200), better results are achieved with larger contextual windows (10 and 20). On the contrary, in embeddings with reduced size (100), the best result is achieved with a small contextual window (5).

Table 2. Comparison of metrics for Arousal with different models and configurations (Dim: 100 and Dim: 200).

	Dim: 100			**Dim: 200**		
	MAE	**RMSE**	**R^2**	**MAE**	**RMSE**	**R^2**
Win: 5						
gbr	**0.1062**	**0.1330**	**0.4585**	0.1064	0.1332	0.4571
lightgbm	0.1062	0.1332	0.4566	**0.1061**	**0.1331**	**0.4573**
rf	0.1062	0.1339	0.4513	0.1067	0.1343	0.4475
Win: 10						
gbr	0.1062	0.1330	0.4581	**0.1063**	**0.1331**	**0.4573**
lightgbm	**0.1058**	**0.1326**	**0.4620**	0.1061	0.1332	0.4571
rf	0.1064	0.1341	0.4493	0.1065	0.1339	0.4506
Win: 20						
gbr	0.1064	0.1332	0.4564	**0.1062**	**0.1331**	**0.4574**
lightgbm	**0.1061**	**0.1330**	**0.458**2	0.1065	0.1337	0.4529
rf	0.1064	0.1349	0.4508	0.1065	0.1342	0.4484

The values of MAE, RMSE, and R^2 for Arousal. 'Dim' corresponds to the embedding dimension, and 'Win' corresponds to the size of the contextual window.

4.3 Cross Validation

The models with the best results in embeddings of size 100 and window size five were selected since this configuration optimized the performance in valence. Thus, LIGHTGBM was chosen for valence and GBR for arousal. Figure 4 shows that the valence predictions are close to the actual values, while the arousal predictions are concentrated in a middle range, moving away from the baseline. Each model was retrained and validated using a 10-fold cross-validation strategy.

The results obtained from this experiment are summarized in Tables 4 and 5. For the arousal dimension, the model achieved a mean MAE of 0.1049, a mean RMSE of 0.1319, and a mean R^2 of 0.4676. In the case of valence, the model achieved a mean MAE of 0.0167, a mean RMSE of 0.0252, and a mean R^2 of 0.9751, with relatively low standard deviations across all folds.

A subset of 'simple' songs (those with reduced progressions and only major and minor chords) was evaluated as a comparison. Figure 5 shows how the predictions follow the previous trends in both dimensions, highlighting a better performance in arousal for these songs.

4.4 Comparison

Table 6 compares the best results of selected state-of-the-art works with the best model proposed in this study. The results highlight that the analysis of harmonic structures achieves better performance than the analysis of purely acoustic

Table 3. Comparison of metrics for Valence with different models and configurations (Dim: 100 and Dim: 200).

	Dim: 100			**Dim: 200**		
	MAE	**RMSE**	**R^2**	**MAE**	**RMSE**	**R^2**
Win: 5						
lightgbm	**0.0175**	**0.0260**	**0.9733**	**0.0332**	**0.0461**	**0.9160**
et	0.0186	0.0277	0.9696	0.0349	0.0486	0.9069
rf	0.0188	0.0280	0.9691	0.0351	0.0488	0.9060
Win: 10						
lightgbm	**0.0328**	**0.0457**	**0.917**6	**0.0181**	**0.0272**	**0.9708**
et	0.0343	0.0478	0.9098	0.0193	0.0289	0.9671
rf	0.0345	0.0483	0.9098	0.0196	0.0294	0.9660
Win: 20						
lightgbm	**0.0327**	**0.0463**	**0.9192**	**0.0175**	**0.0268**	**0.9717**
et	0.0342	0.0473	0.9118	0.0184	0.0283	0.9684
rf	0.0345	0.0478	0.9099	0.0186	0.0286	0.9678

The values of MAE, RMSE, and R^2 for Valence. 'Dim' corresponds to the embedding dimension, and 'Win' corresponds to the size of the contextual window.

Table 4. Cross-validation results for the arousal dimension after data augmentation.

Fold	**MAE**	**RMSE**	**R^2**
0	0.1073	0.1343	0.4541
1	0.1024	0.1276	0.4609
2	0.1023	0.1284	0.4850
3	0.1071	0.1353	0.4693
4	0.1014	0.1282	0.4961
5	0.1025	0.1291	0.4956
6	0.1052	0.1314	0.4574
7	0.1099	0.1384	0.4433
8	0.1056	0.1335	0.4609
9	0.1054	0.1331	0.4534
Mean	**0.1049**	**0.1319**	**0.4676**
Std	**0.0026**	**0.0034**	**0.0175**

The table reports the MAE, RMSE, and R^2 metrics obtained from 10-fold cross-validation for the gbr model for the arousal dimension.

characteristics, at least for recognizing the valence (positivity or negativity) a listener perceives.

Skip-gram predictive embeddings facilitate pattern learning in artificial intelligence models, enhancing the analysis of harmonic structures for emotion recognition. Thus, combining high-level harmonic features informed by music

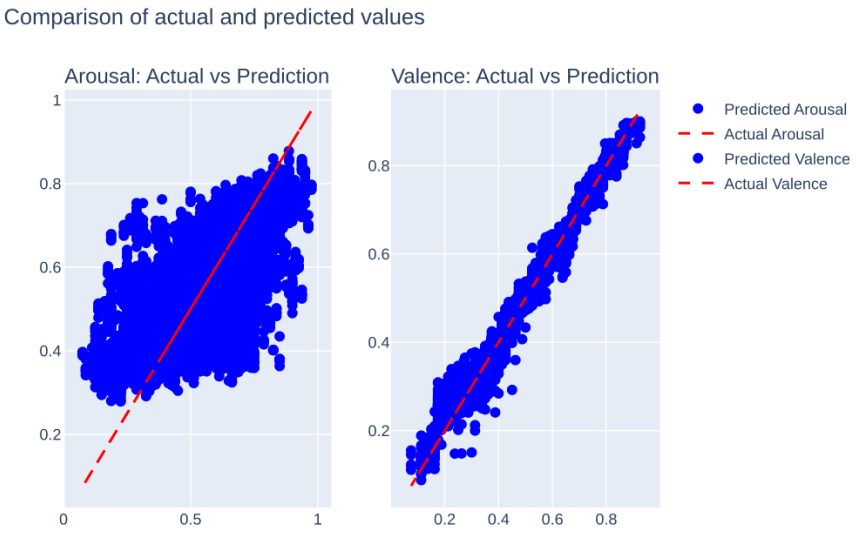


Fig. 4. Behavior of the model in the dimensions of arousal and valence.

Table 5. Cross-validation results for the valence dimension after data augmentation.

Fold	MAE	RMSE	R^2
0	0.0166	0.0255	0.9733
1	0.0166	0.0253	0.9769
2	0.0163	0.0233	0.9786
3	0.0162	0.0239	0.9760
4	0.0165	0.0253	0.9756
5	0.0169	0.0259	0.9740
6	0.0168	0.0251	0.9751
7	0.0168	0.0243	0.9769
8	0.0173	0.0273	0.9710
9	0.0166	0.0257	0.9733
Mean	**0.0167**	**0.0252**	**0.9751**
Std	**0.0003**	**0.0011**	**0.0021**

The table reports the MAE, RMSE, and R^2 metrics obtained from 10-fold cross-validation for the lightgbm model for the valence dimension.

theory with natural language processing techniques proves more efficient than traditional MER approaches for recognizing the valence dimension.

However, low-level acoustic features remain superior for recognizing the arousal dimension. Acoustic-based systems typically consider a wide range of features such as pitch, timbre, and rhythm. In contrast, the harmonic analysis

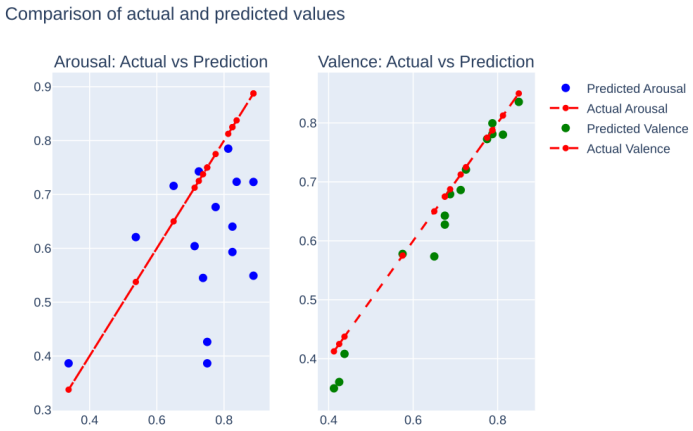


Fig. 5. Comparison between actual and predicted values for the easy songs set.

in this work was limited to only two chord modes (major and minor), focusing on basic chord structures.

Table 6. Comparison of state-of-the-art results and the present work. RMSE and R^2 metrics for Arousal and Valence are shown.

	RMSE		**R^2**		**T.C**
	Arousal	**Valence**	**Arousal**	**Valence**	
Y-H. Cho [5]	0.806	1.104	—	—	Chords
PMEMO [23]	**0.102**	0.124	—	—	Acoustics
EmoMucs [3]	0.2307	0.2373	0.6046	0.4584	Acoustics
Yang [21]	0.1322	0.1066	0.6687	0.4606	Acoustics
Proposed	0.1319	**0.0252**	0.4676	0.9751	Chords

Note: The table shows the best results of each work. Column T.C references the main feature that drives the models. Works marked with --- did not report the R^2 metric.

5 Discussion and Conclusion

The predictive skip-gram embeddings could capture the music-theoretic relationship between major and minor chords, similar to Lahnala's work [12]. When representing the embeddings in the plane, the distribution of the embeddings is almost similar to the circle of fifths of music theory, failing in the distribution of the outer circle (relative minors). However, the data set used

in this work was limited since, unlike Lahnala's or Greer's work [12, 8], only major and minor chords were extracted, resulting in a reduced chord dictionary. Even so, this only affects the representation in the plane since when calculating the cosine similarity (figure 3), we observe how the embeddings manage to capture the relationship between relative major and minor. Thus, skip-gram boosted with data augmentation can capture relationships between chords as music theory dictates.

Finally, it can be observed how analyzing the harmonic structure of a song improves the results obtained by M.E.R systems based on the analysis of low-level acoustic features [23, 21], demonstrating the strong link of harmonic structure in the perception of positivity of emotions in music, furthermore, using only major and minor chords does not affect emotion recognition and boosts the results in the valence dimension because major and minor modes are often associated with feelings such as happiness and sad [11]. The model performs better than the average in the arousal dimension, as indicated by the R^2. However, the result is low compared to works such as [23, 21]. This may be so because the harmonic structure of a song does not reflect other characteristics such as rhythm, color, or degree of energy, thus determining that, in predicting the degree of intensity of an emotion, it is better to work with low-level characteristics. Nonetheless, there is significant potential for improvement, particularly by incorporating more complex chords. Additionally, implementing contextual embeddings could enhance the results further. However, this would necessitate a larger dataset.

In conclusion, analyzing harmonic structures with basic chords improves emotion recognition based on dimensional taxonomies and significantly impacts predicting valence. Nonetheless, the exclusive analysis of harmonic structures leaves aside other factors related to the arousal axis, such as rhythm, which affects the performance of machine learning models in predicting values of this axis.

As future work, we aim to explore multimodal approaches that combine chord embeddings with acoustic features derived from spectrogram representations, such as chromagrams, constant-Q transforms (CQT), and mel-spectrograms. While chord embeddings effectively capture the harmonic structure relevant to valence prediction, spectrogram-based features offer a richer representation of the temporal and dynamic aspects of music. By integrating these modalities, we expect to enhance the model's ability to capture the energy and intensity variations associated with arousal, which are not fully reflected in harmonic content alone.

References

1. Alajanki, A., Yang, Y.-H., Soleymani, M.: Benchmarking music emotion recognition systems. PLOS ONE, (2016)
2. Ali, M.: PyCaret: An open source, low-code machine learning library in Python (April 2020), https://www.pycaret.org, pyCaret version 1.0

3. de Berardinis, J., Cangelosi, A., Coutinho, E.: The multiple voices of musical emotions: source separation for improving music emotion recognition models and their interpretability. International Society for Music Information Retrieval Conference, (2020)
4. Böck, S., Korzeniowski, F., Schlüter, J., Krebs, F., Widmer, G.: madmom: a new Python Audio and Music Signal Processing Library. In: Proceedings of the 24th ACM International Conference on Multimedia. pp. 1174–1178. Amsterdam, The Netherlands (10 2016) doi: 10.1145/2964284.2973795
5. Cho, Y.-H., Lim, H., Kim, D.-W., Lee, I.-K.: Music emotion recognition using chord progressions. In: 2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC). pp. 002588–002593. IEEE (10 2016) doi: 10.1109/SMC.2016.7844628
6. Cui, X., Wu, Y., Wu, J., You, Z., Xiahou, J., Ouyang, M.: A review: Music-emotion recognition and analysis based on eeg signals. Frontiers in Neuroinformatics, vol. 16, pp. 997282 (10 2022) doi: 10.3389/FNINF.2022.997282/BIBTEX
7. Gomez-Canon, J. S., Cano, E., Eerola, T., Herrera, P., Hu, X., Yang, Y.-H., Gomez, E.: Music emotion recognition: Toward new, robust standards in personalized and context-sensitive applications. IEEE Signal Processing Magazine, vol. 38, pp. 106–114 (11 2021) doi: 10.1109/MSP.2021.3106232
8. Greer, T., Singla, K., Ma, B., Narayanan, S.: Learning shared vector representations of lyrics and chords in music. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 3951–3955. IEEE (2019)
9. Han, D., Kong, Y., Han, J., Wang, G.: A survey of music emotion recognition. Frontiers of Computer Science, vol. 16, pp. 166335 (12 2022) doi: 10.1007/s11704-021-0569-4
10. Herrera, E.: Teoría musical y armonía moderna Vol. 2, vol. 2. Antoni Bosch editor (2022)
11. Hevner, K.: Experimental studies of the elements of expression in music. The American Journal of Psychology, vol. 48, pp. 246 (4 1936) doi: 10.2307/1415746
12. Lahnala, A., Kambhatla, G., Peng, J., Whitehead, M., Minnehan, G., Guldan, E., Kummerfeld, J. K., Çamcı, A., Mihalcea, R.: Chord embeddings: Analyzing what they capture and their role for next chord prediction and artist attribute prediction. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 12693 LNCS, pp. 171–186 (2021) doi: 10.1007/978-3-030-72914-1_12
13. Lucia-Mulas, M. J., Revuelta-Sanz, P., Ruiz-Mezcua, B., Gonzalez-Carrasco, I.: Automatic music emotion classification model for movie soundtrack subtitling based on neuroscientific premises. Applied Intelligence, vol. 53, pp. 27096–27109 (11 2023) doi: https://doi.org/10.1007/s10489-023-04967-w
14. McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E., Nieto, O.: librosa: Audio and music signal analysis in python. SciPy, vol. 2015, pp. 18–24 (2015)
15. Panda, R., Malheiro, R., Paiva, R. P.: Novel audio features for music emotion recognition. IEEE Transactions on Affective Computing, vol. 11, pp. 614–626 (10 2020) doi: 10.1109/TAFFC.2018.2820691
16. Posner, J., Russell, J. A., Peterson, B. S.: The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. Development and Psychopathology, vol. 17, pp. 715–734 (7 2005) doi: 10.1017/S0954579405050340

17. Pyrovolakis, K., Tzouveli, P., Stamou, G.: Multi-modal song mood detection with deep learning. Sensors 2022, Vol. 22, Page 1065, vol. 22, pp. 1065 (1 2022) doi: 10.3390/S22031065
18. Russell, J. A.: A circumplex model of affect. Journal of Personality and Social Psychology, vol. 39, pp. 1161–1178 (12 1980) doi: 10.1037/H0077714
19. Steinbeis, N., Koelsch, S., Sloboda, J. A.: The role of harmonic expectancy violations in musical emotions: Evidence from subjective, physiological, and neural responses. Journal of Cognitive Neuroscience, vol. 18, pp. 1380–1393 (8 2006) doi: 10.1162/jocn.2006.18.8.1380
20. Thayer, R. E.: The Biopsychology of Mood and Arousal. Oxford University PressNew York, NY (9 1990)
21. Yang, J.: A novel music emotion recognition model using neural network technology. Frontiers in Psychology, vol. 12, pp. 760060 (9 2021) doi: 10.3389/FPSYG.2021.760060
22. Zhang, F., Meng, H., Li, M., Cui, R., Liu, C.: Music emotion recognition based on chord identification. In: The International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery. pp. 956–963. Springer (2020)
23. Zhang, K., Zhang, H., Li, S., Yang, C., Sun, L.: The pmemo dataset for music emotion recognition. In: Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval. pp. 135–142. ICMR '18, ACM, New York, NY, USA (2018) doi: 10.1145/3206025.3206037

ISSN 1870-4069

Sesgos inductivos relacionales en mecanismos de atención

Víctor Mijangos [1], Ximena Gutierrez-Vasques [2], Verónica E. Arriola [1], Ulises Rodríguez-Domínguez [1], Alexis Cervantes [1], José Luis Almanzara [1]

[1] Universidad Nacional Autonoma de Mexico,
Facultad de Ciencias,
México

[2] Universidad Nacional Autonoma de Mexico,
CEIICH,
México

{vmijangosc, v.arriola, ulises.rodriguez.dominguez, alexis.cervantes, jose-luis}@ciencias.unam.mx, xim@unam.mx

Resumen. El aprendizaje inductivo busca construir modelos generales a partir de ejemplos específicos, siendo guiado por sesgos inductivos que influyen en la selección de hipótesis y determinan la capacidad de generalización. En este trabajo, nos centramos en caracterizar los sesgos inductivos relacionales presentes en los mecanismos de atención, entendidos como suposiciones sobre las relaciones subyacentes entre los datos. Desde el marco del aprendizaje profundo geométrico, analizamos los mecanismos de atención más comunes en términos de sus propiedades de equivariancia respecto a subgrupos de permutaciones, lo que nos permite proponer una clasificación basada en sus sesgos relacionales.

Palabras clave: Mecanismos de atención, transformadores, sesgos inductivos, aprendizaje profundo geométrico.

Relational Inductive Biases in Neural Attention Mechanisms

Abstract. Inductive learning aims to build general models from specific examples, guided by inductive biases that influence hypothesis selection and determine generalization capability. In this work, we focus on characterizing relational inductive biases present in attention mechanisms, understood as assumptions about underlying relationships between data elements. Within the framework of geometric deep learning, we analyze common attention mechanisms in terms of their equivariance properties with respect to permutation subgroups, which allows us to propose a classification based on their relational biases.

Keywords: Attention mechanisms, transformers, inductive biases, geometric deep learning.

1. Introducción

Uno de los paradigmas más comunes en los métodos actuales de aprendizaje de máquina es el *aprendizaje inductivo* cuyo objetivo es construir una función general, o hipótesis, a partir de ejemplos particulares observados [20,21]. De esta manera, dado un conjunto de ejemplos de entrenamiento (pares de entrada-salida), el modelo busca una hipótesis, dentro de un espacio de posibles hipótesis, que se ajuste bien a los datos de entrenamiento y que pueda generalizarse bien a instancias que no se han visto antes.

Las suposiciones realizadas por un algoritmo de aprendizaje para proponer hipótesis, que estén definidas para todo el dominio del problema y no sólo para los valores de los ejemplares observados, constituyen un *sesgo inductivo*. Son estas suposiciones las que le dotan de potencial para generalizar a datos no vistos [21]. Otra forma de sesgo inductivo son aquellas suposiciones que prefieren ciertas hipótesis sobre otras. Por lo tanto, los sesgos inductivos juegan un papel clave en la capacidad de generalización de los modelos de aprendizaje de máquina.

A pesar de que los sesgos inductivos son uno de los componentes que permiten el aprendizaje de diversas tareas, e.g., procesamiento del lenguaje natural (NLP), reconocimiento y generación de imágenes, es difícil encontrar trabajos que aborden formalmente los fundamentos de los sesgos inductivos, particularmente en el aprendizaje profundo.

En este artículo nos centraremos en exponer el funcionamiento de uno de los tipos de sesgos inductivos más prominentes, los *sesgos relacionales*, que son aquellos que explotan la estructura relacional inherente a los datos [3]. Nuestro análisis estará aplicado a las redes neuronales de tipo transformador (*transformer*) y sus mecanismos de atención. Proponemos adoptar las estrategias del aprendizaje profundo geométrico para describir los sesgos relacionales, pues proveen un marco de estudio robusto para modelar estructuras relacionales en los datos, así como transformaciones y simetrías.

2. Marco teórico

2.1. Sesgos inductivos relacionales

Los sesgos inductivos son suposiciones previas, *a priori*, que ayudan a que el aprendizaje pueda elegir mejor una hipótesis sobre otra [20]. Tom Mitchell [21] propone plantearse la pregunta de cuáles son los *a priori* necesarios para que el algoritmo de aprendizaje pueda realizar un proceso deductivo para generalizar sobre una nueva instancia x. Estas suposiciones *a priori* son precisamente lo que se entiende por sesgos inductivos.

Definición 1 (Sesgo inductivo) *Sea $f^* : X \to Y$ una función objetivo arbitraria y sea $D = \{(x, f^*(x)) : x \in X\}$ un conjunto de datos de entrenamiento y A un algoritmo de aprendizaje. Un sesgo inductivo de A es un conjunto mínimo de suposiciones B tal que para todo $x \in X$:*[1]

$$(B \wedge D \wedge x) \vdash A(x, D)$$

[1] Usamos la notación de Mitchell [21] tomando a B y D como conjunciones sobre sus elementos vistos como proposiciones.

. Donde $A(x,D)$ es la predicción del algoritmo entrenado sobre D en la instancia x, $\vdash$ denota la inferencia de $A(x,D)$ a partir de $(B \wedge D \wedge x)$:

En particular, nos interesan los sesgos inductivos relacionales [3] para analizar datos conformados por conjuntos de entidades, que se encuentran estructuralmente relacionadas entre sí. Por ejemplo, sea X un conjunto de enunciados en español, cada ejemplar x está constituido por el conjunto de palabras $\{x_1,...,x_n\}$ en el enunciado.

Definición 2 (Sesgo inductivo relacional) *Dado un conjunto de entidades que conforman un ejemplar $x = \{x_1,...,x_n\}$, un sesgo inductivo relacional es una suposición acerca de las relaciones entre dichas entidades. Es decir, es una estructura relacional (x,G), tal que $G = (V,E)$ es una gráfica relacionando los elementos de x.*

2.2. Aprendizaje profundo geométrico

Para estudiar los sesgos inductivos relacionales dentro de los **mecanismos de atención**, es crucial adoptar un marco teórico que permita caracterizar y relacionar los diferentes mecanismos que se han propuesto. Bronstein et al. [6] adoptan un enfoque basado en características geométricas y sus grupos de simetrías. Papillon et al. [22] han extendido este enfoque a una perspectiva topológica. Gavranović et al [13] proponen un marco más general basado en teoría de categorías para englobar tanto perspectivas geométricas y topológicas (que llaman "descendentes" o *top-down*) como marcos "ascendentes" (*bottom-up*), que parten de la construcción de arquitecturas con base a métodos de diferenciación automática [1], [23], [5]. Nos enfocaremos en las relaciones establecidas por una gráfica que se asume dentro de los modelos de atención. Adoptamos la teoría del aprendizaje geométrico profundo [7], ya que este modelo teórico, al enfocarse en estructuras geométricas del dominio de datos, es ideal para expresar las relaciones entre las entidades de los datos.

El aprendizaje profundo geométrico adopta una metodología de estudio basada en el programa de Felix Klein para la geometría [16]: se define un conjunto o dominio de elementos y grupos de transformaciones asociadas a este dominio. Con base en esta idea, el objetivo del aprendizaje geométrico profundo es determinar características (principalmente relacionales) sobre el dominio de datos y determinar el tipo de capas ocultas que pueden aprovechar dichas características.

La idea esencial detrás del aprendizaje profundo geométrico radica en estudiar a los ejemplares x y las relaciones entre sus entidades x_i. Estas suelen representarse a través de gráficas $G = (V,E)$ donde los vértices en V se asocian a las entidades y E representa sus relaciones. El dominio de los datos y la tarea determinan el tipo de capas de un modelo profundo. Estas capas deben preservar la información relevante para resolver las tareas. Para formalizar esto, es primordial el concepto de función equivariante.[2]

[2] Un ejemplo de especial interés son las capas convolucionales. En estas, el dominio consiste en imágenes, que pueden entenderse como elementos $x \in \mathbb{R}^{H\times W\times C}$, donde H es la altura, W la anchura y C el número de canales. La acción de una traslación $g \in \mathscr{G}$ puede representarse por una matriz de traslación $T = \rho(g)$, de tal forma que Tx es una traslación de la imagen x. Una convolución $f : \mathbb{R}^{H\times W\times C} \to \mathbb{R}^{H\times W\times C}$ es equivariante ante traslaciones en tanto se puede comprobar la igualdad $f(Tx) = Tf(x)$.

Definición 3 (Función equivariante) *Sea $\mathscr{G}$ un grupo de simetrías sobre un conjunto X, una función $f : X \to X$ se dice que es $\mathscr{G}$-equivariante si para toda acción del grupo $g \in \mathscr{G}$ se cumple que:*

$$f\big(\rho(g)x\big) = \rho(g)f(x). \tag{1}$$

Donde $\rho(g)$ es la representación (matricial) de la acción g.

Otro concepto importante dentro del marco teórico es el de invarianza y las funciones invariantes donde se tiene que $f\big(\rho(g)x\big) = f(x)$. Dadas las características de los mecanismos de atención nos enfocaremos en funciones $\mathscr{G}$-equivariantes, o simplemente equivariantes. El marco del aprendizaje profundo geométrico permite definir una gran cantidad de capas principalmente para arquitecturas de redes neuronales de gráficas. Veremos que las capas de atención caben bajo el concepto de una capa gráfica.

Definición 4 (Capa gráfica) *Una capa gráfica (o de envío de mensajes) depende de una gráfica $G = (V,E)$ con un sistema de vecindades $\mathscr{N}_v$ para cada vértice asociado a una entidad x_v. De tal forma que la nueva entidad oculta h_v es de la forma:*

$$h_v = \phi\left(x_v, \bigoplus_{u \in \mathscr{N}_v} \psi(x_v, x_u)\right). \tag{2}$$

En donde se distinguen los siguientes elementos:

1. *Función de mensaje $\psi(x_v, x_u)$: Genera un mensaje (generalmente un vector) con base en x_v y sus vecinos x_u.*
2. *Función de agregación $\bigoplus$: Determina cómo se combinan los mensajes para la actualización; se pide que sea un operador conmutativo.*
3. *Función de actualización $\phi(\cdot,\cdot)$: Define la forma final que tendrá la nueva representación h_v con base en la representación original x_v y la estructura relacional.*

En la Sección 3, estudiamos las equivarianzas de los mecanismos de atención a partir de subgrupos de permutaciones finitas S_n. Para esto tomamos como punto de partida el teorema de Cayley que apunta a que todo grupo es isomorfo a un subgrupo de permutaciones [8].

2.3. Mecanismos de atención

Los mecanismos de atención fueron introducidos, para el problema de traducción automática en redes recurrentes, por Bahdanau et al. [2]. Posteriormente, Vaswani et al. [27] propusieron reemplazar las recurrencias por mecanismos de atención que trabajaran sobre los mismos datos de entrada, lo que llamó auto-atención (*self-attention*). A partir de estos trabajos se han definido nuevas arquitecturas y capas de atención, de las cuales aquí presentamos las más comunes, comenzando por la auto-atención [27].

Definición 5 (Auto-atención) *Dado un conjunto de entidades de un ejemplar* $\{x_1,..,x_n\}$ *con* $x_i \in \mathbb{R}^d$*, un mecanismo de auto-atención es una capa de la forma:*

$$h_i = \sum_j \alpha(x_i,x_j)\psi_v(x_j), \tag{3}$$

donde $\alpha(x_i,x_j)$ *son los pesos de atención definidos como:*

$$\alpha(x_i,x_j) = Softmax\Big(\frac{\psi_k(x_j)^T\psi_q(x_i)}{\sqrt{d}}\Big). \tag{4}$$

Denotamos con ψ_q, ψ_k *y* ψ_v *a las proyecciones aplicadas a las entidades de entrada, generalmente definidas por una función lineal o afín.*

Los transformadores [27] integran un tipo de atención similar a la de Bahdanau [2] que representa las entidades de entrada (en el codificador) con las de la salida (el decodificador). Esta atención codificador-decodificador, puede ser definida bajo los conceptos de la Definición 5 restringiendo la forma en que se determinan las relaciones.

Definición 6 (Atención codificador-decodificador) *Supóngase una bipartición determinada por los conjuntos de entidades* $X = \{x_1,...,x_m\}$ *e* $Y = \{y_1,...,y_n\}$*. Definimos la atención codificador-decodificador como el mecanismo que para todo* y_i*, con* $i \in \{1,...,n\}$*, obtiene las representaciones ocultas como:*

$$h_i = \sum_{j=1}^{m} \alpha(y_i,x_j)\psi_v(x_j). \tag{5}$$

Donde $\alpha(y_i,x_j)$ *se define de forma similar a la Ecuación 4.*

Cuando los mecanismos de atención se utilizan para predecir un elemento (entidad), dado un conjunto previo (como en la generación de texto), el entrenamiento de estos modelos no puede asumir que los elementos previos dependen de los futuros, que aún no conoce. Para lidiar con esto, los decodificadores en los transformadores implementan mecanismos de atención enmascarada [27].

Definición 7 (Atención enmascarada) *Dado un conjunto de entrada de entidades ordenadas* $x_1,..,x_n$*, un mecanismo de atención enmascarada es una capa que obtiene representaciones de la forma:*

$$h_i = \sum_{j\leq i} \alpha(x_i,x_j)\psi_v(x_j), \tag{6}$$

donde los pesos de atención $\alpha(x_i,x_j)$ *se estiman como en la Ecuación 4.*

En la definición previa hemos definido un orden para simplificar la expresión en la sumatoria. La idea de desconectar nodos dentro de la gráfica que define las relaciones entre las entidades en un mecanismo de atención puede llevarse todavía más lejos. Por ejemplo, Child et al. [9] sugieren un mecanismo de atención por pasos donde las relaciones están acotadas por distintas restricciones, como un límite a los elementos previos con los que se puede conectar una entidad.

Definición 8 (Atención por pasos) *La atención por pasos es un tipo de atención dispersa donde, dada una entrada con entidades ordenadas $x_1, ..., x_n$, se obtienen sus representaciones como:*

$$h_i = \sum_{t \leq j \leq i} \alpha(x_i, x_j) \psi_v(x_j), \tag{7}$$

donde $t = \max\{0, i-k\}$ para una constante k y $\alpha(x_i, x_j)$ es como en la Ecuación 4.

Otra forma general de obtener mecanismos de atención es asumir que las relaciones entre las entidades de cada ejemplar son arbitrarias, y están definidas por la matriz de adyacencia de una gráfica [28].

Definición 9 (Atención en gráficas) *Dadas las entidades $\{x_1, .., x_n\}$ de un dato con información sobre sus vecinos $\mathcal{N}_i$ para toda i, un mecanismo de atención en gráficas es una capa de la forma:*

$$h_i = \sum_{j \in \mathcal{N}_i} \alpha(x_i, x_j) \psi_v(x_j), \tag{8}$$

donde $\alpha(x_i, x_j)$ son los pesos de atención como en la Ecuación 4.

La atención dispersa y la atención en gráficas se diferencian en que la atención dispersa asume relaciones arbitrarias para todos los ejemplares del dominio de datos, mientras que en la atención en gráficas, las relaciones dependen de cada dato particular. Este último mecanismo es la forma más general de un mecanismo de atención de donde se pueden derivar los mecanismos definidos anteriormente.

2.4. Atención como núcleo generalizado

Tsai et al. [26] proponen clasificar los mecanismos de atención con base en un núcleo (*kernel*), centrado en la función $\alpha(x_i, x_j)$ de la Ecuación 4. Bajo esta perspectiva, los pesos de atención dependen de una función $k : X \times X \to \mathbb{R}$, donde X es el espacio de rasgos para los mecanismos de atención (tanto rasgos posicionales como no posicionales [27]). En este marco, $k(\cdot, \cdot)$ es un núcleo, que en los mecanismos de atención es exponencial, y ya que las proyecciones ψ_q, ψ_k no son simétricas en general, se considera a k como un núcleo generalizado (no simétrico).

La visión basada en kenerls [26] no considera los sesgos inductivos que pueden existir dentro de estos mecanismos. En lo que sigue, asumimos que los pesos de atención $\alpha(x_i, x_j)$ están determinados por el softmax de los productos $k(x_i, x_j)$ bajo un núcleo generalizado. Esto es:

$$\alpha(x_i, x_j) = Softmax_K\big(K_{i,j}\big), \tag{9}$$

No abordamos a profundidad las consecuencias que tiene el uso de diferentes tipos de núcleos [26], puesto que nuestra propuesta se enfoca a sesgos inductivos relacionales. Aunque cabe señalar que esta visión daría pie a sesgos relacionales no necesariamente binarios. Esta es una perspectiva que dejamos como trabajo a futuro. Señalamos que nuestra propuesta no es contraria, sino complementaria a la visión de estos mecanismos en el marco de los núcleos generalizados.

3. Resultados

A partir de la revisión de los diferentes mecanismos de atención (Sección 2.3) se pudo observar que estos mecanismos comparten un núcleo general en donde se tiene una agregación de forma aditiva de los valores de entidades vecinas ponderados por una probabilidad generalmente estimada por una función softmax. Este hecho ya introduce un sesgo inductivo relacional dentro de los mecanismos de atención.

Proposición 1 (Sesgo de relaciones estocásticas) *Las capas de atención asumen que las relaciones entre las entidades $x_1,...,x_n$ son estocásticas.*

La matriz de atención, que denotaremos como α, define la matriz estocástica asociada, en donde cada entrada $\alpha_{i,j} := \alpha(x_i,x_j) = p(x_j|x_i)$. Esta probabilidad está dada en términos de la función softmax (Ecuación 9).

Cabe señalar que los mecanismos de atención se enmarcan dentro de capas de la forma expuesta en la Ecuación 2 (Definición 4): i) la función de mensaje está determinada como $\psi(x_i,x_j) = \alpha(x_i,x_j)\psi_v(x_j)$; ii) la agregación se da por medio de la suma sobre todas las entidades de un dato; y iii) la función de actualización obtiene las nuevas representaciones como la agregación de los mensajes. Resaltando el marco gráfico y estocástico, proponemos una definición general de capa de atención.

Definición 10 (Capa de atención) *Una capa de atención es un tipo de capa gráfica (Definición 4) que estima la representación de un conjunto de entidades $\{x_1,x_2,...,x_n\}$ con sistema de vecindades $\{\mathcal{N}_i : i = 1,2,...,n\}$ a partir del valor esperado sobre una distribución p de estas vecindades:*

$$h_i = \mathbb{E}_{p\sim\mathcal{N}_i}\big[\psi_v(x)\big],$$

Donde la esperanza se estima sobre las relaciones de una matriz de adyacencia dada como (Ecuación 9):

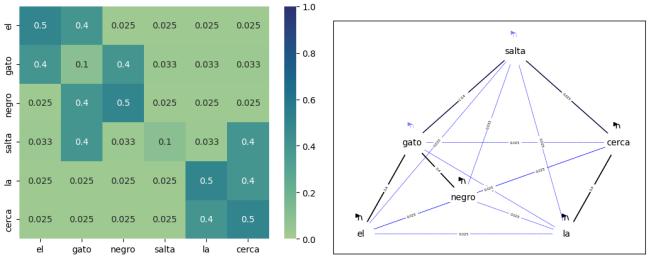
$$\alpha(x_i,x_j) = Softmax_K(K_{i,j}).$$

$\psi_v(x)$ es una proyección de los datos sobre el espacio de valores.

En lo que sigue, nos basamos en esta definición para mostrar los sesgos inductivos relacionales a los que responde cada uno de los mecanismos de atención. En particular, nos enfocaremos en la forma en que las relaciones se manifiestan en la matriz de atención, pues, como hemos señalado, es esta matriz la que determina las relaciones entre entidades. Para esto, nos centramos en los procesos de enmascaramiento.

Lemma 1. *Las relaciones que subyacen al dominio de datos en un mecanismo de atención, se manifiestan en la matriz de atención como un proceso de enmascaramiento.*

Demostración. El proceso de enmascaramiento consiste en eliminar ciertas entradas de la matriz de atención α. Sea K la matriz de productos (Ecuación 9) al que se aplica la función softmax. Una entrada se enmascara a partir de la asignación $K_{i,j} = -\infty$ antes de aplicar el softmax, de tal forma que se tiene que $\alpha_{i,j} = 0$. Claramente esto representa



(a) Matriz de atención (b) Gráfica relacional

Fig. 1. Ejemplo de las relaciones establecidas por medio de un mecanismo de auto-atención.

una desconexión en la estructura gráfica relacional. Si $G = (V,E)$, podemos fácilmente definir el proceso de enmascaramiento como:

$$K_{i,j} = \begin{cases} k(x_i,x_j) & \text{si } (i,j) \in E \\ -\infty & \text{si } (i,j) \notin E. \end{cases}$$

De esta forma, el enmascaramiento está determinado por el tipo de relaciones que se asumen en el dominio de datos.

Teorema 1 (Sesgo relacional en auto-atención) *Las capas de auto-atención asumen un sesgo inductivo relacional con base en una gráfica completamente conectada [6].*

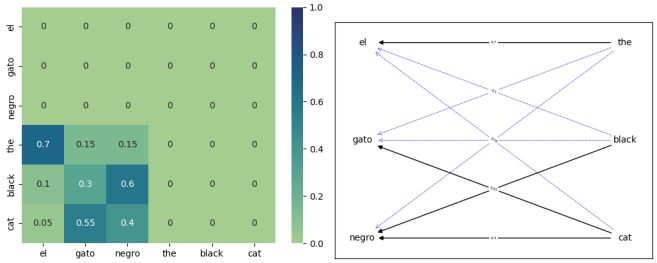
Demostración. En una capa de auto-atención (Definición 5), la agregación, que en las redes de transmisión de mensajes se aplica sobre los vecinos en la gráfica del nodo que representa a la entidad x_i, se aplica sobre todas las entidades de entrada.

Con base en el Lema 1, si K es la matriz de productos, es claro que para todo x_i, x_j entidades de los datos de entrada se tiene que $K_{i,j} = k(x_i,x_j) > -\infty$, por lo que en la matriz de atención $\alpha_{i,j} > 0$. Esto implica que no se realizan desconexiones en la gráfica subyacente; es decir, se asume una gráfica completamente conectada.

Por tanto, las capas de auto-atención asume un sesgo inductivo relacional donde las entidades de un datos se relacionan todas entre sí, esto es, $E = X \times X$ (véase la Figura 1). Con respecto a las capas de atención codificador-decodificador (Definición 6), éstas se conforman a partir de una agregación con respecto a las entidades de entrada y buscan únicamente representar entidades de salida.

Teorema 2 (Sesgo relacional atención codificador-decodificador) *Las capas de atención codificador-decodificador asumen un sesgo inductivo relacional con base en una gráfica bipartita.*

Demostración. En una capa de atención codificador decodificador se tiene un conjunto de datos $x_1, x_2, ..., x_n, y_1, y_2, ..., y_m$ con una partición de las entidades $X = \{x_1, ..., x_n\}$ y



(a) Matriz de atención (b) Gráfica relacional

Fig. 2. Ejemplo de las relaciones establecidas por medio de un mecanismo de atención codificador decodificador.

$Y = \{y_1, ..., y_m\}$ donde $X \cap Y = \emptyset$. La matriz de productos K entonces está determinada de la siguiente forma:

$$K_{i,j} = \begin{cases} k(y_i, x_j) \text{ si } y_i \in Y \wedge x_j \in X \\ -\infty \text{ en otro caso.} \end{cases}$$

Esto es, $\alpha_{i,j} > 0$ si y sólo si $x_i \in X$ y $y_j \in Y$, por lo que la matriz de atención sólo conecta el bloque inferior izquierdo (Figura 2).

La partición está determinada por las entidades del dato de entrada y del dato de salida; generalmente, se trata de una gráfica dirigida que va de las entradas hacia las salidas. Esta atención se puede aplicar tanto a transformadores como a redes recurrentes.

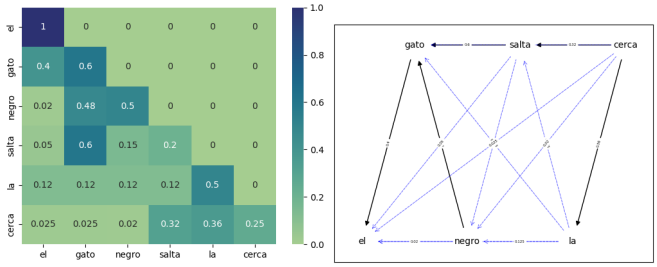
El tercer mecanismo a revisar es la atención enmascarada (Definición 7). Estos tipos de atención asumen un orden en las entidades de entrada, de tal forma que, al pensarse como una gráfica, una conexión se da entre dos nodos si y sólo si una de las entidades precede a otra en este orden.

Teorema 3 (Sesgo relacional en atención enmascarada) *Las capas de atención enmascarada asumen un sesgo inductivo relacional de orden total.*

Demostración. Sea $x = \{x_1, x_2, ..., x_n\}$ el conjunto de entidades de entrada, y defínase un orden $O(x) = \{(x_i, x_{i+1}) : i = 1, ..., n-1\}$ sobre las entidades. Al tomar la matriz de productos K del mecanismo de atención, consideramos el enmascaramiento determinado por las entradas de la matriz en la siguiente forma:

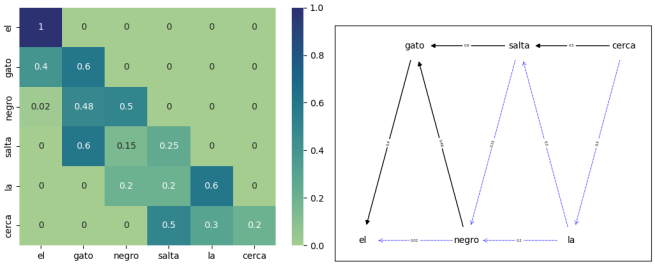
$$K_{i,j} = \begin{cases} k(x_i, x_j) & \text{si } j \leq i \\ -\infty & \text{si } j > i. \end{cases}$$

De tal forma que la matriz de atención tendrá entradas $\alpha_{i,j} \neq 0$ si y sólo si el elemento x_j precede al elemento x_i en $O(x)$, mientras que las otras entradas representan desconexiones en la gráfica. Claramente, las relaciones diferentes de 0 están en la parte triangular superior de la matriz de atención (Figura 3).



(a) Matriz de atención (b) Gráfica relacional

Fig. 3. Ejemplo de las relaciones establecidas por medio de un mecanismo de atención enmascarada.



(a) Matriz de atención (b) Gráfica relacional

Fig. 4. Ejemplo de las relaciones establecidas por medio de un mecanismo de atención por pasos.

Ya que se cumple: i) para toda i, $i \leq i$ por identidad; ii) si $j \leq i$ y $i \leq j$ entonces $i = j$ (de lo contrario habría conexiones en la parte triangular superior de la matriz de adyacencia); y iii) cuando $k \leq j$ y $j \leq i$, entonces $k \leq i$ pues i se conecta con todos los precedentes, podemos concluir que el orden es total.

El tipo de relaciones en las capas de atención enmascarada introducen un sesgo inductivo relacional donde se asume que las entidades no se conectan con elementos subsecuentes. Esto, como es bien sabido, es útil para la representación de secuencias.

Tanto la atención enmascarada como la atención dispersa por pasos (Definición 8) pueden verse como formas de relacionar las entidades de una gráfica únicamente con elementos previos. Esto define una gráfica dirigida cuya matriz de adyacencia cuenta únicamente con la parte triangular inferior. En el caso de la atención por pasos, también se presentan ceros en la pare inferior de la matriz de adyacencia pero acotado a un número dado de elementos previos (véase Figura 4).

Teorema 4 (Sesgo relacional en atención por pasos) *Las capas de atención por pasos asumen un sesgo inductivo donde un elemento se conecta con los p elementos previos para un p fijo y dado un orden previo.*

Demostración. Se puede observar que la atención por pasos es similar a la atención enmascarada y también requieren de un orden $O(x)$ sobre las entidades. Bajo este orden, se pueden definir las entradas de K como:

$$K_{i,j} = \begin{cases} k(x_i, x_j) & \text{si } t \leq j \leq i \\ -\infty & \text{si } t > j > i \end{cases}$$

Definimos aquí $t = \max\{0, i-p\}$ para algún $p \geq 0$. Las relaciones entre las entidades sólo se dan con los t elementos previos, lo que define una gráfica acíclica dirigida (DAG por sus siglas en inglés) que determina el sesgo inductivo relacional en este tipo de mecanismos.

3.1. Sensibilidad a equivarianzas

Para analizar el tipo de equivarianzas ante las cuales son sensibles las capas de atención, nos basamos en subgrupos de permutaciones [8]. Con respecto a las capas de auto-atención tenemos el siguiente resultado ya presentado en [6].

Teorema 5 (Equivarianza en auto-atención) *Las capas de auto-atención son equivariantes ante permutaciones.*

En la atención enmascarada y la atención por pasos, podemos ver que existe una dependencia de un orden previamente establecido.

Teorema 6 (Equivarianza en atención enmascarada) *Las capas de atención enmascarada son equivariantes ante traslaciones.*

Demostración. Una traslación es una función $\sigma(i) = i + m$ para alguna m constante. Como hemos mostrado, las capas de atención enmascarada asumen un orden $O(x)$ sobre las entidades de entrada para crear el enmascarameinto de la matriz de atención α con base en la matriz de productos K. Bajo la traslación es fácil observar que la definición de cada entrada de K:

$$K_{\sigma(i),\sigma(j)} = \begin{cases} k(x_{\sigma(i)}, x_{\sigma(j)}) & \text{si } \sigma(j) \leq \sigma(i) \\ -\infty & \text{si } \sigma(j) > \sigma(i) \end{cases}$$

preserva el orden (si $j \leq i$ entonces $\sigma(i) \leq \sigma(j)$).

Teorema 7 (Equivarianza en atención por pasos) *Las capas de atención por pasos son equivariantes ante traslaciones.*

Demostración. El Teorema 4 introdujo un orden sobre las entidades que, sin embargo, no define relaciones de orden total, pues tenemos que para un p fijo si $j < i - p$ entonces la entidad j no se relaciona con la entidad i, aunque sí puede relacionarse con entidades relacionadas con i, por lo que no tenemos una relación transitiva. Sin embargo, como hemos señalado, la atención por pasos define una gráfica acíclica dirigida (DAG) en donde podemos definir un único orden topológico que hace de esta gráfica un orden total. Por tanto, al igual que con la atención enmascarada, una traslación no modifica las relaciones, pues el orden total resultante con la traslación se preserva.

Corolario 1 *Los sesgos inductivos relacionales de las capas de atención por pasos engloban los de la atención enmascarada.*

La atención por pasos generaliza a la atención enmascarada en tanto basta escoger $t = \text{máx}\{n : n = |x|\}$, es decir, considerar los elementos previos como el valor máximo que puede tener una secuencia.

Teorema 8 (Equivarianza en atención codificador-decodificador) *La atención codificador decodificador es equivariante ante permutaciones en bloques.*

Demostración. Una permutación en bloque es un epimorfismo $\sigma : X \to X$ tal que, dada una relación de equivalencia $\sim$, se tiene que si $x \sim y$ en X, entonces $\sigma(x) \sim \sigma(y)$. Como se señaló en el Teorema 2, la atención codificador-decodificador asume una partición $X \cup Y$ de los datos, de tal forma, que existe una relación de equivalencia $\sim$ que subyace a dicha partición. Claramente si $x_i, x_j \in X$ entonces $x_i \sim x_j$ por lo que $\sigma(x_i) \sim \sigma(x_j)$ y, por tanto, $\sigma(x_i)$ y $\sigma(x_j)$ siguen siendo elementos de X. De forma similar, las permutaciones sobre elementos de Y permanecen en Y. Además los conjuntos X e Y siguen siendo disjuntos ante este tipo de permutaciones, por lo que la partición prevalece.

3.2. Discusión

Proponemos una jerarquía de mecanismos de atención comunes, basada en los sesgos relacionales. Otros trabajos han caracterizado a los transformadores de acuerdo al tipo de arquitectura en que se presentan [18], [19] sin entrar en detalles en las características de la atención. El trabajo de Tsai et al. [26], por su parte, presenta una metodología para la elección de mecanismos de atención, aunque no aborda la forma en que los diferentes tipos de enmascaramiento se relacionan o qué tipo de sesgos introducen.

El Cuadro 1 muestra la clasificación de los mecanismos de atención desde el más general al más particular con base al sesgo relacional y a las equivarianzas que presentan. La atención en gráficas es más general que la atención dispersa. La auto-atención relaciona todas las entidades entre sí, mientras que la atención enmascarada sólo permite relaciones con elementos previos, y la atención por pasos pone una cota al número de elementos previos. Finalmente, la atención codificador-decodificador requiere una bipartición.[3]

La tabla previa también incluye el tipo de datos que estos modelos podrían trabajar. La atención en gráficas es útil para conjuntos de datos donde cada dato tiene una estructura gráfica independiente; por ejemplo, se ha usado para aplicaciones de clasificación de nodos, así como para superficies 3d descritas por triangulaciones; un ejemplo de esto el es modelo de GAT [28]. Las redes de auto-atención suelen usarse en modelos del lenguaje auto-codificados como BERT [11], donde se asume una relación bidireccional. Por su parte, la atención enmascarada se utiliza en modelos auto-regresivos como GPT [10] en los que se asume que se desconoce los elementos siguientes, por lo que sólo hay relaciones hacia atrás. La atención dispersa es utilizada

[3] La implementación de diferentes capas de atención y de transformadores puede encontrarse en: https://victormijangosdelacruz.github.io/MecanismosAtencion/.

Tabla 1. Clasificación de los mecanismos de atención más comunes.

Mecanismo de atención	Relaciones	Grupo de simetrías	Tipo de datos
Atención gráfica	Depende del dato	Dinámica	Con estructuras gráficas por cada dato
Atención dispersa	Arbitraria	Arbitraria	Con relación gráfica específica
Auto-atención	Completamente conectada	Permutaciones	Relacionales o bidireccionales
Enmascarada	elementos previos	Traslación	Secuenciales
Por pasos	p elementos previos	Traslación	Secuenciales acotados
Codificador-decodificador	Bipartición	Sistema de bloques	Particionales

por modelos como Unlimiformer [4] el cuál procesa bloques de texto (acotando en base a la cercanía) y se aplica en clasificación y análisis textuales. Modelos como ViT [12][17] o S4 [14] utilizan atención por pasos en el procesamiento de imágenes como secuencias de tókens, transformando secciones de la imagen o parches. Finalmente T5 [24] incorpora atención codificador-decodificador entre la partición de los datos en elementos de entrada y de salida. Se ha aplicado al lenguaje para traducción o resumen automático. El Cuadro 2 resume estas arquitecturas y sus aplicaciones.

Una de las principales limitantes de los transformadores es que el funcionamiento de estos radica en mecanismos de atención que justamente asumen gráficas completamente conectadas, requiriendo de una gran cantidad de datos para poder aprender relaciones adecuadas. Agregar restricciones a las relaciones que se pueden presentar introduce un sesgo relacional que puede ayudar a mejorar el rendimiento de las arquitecturas basadas en estas capas, siempre y cuando estos sesgos inductivos respondan a la estructura de los datos.

4. Conclusiones y trabajo a futuro

En este trabajo hemos presentado una aproximación teórica a los sesgos inductivos relacionales dentro de los mecanismos de atención. Para esto, nos hemos basado en la teoría del aprendizaje profundo geométrico [6] que nos ha permitido estudiar el tipo de subgrupos de permutaciones ante los cuales éstos mecanismos son equivariantes. Hemos así conformado una clasificación según la suposición que cada mecanismo hace sobre las relaciones subyacentes en las entidades del dominio de los datos. Nuestro análisis proporciona una comprensión más profunda de cómo funcionan los mecanismos de atención y cómo pueden procesar fenómenos complejos, como variaciones sintácticas en lenguas naturales.

Asimismo, extendimos la caracterización de Tsai et al. [26] del enmascaramiento y las relaciones en los mecanismos de atención. Sin embargo, en los mecanismos de atención las relaciones se dan no sólo entre pares de entidades (una gráfica común), sino que pueden existir relaciones de mayor orden (tripletas, cuadripletas, etc.). Las no linealidades en estos mecanismos, como las evidentes desde la perspectiva de los kernels generalizados [26], pueden ser una fuente para dichas relaciones de mayor

Tabla 2. Modelos con uso de atención para lenguaje y visión.

Modelo	**Tipo de Atención**	**Aplicaciones/Efectos**
GAT [28]	En gráficas	Clasificación de nodos, clustering, sistemas de recomendación, modelos 3d.
BERT [11]	Auto-atención	Análisis de sentimiento, clasificación de texto, preguntas y respuestas.
GPT [**?**]	Enmascarada	Generación de texto, traducción automática, resumen automático. Permite el pre-entrenamiento generativo de un modelo grande de lenguaje a través de múltiples tareas de manera no supervisada.
Unlimiformer [4]	Atención dispersa	Análisis de texto, clasificación de texto, procesamiento de lenguaje natural. Permite aprender secuencias de largo alcance con recursos computacionales limitados.
ViT [12]	Por pasos	Clasificación de imágenes, detección de objetos, segmentación de imágenes. Captura patrones de atención espaciales tanto agnósticos (e.g., en primeras capas) como sensibles (e.g., en capas más profundas) al contenido.
T5 [24]	Codificador decodificador	Traducción automática, resumen automático, generación de texto. Permite el entrenamiento del modelo a través de múltiples tareas de manera no supervisada.

orden. Como ejemplo, en el trabajo reciente de [25] se pudo establecer que los transformadores visuales consideran interacciones espaciales de alto orden dentro de cada bloque básico de sus capas. En [15] interpretan la interacción entre tókens de texto en un modelo de transformador, construyendo relaciones jerárquicas y mostrando que éstas interacciones no se dan a pares sino que toman lugar a órdenes mayores.

Como trabajo futuro resulta un reto investigar a fondo las relaciones de mayor orden que pueden darse en éstos mecanismos para poder tener una caracterización completa de los mismos. Otro aspecto relevante es investigar las relaciones que existen con otro tipo de capas como las convolucionales, que pueden verse como un subconjunto de las capas de atención [6]. Ésto puede proporcionar una comprensión más completa de los mecanismos de atención y su papel en las arquitecturas de transformadores.

Agradecimientos. Agradecemos los comentarios de los revisores que ayudarán a la calidad del presente trabajo. También queremos agradecer a los proyectos PAPIIT TA100924 y TA100725 de la UNAM.

Referencias

1. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., et al.: Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:1603.04467, (2016)
2. Bahdanau, D.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473, (2014)

3. Battaglia, P. W., Hamrick, J. B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., Tacchetti, A., Raposo, D., Santoro, A., Faulkner, R., et al.: Relational inductive biases, deep learning, and graph networks. arXiv preprint arXiv:1806.01261, (2018)
4. Bertsch, A., Alon, U., Neubig, G., Gormley, M. R.: Unlimiformer: Long-range transformers with unlimited length input (2023), https://arxiv.org/abs/2305.01625
5. Bradbury, J., Frostig, R., Hawkins, P., Johnson, M. J., Leary, C., Maclaurin, D., Necula, G., Paszke, A., VanderPlas, J., Wanderman-Milne, S., et al.: Jax: composable transformations of python+ numpy programs, (2018)
6. Bronstein, M. M., Bruna, J., Cohen, T., Veličković, P.: Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. arXiv preprint arXiv:2104.13478, (2021)
7. Bronstein, M. M., Bruna, J., LeCun, Y., Szlam, A., Vandergheynst, P.: Geometric deep learning: going beyond euclidean data. IEEE Signal Processing Magazine, vol. 34, no. 4, pp. 18–42 (2017)
8. Cayley, A.: On the theory of groups, as depending on the symbolic equation θn= 1. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, vol. 7, no. 42, pp. 40–47 (1854)
9. Child, R., Gray, S., Radford, A., Sutskever, I.: Generating long sequences with sparse transformers. arXiv preprint arXiv:1904.10509, (2019)
10. Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q., Salakhutdinov, R.: Transformer-XL: Attentive language models beyond a fixed-length context. In: Korhonen, A., Traum, D., Màrquez, L. (eds) Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 2978–2988. Association for Computational Linguistics (2019)
11. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). pp. 4171–4186 (2019)
12. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale (2021), https://arxiv.org/abs/2010.11929
13. Gavranović, B., Lessard, P., Dudzik, A., von Glehn, T., Araújo, J. G. M., Veličković, P.: Position: Categorical deep learning is an algebraic theory of all architectures (2024), https://arxiv.org/abs/2402.15332
14. Gu, A., Goel, K., Ré, C.: Efficiently modeling long sequences with structured state spaces (2022), https://arxiv.org/abs/2111.00396
15. Hao, Y., Dong, L., Wei, F., Xu, K.: Self-attention attribution: Interpreting information interactions inside transformer. In: The Thirty-Fifth AAAI Conference on Artificial Intelligence. AAAI Press (2021)
16. Kisil, V. V.: Erlangen program at large: an overview. Advances in applied analysis, pp. 1–94 (2012)
17. Li, Y., Wang, J., Dai, X., Wang, L., Yeh, C.-C. M., Zheng, Y., Zhang, W., Ma, K.-L.: How does attention work in vision transformers? a visual analytics attempt. IEEE Transactions on Visualization and Computer Graphics, vol. 29, no. 6, pp. 2888–2900 (2023)
18. Liu, Y., Zhang, Y., Wang, Y., Hou, F., Yuan, J., Tian, J., Zhang, Y., Shi, Z., Fan, J., He, Z.: A survey of visual transformers. IEEE Transactions on Neural Networks and Learning Systems, (2023)
19. Lu, D., Xie, Q., Wei, M., Gao, K., Xu, L., Li, J.: Transformers in 3d point clouds: A survey. arXiv preprint arXiv:2205.07417, (2022)
20. Mitchell, T. M.: The need for biases in learning generalizations. Readings in Machine Learning, (1980)

21. Mitchell, T. M., Mitchell, T. M.: Machine learning, vol. 1. McGraw-hill New York (1997)
22. Papillon, M., Sanborn, S., Hajij, M., Miolane, N.: Architectures of topological deep learning: A survey of message-passing topological neural networks. arXiv preprint arXiv:2304.10031, (2023)
23. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems, vol. 32 (2019)
24. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P. J.: Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of machine learning research, vol. 21, no. 140, pp. 1–67 (2020)
25. Rao, Y., Zhao, W., Tang, Y., Zhou, J., Lim, S.-N., Lu, J.: Hornet: efficient high-order spatial interactions with recursive gated convolutions. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. NIPS '22, Curran Associates Inc. (2022)
26. Tsai, Y.-H. H., Bai, S., Yamada, M., Morency, L.-P., Salakhutdinov, R.: Transformer dissection: An unified understanding for transformer's attention via the lens of kernel. In: Inui, K., Jiang, J., Ng, V., Wan, X. (eds) Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 4344–4353. Association for Computational Linguistics, Hong Kong, China (Nov 2019) doi: 10.18653/v1/D19-1443 , https://aclanthology.org/D19-1443/
27. Vaswani, A.: Attention is all you need. arXiv preprint arXiv:1706.03762, (2017)
28. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., Bengio, Y.: Graph attention networks (2018), https://arxiv.org/abs/1710.10903

ISSN 1870-4069

Social Media Sentiment Analysis

Mariana Edith Antonio Aranda, Brenda Sunuami González López,
María Guadalupe Pineda Arizmendi

Tecnológico de Estudios Superiores de Tianguistenco,
Mexico

{mariana.editharanda15, brenda.sunuami.gonzalez.lop,
mariaguadalupe.pineda.arizmendi}@gmail.com

Abstract. The purpose of this paper is to analyze the feelings that people express on social networks such as X (formerly Twitter), Facebook and Instagram. To achieve this, natural language processing techniques and machine learning algorithms were applied. With these tools, the publications were classified into three types of feelings: positive, negative and neutral. Data already available in the English language were used. For the analysis we used the VADER tool (an analyzer based on dictionaries and rules) and an algorithm based on Vector Support Machines. The results show that the model works best with neutral and positive publications, but has some difficulties in identifying negative ones. This type of analysis can be used to better understand public opinion and help make decisions in areas such as communication, marketing or attention to social problems.

Keywords: Machine learning, vader, natural language processing, vector support machines.

1 Introduction

Social networks have become a widely used medium for expressing opinions, emotions and experiences. Every day, millions of people post content that reflects their mood, personal experiences or reactions to social events. This has generated a large volume of data that can be used to understand collective feeling.

Sentiment analysis is a tool that, through the use of natural language processing and machine learning models, identifies whether a publication expresses a positive, negative or neutral opinion. This classification is useful for companies, institutions or researchers who want to know people's views on a particular subject.

On various platforms such as X, Facebook and Instagram users make a lot of posts that reflect different feelings, generating a large amount of textual data that is useful for their analysis.

The problem is that, although these publications contain valuable information, their analysis and efficient classification faces several challenges. For example, the language used in social networks is extremely varied, with informal expressions, abbreviations, emojis, etc., making it difficult to interpret feelings.

The use of natural language processing techniques and machine learning algorithms allows large volumes of data to be automatically processed and classified, identifying whether the post is expressed in categories as positive, negative and neutral.

The general objective of this work is to implement artificial intelligence techniques, such as natural language processing algorithms and machine learning, to analyze publications made on X platforms, Facebook and Instagram and classify them based on three polarities: positive, negative and neutral.

2 Theoretical Framework

Several authors have addressed the analysis of feelings using different techniques. Salgado and Trujillo (2024) used neural networks and SVM and Bayesian classifiers, achieving an accuracy of 80% on Twitter. Lazo and Rodas (2024) evaluated student comments on social networks of universities in Cuenca, finding a predominantly positive perception.

Moreno, Ávila and Ramírez applied techniques such as Python, NLTK and TextBlob to identify business opportunities through sentiment analysis on Twitter. Cardoso, Talame, Amor and Neil (2019) categorized tweets on emotions as fear, anger and happiness using NoSQL databases. Granados (2020) developed a model based on recurrent neural networks and GRU, using the TASS corpus to detect opinion trends in Spanish.

Maldonado (2022) used TextBlob and Tweepy in policy and marketing contexts, evaluating the accuracy of models. Henríquez, Pla, Hurtado and Guzmán (2017) applied SVM along with ontologies to classify opinions on products and services. Aguado et al. (2012) used language rules and tools like Freeling 3.0 to classify emotions in Spanish.

Calvo Madurga (2020) compared models of emotional classification in Spanish using techniques such as Bag of Words and Word Embeddings. Scotto (2021) developed a model based on polarity dictionaries for texts in Spanish.

Rojo, Pollo y Britos (2020) adapted a corpus to the Spanish of Rio de Janeiro in order to improve the analysis of feelings on Twitter. Fernández, Gutiérrez, Gómez and Martínez (2015) created a web application for real-time monitoring of opinions using sentiment dictionaries and machine learning.

3 Methodology

The proposed method describes a detailed process for performing sentiment analysis on social media, ranging from data collection and preparation to the interpretation of results. This process involves several interconnected stages, such as the cleaning and preparation of textual data, its processing using linguistic techniques, and the application of machine learning models to classify the content according to its polarity positive, negative or neutral. (See Fig. 1).

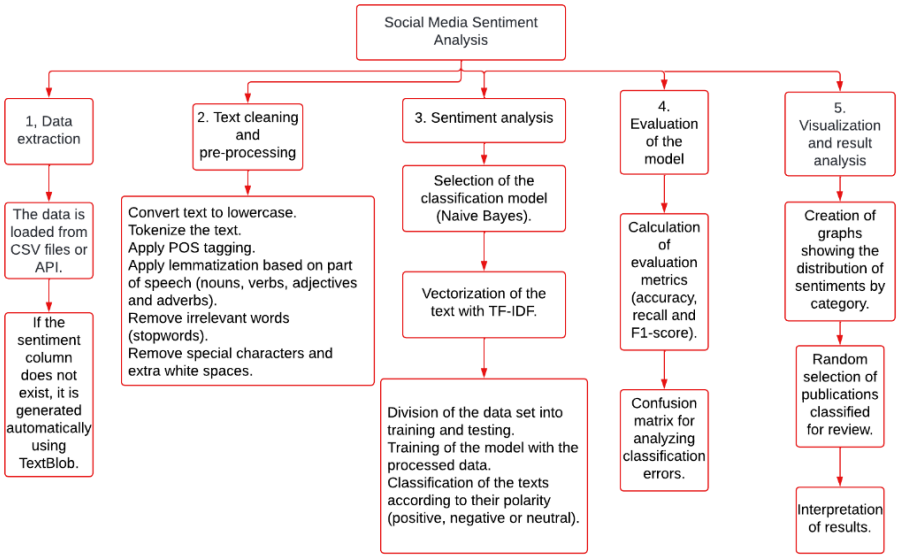


Fig. 1. Diagram of the proposed method.

4 Results

We work with three datasets, each platform consisting of a number of publications as shown in Table 1.

As mentioned before, posts are not tagged, so Vader is used to define the polarity of each post. A small count was made to verify how many positive, negative and neutral posts there are on each platform. Thus, giving the following data shown in Table 2.

The SVM model had an acceptable performance in general, highlighting its effectiveness in the classification of neutral publications, although with some difficulties in identifying negative texts. The evaluation metrics and confusion matrix obtained in the model test are presented below:

Accuracy: 0.60, indicating that approximately 60% of the total predictions were correct.

The model divided the data into 80% for training (2,400 data) and 20% for testing (600 data).

Evaluation metrics such as Accuracy (Precision), Recovery (Recall) and F1-Score were used as show in Table 3. The model makes predictions about the polarity of the texts, as a first result it has the evaluation of metrics show in Table 3.

Precision (Precision): Indicates the proportion of instances classified as a specific category that actually belong to that category.

Negative:0.83 this means that 83% of posts classified as negative are actually negative. Neutral: 0.60 this means that 60% of the publications which were classified as neutral are actually so. Positive: 0.59 this means that 59% of the publications that were classified as positive are actually positive.

Table 1. Number of posts from each social network.

	X	Facebook	Instagram
Number of posts	1,000	1,000	1,000

Table 2. Number of positive, negative and neutral publications on each platform.

	X	Facebook	Instagram
Positive publications	282	474	338
Negative publications	170	149	53
Neutral publications	548	377	609

Table 3. Evaluation metrics.

	precision	recall	f1-score	support
negative	0.83	0.12	0.21	83
neutral	0.60	0.89	0.72	311
positive	0.59	0.36	0.45	206

Recall (Recovery or Sensitivity): Measures the proportion of true instances of a class that the model was able to identify correctly Negative: 0.12 this means that 12% of all really negative posts were correctly identified, Neutral: 0.89 this means that 89% of the really neutral posts were correctly classified. Positive: 0.36 this means that 36% of the really positive posts were rated correctly.

F1-Score: It is the harmonic mean between precision and recall, used when classes are unbalanced.

Negative: 0.21 this suggests that the model does not classify negative posts very well. Neutral: 0.72 this means that there is an acceptable performance in the category. Positive: 0.45 indicating performance similar to the neutral class.

Support: Number of real instances in each category.

- Negative: 83 posts are negative.
- Neutral: 311 publications are neutral.
- Positive: 206 publications are positive.

A confusion matrix was used to analyse classification errors. (See Fig. 2).

Class "negative":

- 10 publications were correctly classified as "negative".
- 55 publications that were actually "negative" were incorrectly classified as "neutral".
- 18 publications that were "negative" were incorrectly classified as "positive".

Class "neutral":

- 278 publications were correctly classified as "neutral".
- 0 publications that were "neutral" were incorrectly classified as "negative".

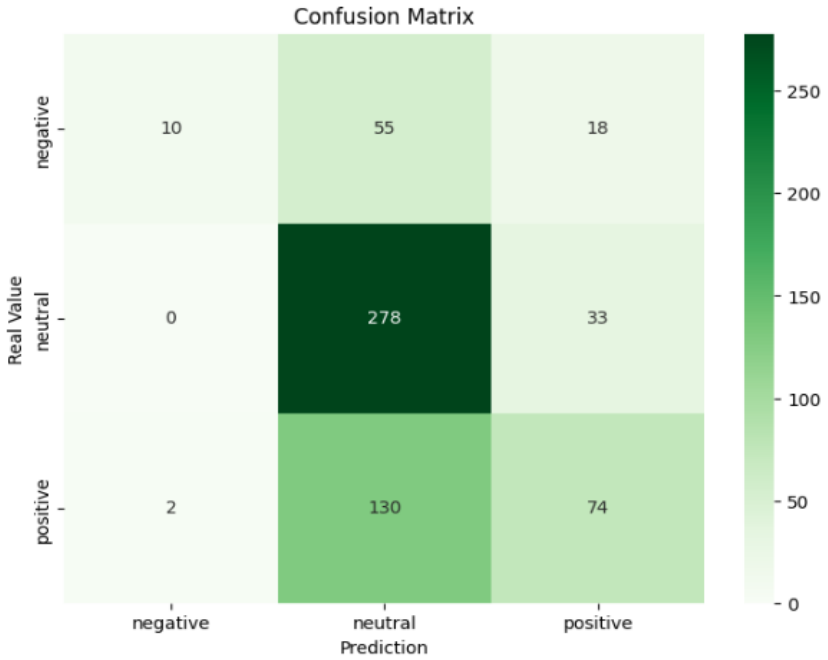


Fig. 2. Confusion Matrix.

- 33 publications that were "neutral" were incorrectly classified as "positive".

Class "positive":

- 74 publications were correctly classified as "positive".
- 130 publications that were "positive" were incorrectly classified as "neutral".
- 2 publications that were "positive" were incorrectly classified as "negative".

As extra information, several publications were shown before and after being preprocessed, in order to show more detail how the preprocessing part works.
The first example will show in detail how preprocessing works, then for the following examples only the original and preprocessed text will be shown without details.

Original text: Hello everybody! Im back on the job, back from the children camp (someone had to look after my sis pupils)... and totally exhausted.

Details of text pre-processing:

- **Lower case conversion:** The text is transformed to lowercase.

hello everybody! i`m back on the job, back from the children camp (someone had to look after my sis' pupils)... and totally exhausted"Tokenization: the text is divided into individual words.

- **Post tagging:** each word is tagged with its grammatical category.

“hello”, “everybody” → Nouns or greetings.

“im”` → Stopword (removed).
“back” → Adverb.
“on”, “the”, “from”, “to” → Stopwords (removed).
“job”, “camp”, “pupils” → Nouns.
“children” → Adjective (headword: “child”).
“someone”, “my”, “sis”' → Stopwords (removed).
“look” → Verb.
“had”, “after” → Stopwords (removed).
“totally” → Adverb.
“exhausted” → Adjective (keyword: “exhaust”).

- **Lemmatization:** words are reduced to their base form.

"children" → "child".
"pupils" → "pupil".
"exhausted" → "exhaust".

- **Words and special characters** that do not contribute meaning for analysis are eliminated.

Irrelevant words such as “im”, “on”, “the”, “from”, “to”, “my”, “and”, “had”, “after” are removed.
Characters such as “!”, “,”, “(”, “)”, “...” and other punctuation marks are removed.
Result of the preprocessed text: hello everybody back job back child camp someone look si pupil totally exhaust.

Examples of X

- Sentiment: positive

  Original text: You are welcome, xuxu

  Processed text: welcome xuxu

  Original Text: mmmmm my hair smells guuud. the wonders of 'pantien' ;D

  Processed Text: mmmmm hair smell guuud wonder

  Original Text: Pepsi throwback, you taste so good in my belly.

  Processed Text: pepsi throwback taste good belly

- Sentiment: negative

  Original Text: Irony: Inventor of Ford Mustang can`t keep his car http://tinyurl.com/lpmvtk via:

  Processed Text: irony inventor ford mustang keep car http via

  Original Text: meh! You should try the one on commercial drive with all the cats

  Processed Text: meh try one commercial drive cat

  Original Text: Consider yourself lucky. It hasn`t rained here in ages. It`s depressing.

  Processed Text: consider lucky rain age depressing

- Sentiment neutral:

Original Text: just woke up, I`m starving

Processed Text: wake starve

Original Text: at the drive ins with daa crewww

Processed Text: drive ins daa crewww

Facebook Examples

- Sentiment: positive

Original Text: Breaking: Edwin Díaz will be suspended 10 games after his ejection for having a foreign substance on his hand, per Jeff Passan. New York Mets | MLB

Processed Text: breaking edwin díaz suspend 10 game ejection foreign substance hand per jeff passan new york mets mlb.

Original Text: Welcome back to another edition of Bet.! Doug Kezirian is joined by Ohm Youngmisuk and Nick Friedell as they preview Game 3 of the NBA Finals. Will the Heat continue to thrive in their role as the underdog? Plus, Ian Parker joins the show to preview UFC 289.

Processed Text: welcome back another edition doug kezirian join ohm youngmisuk nick friedell preview game 3 nba final heat continue thrive role underdog plus ian parker join show preview ufc 289.

- Sentiment: negative

Original Text: Game 4. Denver Nuggets up 4 at the half 👀 Gordon and Jokic lead all scorers with 16 points 🤝

Processed Text: game denver nugget 4 half gordon jokic lead scorer 16 point
Original Text: Breaking: Tonight's Chicago White Sox-New York Yankees and Detroit Tigers-Philadelphia Phillies games have been postponed due to poor air quality in the NY and Philly areas. Both games have been rescheduled for Thursday.
Processed Text: breaking tonight chicago white york yankee detroit phillies game postpone due poor air quality ny philly area game reschedule thursday

- Sentiment: neutral

Original Text Philadelphia Phillies third baseman Alec Bohm will be participating in the 2024 Home Run Derby, the team announced on Friday 🤩

Processed Text:philadelphia phillies third baseman alec bohm participate 2024 home run derby team announce friday

Original Text The Emirates NBA Cup West groups are set 👏

Processed Text:emirate nba cup west group set

Instagram Examples

- Sentiment: positive

Original Text awww @shania.mooree so cute

Processed Text:awww cute

Original Text I have them & they’re so good!!! 😍😍😍

Processed Text:good

- Sentiment: negative

Original Text I have oily skin, and I have such a hard time finding one. I just wanna have a velvety matte skin!! And for not have my makeup separate

Processed Text:oily skin hard time find one wan na velvety matte skin makeup separate

Original Text Why do you half of your plans have an entryway coming into a living room or kitchen with no closet?

Processed Text:half plan entryway come living room kitchen closet

- Sentiment: neutral

Original Text:My fave is demolition

Processed Text: fave demolition

Original Text:Want it

Processed Text: want

Original Text:Do you guys do commercial roofing?

Processed Text: guy commercial roofing

5 Conclusions

The model performs well in the "neutral" category, both in precision and recall. Has difficulties in the "negative" category, where recall is very low (it identifies only 12% of real ones). In the "positive" category, performance is acceptable, but there is still room for improvement. The overall accuracy of the model was approximately 60%, indicating reasonable performance.

This work demonstrates that sentiment analysis using natural language processing (NLP) and machine learning (ML) techniques on various digital platforms is an extremely useful tool for interpreting the emotions expressed by online users. Through this approach, it is possible to gain a clearer and more accurate understanding of users' opinions and attitudes towards certain topics, products, or services. The results obtained show that social media, blogs, forums, and other online platforms constitute a rich and dynamic source of emotional data that can be classified into sentiment categories such as positive, negative, and neutral. This classification is essential for companies, institutions, and analysts looking to understand public reactions and improve their communication or marketing strategies.

The implementation of advanced text processing techniques such as tokenization, lemmatization, and stopword removal significantly improves the quality of text preprocessing, which in turn makes sentiment classification more accurate and effective:

- Tokenization allows the text to be divided into meaningful units such as words, phrases, or even characters, making it easier to identify key elements in the analysis.
- Lemmatization helps reduce words to their base or root form, preventing the distortion of meaning due to morphological variations in words, which improves data understanding.
- Stopword removal, by removing words that do not carry significant semantic value, allows the system to focus on terms that truly influence the expressed sentiment.

When applied correctly, these techniques not only optimize the accuracy of the analysis but also reduce noise in the data, making machine learning models more effective in predicting the polarity of emotions (positive, negative, or neutral) in an automated manner. This type of analysis can be used in a wide range of applications, such as improving customer experience, evaluating advertising campaigns, analyzing public opinions, or even detecting crises on social media. In summary, the use of NLP and ML in sentiment analysis represents a key tool for understanding user emotions and attitudes in the digital age.

References

1. Salgado, N., Trujillo, G.: Sentiment Analysis in Social Network Data: Application of Natural Language Processing and Machine Learning Techniques to Analyze Opinions and Sentiments in Social Network Data in the Context of Information Systems. Dominio de las Ciencias, 10(1), pp. 314–327 (2024) doi: 10.23857/dc.v10i1.3714.
2. Lazo-Calle, A.A., Rodas-Calle, E.L.: Sentiment Analysis in the Social Networks of the Universities of Cuenca. Institutional Repository of the Universidad Politécnica Salesiana (2024) http://dspace.ups.edu.ec/handle/123456789/26900
3. Moreno, L., Ávila, F., Ramírez, A.M.: Business Opportunities. Analysis-of-Feelings-on-Twitter-Social-Networks.pdf
4. Cardoso, A.C., Talame, L., Amor, M.: Opinion Mining: Sentiment Analysis in a Social Network. Institutional Repository of the UNLP. Opinion Mining: Sentiment Analysis in a Social Network (2019)
5. Granados, J.D.: Application of Machine Learning Techniques to Analyze the Polarity of Sentiments in Text to Detect Trends of Opinion on Online Platforms. [Degree project, Santo Tomás de Aquino University, Faculty of Electronic Engineering, Bogotá D.C]. Academic Repository (2020)
6. Maldonado, E.S.: Sentiment Analysis on the Twitter Social Network using Natural Language Processing. [Degree Thesis, National University of Chamborazo, Riobamba, Ecuador]. UNACH Digital Repository: Sentiment Analysis on the Twitter Social Network Using Natural Language Processing (2022)
7. Henríquez, C., Pla, F., Hurtado, L.F.: Sentiment Analysis at the Aspect Level Using Ontologies and Machine Learning. Natural Language Processing, (59), pp. 49-56 (2017)
8. Aguado, G., Barrios, M., Socorro, M.: Sentiment Analysis in a Social Network Corpus. Degree thesis. Polytechnic University of Madrid, Complutense University of Madrid (2012)
9. Calvo Madurga, A.: Analysis of Feelings and Emotions on Social Networks Using ML. [Final degree project, University of Valladolid], UVa. Analysis of Feelings and Emotions on Social Networks Using ML (2020)

10. Scotto, J.: Sentiment Analysis of Opinions on Social Networks Using Natural Language Processing Techniques. Degree thesis, University of Belgrano, Buenos Aires, Argentina, Faculty of Engineering and Computer Technology, Computer Engineering (2021)
11. Rojo, V., Pollo-Cattaneo., Ma. F., Britos, P.: Sentiment Analysis on Twitter: Development of Resources in Rioplatense Spanish from Argentina. Institutional Repository of the UNLP (2020)
12. Fernández, J., Gutiérrez, Y., Gómez, J.: Social Rankings: Visual Analysis of Sentiments in Social Networks. Spanish Society for the Processing of Natural Language, (55), pp. 199–202 (2015)

ISSN 1870-4069

Modelo transformer para la identificación de la nefropatía diabética en pacientes Mexicanos

Luis Ramón Tercero Martínez González, José Adán Hernández Nolasco

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

232h21004@alumno.ujat.mx, adan.hernandez@ujat.mx

Resumen Recientes avances en inteligencia artificial han permitido el desarrollo de modelos de aprendizaje profundo para la identificación de enfermedades. La nefropatía diabética es una de las principales complicaciones de la diabetes mellitus tipo 2, y su identificación es fundamental para prevenir el deterioro renal. Este estudio propone un modelo transformer como apoyo del diagnóstico de la nefropatía diabética en pacientes mexicanos. Es un modelo que utiliza el mecanismo de autoatención, el cual permite capturar relaciones complejas entre variables clínicas. Se aplicaron técnicas de preprocesamiento para garantizar la calidad de los datos y se utilizó una técnica de incrustación para representar similitudes entre variables. Así mismo, fue posible obtener una interpretación del resultado final del modelo, por medio de la extracción de los pesos de atención, obtenidos de los cálculos matemáticos internos del modelo. Además, se logró en la evaluación una precisión de 0.99 % en la clasificación de la enfermedad. Los hallazgos sugieren que este enfoque moderno puede ser una herramienta de apoyo efectiva y confiable para el diagnóstico de la nefropatía diabética en la población mexicana.

Palabras clave: autoatención, incrustaciones, datos tabulares, transformadores.

Transformer Model for the Identification of Diabetic Nephropathy in Mexican Patients

Abstract. Recent advances in artificial intelligence have enabled the development of deep learning models for disease detection. Diabetic nephropathy is one of the main complications of type 2 diabetes mellitus, and its early identification is essential to prevent kidney deterioration. This study proposes a transformer-based model as a clinical decision support tool for diagnosing diabetic nephropathy in Mexican patients. The model leverages the self-attention mechanism, which captures complex relationships among clinical variables. Preprocessing techniques

were applied to ensure data quality, and an embedding technique was used to represent similarities between variables. Furthermore, interpretability was achieved by extracting attention weights from the model's internal computations. The evaluation yielded a classification accuracy of 99%, suggesting that this modern approach can serve as an effective and reliable tool to assist in the diagnosis of diabetic nephropathy within the Mexican population.

Keywords: self-attention, embeddings, tabular data, transformers.

1. Introducción

La inteligencia artificial (IA) es una rama de la informática que permite a las computadoras realizar tareas que normalmente requieren inteligencia humana, se ha convertido en un campo clave para resolver problemas complejos en diversas disciplinas. Además, el aprendizaje automático (AA), es una rama de la IA que permite a las computadoras aprender y mejorar de forma autónoma. Las redes neuronales son una sub rama del AA y si estas tienen un numero grande de capas , se le denomina aprendizaje profundo (AP). EL AP destaca por su capacidad para extraer patrones a partir de grandes volúmenes de información, con impacto en sectores como la salud. Por lo tanto, permite abordar desafíos como el diagnóstico de enfermedades. De manera general, la arquitectura del AP incluyen una capa de entrada para recibir datos, capas intermedias ocultas para procesarlos y una capa de salida que genera el resultado [1].

En particular, las arquitecturas de AP han demostrado un alto rendimiento en la identificación de características relevantes para la predicción de enfermedades, alcanzando en precisión y generalización a los algoritmos de AA; en especifico las arquitecturas basadas en transformer [2], principalmente utilizados en procesamiento de lenguaje natural. Él autor menciona en su estudio, que la arquitectura transfomer del AP ha sido aplicado con éxito para procesar datos tabulares y han mostrado un rendimiento mayor al de los algoritmos de AA como los árboles [3]. Su capacidad para modelar relaciones no lineales sin necesidad de predefinir interacciones entre variables es ideal para el diagnóstico de enfermedades.

Sin embargo, la evolución del transformer para datos tabulares ha avanzado desde redes totalmente conectadas hasta arquitecturas híbridas capaces de manejar su complejidad [4]. Algunas arquitecturas emplean atención secuencial para la selección de características por instancia, mejorando la interpretabilidad, mientras que otros combinan autoatención y atención entre muestras para capturar interacciones complejas y optimizar la escalabilidad [5]. Una estrategia eficiente en este contexto es la integración de mecanismos de atención con perceptrones multicapa (MLP), como en los modelos híbridos, cuya innovación principal radica en la técnica de crear embeddings cuya función es crear vectores numéricos a los datos de entrada, lo que permite representar los datos y capturar relaciones complejas [6].

Dentro de las múltiples aplicaciones de las arquitecturas transformer en el ámbito de la salud, el diagnóstico de nefropatía diabética —una complicación grave asociada a la diabetes mellitus tipo 2 (DM2)— representa un desafío crítico. En países como México, donde la DM2 es una de las principales causas de insuficiencia renal crónica, este problema se ve agravado por la escasez de bases de datos clínicas estructuradas y la limitada adopción de herramientas de inteligencia artificial. Ante esta necesidad, el presente estudio propone una arquitectura basada en transformers capaz de capturar representaciones latentes a partir de datos clínicos tabulares, con el objetivo de mejorar la identificación temprana de factores de riesgo y apoyar el diagnóstico oportuno de la nefropatía diabética.

1.1. Trabajos relacionados

En las últimas décadas, el AP ha tenido buenos resultados en la clasificación de enfermedades crónicas, como la nefropatía diabética, entre otras. Con el fin de proporcionar una visión de las investigaciones más recientes en esta aplicación. El estudio de Zisser et al [7] propone el modelo STRAFE, que explora el uso de transformers para mejorar la predicción del deterioro en pacientes con enfermedad renal crónica (ERC), reportando que el modelo mejora la identificación de pacientes de alto riesgo. Asimismo, la investigación de Arafat et al [8] propone un modelo de aprendizaje profundo para la detección automatizada de ERC, obteniendo resultados con una precisión del 99%, superando otros métodos recientes.

Los autores Singh et al [9] presentan un modelo para la mejora del diagnóstico temprano de la ERC, este trabajo utilizó AP y la técnica de selección de características mediante Recursive Feature Elimination (RFE). Con este enfoque, se identificaron variables clave que fueron utilizadas en distintos modelos de clasificación, logrando que el modelo propuesto superara a SVM, KNN, regresión logística, Random Forest y Naïve Bayes, alcanzando una precisión del 100%.

A nivel internacional, empresas como Amazon, Google, Facebook y NVIDIA han desarrollado arquitecturas híbridas que combinan AP con mecanismo de atención, optimizando el análisis de datos tabulares. Arquitecturas como GatedTabTransformer que incorporan mecanismos de gating para mejorar la selección de características [10], mientras que SAINT introduce atención entre muestras, favoreciendo la interpretabilidad de relaciones inter-muestra [11], ARM-Net emplean neuronas exponenciales y atención para un aprendizaje más eficiente de interacciones explícitas [12], TabTransformer captura interacciones complejas entre variables mediante bloques de atención y feed-forward [13], y TabNet se distingue por su preentrenamiento autosupervisado [14].

Sin embargo, se observa que muchas de las propuestas analizadas se centran en comparar los algoritmos de clasificación, sin considerar la importancia de las características significativas resultantes, la interpretabilidad del modelo y cabe resaltar que ninguno de los mencionados utiliza datos de pacientes mexicanos. Por lo tanto, en este estudio, la arquitectura propuesta es entrenada con datos

de pacientes mexicanos, convirtiéndola en una herramienta clave para el apoyo del diagnóstico de la ND. Asimismo, se observa un crecimiento en el desarrollo de estas arquitecturas híbridas, con una tendencia a usar arquitecturas más sofisticada y eficientes desde 2019 hasta 2023. Esto sugiere que la integración de mecanismos de atención en modelos tabulares ha ido perfeccionándose con el tiempo, priorizando tanto la interpretabilidad como la eficiencia en el manejo de datos estructurados, numéricos y categóricos.

1.2. Objetivo de estudio

Este estudio tiene como propósito identificar a pacientes con DM2 que presentan riesgo de desarrollar ND mediante el uso de arquitecturas basado en transformer. Se emplea un clasificador entrenado con el conjunto de datos DiabetIA [17], el cual fue obtenido a través de un proyecto de acceso abierto del Consejo Nacional de Humanidades, Ciencias y Tecnologías (CONACYT). Los datos provienen del Sistema de Informática de Medicina Familiar (SIMF) del Instituto Mexicano del Seguro Social (IMSS) en Michoacán.

2. Metodología

Siguiendo la metodología CRISP-DM, se extrajeron de la base de datos DiabetIA, los datos crudos de los pacientes con DM2 y asimismo aplicando filtros, los datos de pacientes positivos y negativos con ND, constituyendo así la muestra de estudio. La muestra utilizada incluye información sobre otras comorbilidades asociadas a la DM2, tales como retinopatía, neuropatía y pie diabético. Los datos de la muestra estaban organizados en cinco categorías: demográficos, medicamentos, laboratorios, medidas y diagnóstico médico. A través de esta estructura, el estudio se centra especificamente en la clasificación de pacientes con ND, aplicando modelos basados en Transformers.

2.1. Preprocesamiento de la base de datos

La muestra inicial contenía 239 columnas y 50,936 observaciones, ya etiquetadas y anonimizadas. No obstante, se eliminaron las filas con más del 50% de datos faltantes y la columna CURP, por no aportar valor predictivo al modelo. Como resultado, se conservaron 238 columnas y 19,857 observaciones. La base de datos presentaba un desbalance significativo entre clases, además de valores nulos. Para mitigar el sesgo en el entrenamiento del modelo, se aplicó una estrategia de submuestreo: se seleccionaron las 5,000 observaciones más completas del grupo de pacientes negativos (de un total de 18,689) y se concatenaron con las 3,557 observaciones correspondientes a pacientes positivos. Este conjunto balanceado de 8,557 observaciones se utilizó como base para la fase de entrenamiento y evaluación del modelo.

Además, se realizó un análisis de datos, las distribuciones, correlaciones entre variables, valores atípicos y sesgos de las variables numéricas y categóricas para

comprenderlos y nos ayuda a seleccionar las técnicas y modelo adecuado a usar. Se identificaron valores nulos en variables y en las variables categóricas, varias columnas tenían más del 70% de valores nulos, por lo que fueron eliminadas. La base de datos al final quedo con 49 columnas (25 numéricas y 24 categóricas) y 8,557 observaciones (5,000 negativas y 3,557 positivas).

2.2. Imputación de datos tabulares

Las medidas estadísticas media, mediana y moda, son ampliamente utilizadas para la imputación simple de datos tabulares, pero de acuerdo a la literatura, introducen sesgo en los datos. Por lo cual se decidió utilizar la técnica de *Iterative Imputer* [18], este enfoque trabaja de forma iterativa, modelando las relaciones entre variables mediante un esquema basado en árboles de decisión. De esta manera, cada valor ausente se predice utilizando la información más relevante de los demás variables, refinando progresivamente las estimaciones en cada iteración. Además, la imputación se realizó por separado para pacientes positivos y negativos, centrándose exclusivamente en las variables numéricas, ya que las categóricas estaban completas.

2.3. Selección de características

Se emplearon dos técnicas: LASSO (*Least Absolute Shrinkage and Selection Operator*), efectiva en estudios clínicos [15], y RFE (*Recursive Feature Elimination*) con *RandomForestClassifier*, estas técnicas son utilizadas con éxito en diagnósticos basados en datos clínicos [16]. Se obtuvieron seis grupos de 20, 30 y 40 variables mixtas, respectivamente, para entrenar el modelo en distintos casos de estudio. LASSO tiende a seleccionar más variables categóricas, mientras que RFE favorece las numéricas, eligiendo aquellas con mayor impacto en el rendimiento. Variables como el cultivo de orina, la glucosa en sangre, la creatinina, la hemoglobina y la albúmina aparecieron de manera consistente en todos los grupos, lo que refuerza su relevancia para la clasificación de la ND. La convergencia de ambas técnicas destacan la importancia de ciertos biomarcadores en la clasificación. En la siguiente sección se evalúa el grupo de 20 por Lasso y RFE, ya que este grupo tuvo mejor desempeñó y resultados en el modelo basado en Transformers.

3. Modelo de clasificación transformers

3.1. Arquitectura del modelo

La arquitectura propuesta está basada en el modelo transformer adaptado para el procesamiento de datos tabulares. Esta incorpora tres mecanismos fundamentales: el *Self-Attention*, que asigna pesos a las entradas según su relevancia contextual; la *atención multi-cabeza*, que permite capturar múltiples patrones de interacción de forma paralela; y la *Gated Linear Unit (GLU)*, que

incrementa la capacidad expresiva del modelo al combinar transformaciones lineales con una compuerta sigmoide. En conjunto, estos componentes permiten capturar relaciones complejas y filtrar información irrelevante o ruidosa en los datos.

Para una implementación eficiente, estos mecanismos se integran dentro de una clase denominada *CustomTransformerEncoderLayer*, la cual constituye el núcleo del codificador propuesto. Esta clase se repite en múltiples capas, lo que permite una representación jerárquica y profunda de las variables de entrada.

En la etapa de entrada, el modelo genera representaciones vectoriales (embeddings) tanto para variables categóricas como continuas. Las variables continuas son transformadas a una dimensión fija d mediante capas lineales, mientras que las categóricas son proyectadas a través de embeddings aprendibles. Estas representaciones son combinadas y procesadas en secuencia por el codificador transformer multicapa, seguido por una capa densa encargada de realizar la clasificación.

Con el fin de brindar mayor claridad sobre la estructura matemática del modelo y justificar su implementación, a continuación se presentan las expresiones formales que definen los principales componentes del codificador. Estas ecuaciones describen cómo se realiza el procesamiento secuencial de la información y cómo se integran los distintos mecanismos que conforman la arquitectura del modelo propuesto.

Gated Linear Unit (GLU) La capa GLU mejora la flexibilidad combinando transformaciones lineales con una compuerta sigmoide y su fórmula se muestra en la ecuación 1:

$$o = \sigma(W_z x + b_z) \odot \tanh(W_h x + b_h) + (1 - \sigma(W_z x + b_z)) \odot x, \tag{1}$$

donde:

- x representa el vector de entrada.
- W_z y W_h son matrices de pesos para las transformaciones lineales asociadas a la compuerta sigmoide y a la función tangente hiperbólica, respectivamente.
- b_z y b_h son los vectores de sesgo correspondientes.
- $\sigma(\cdot)$ denota la función sigmoide, que actúa como compuerta para controlar el flujo de información.
- $\tanh(\cdot)$ es la función tangente hiperbólica, que introduce no linealidad.
- $\odot$ indica el producto elemento a elemento (producto Hadamard) entre vectores del mismo tamaño.
- o es el vector de salida de la capa GLU, que combina información filtrada y residual para una representación más expresiva.

Mecanismo de Self-Attention El mecanismo de *Self-Attention* asigna pesos a las entradas según su relevancia y su fórmula se muestra en la ecuación 2:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \tag{2}$$

Donde:

- Q, K y V representan los tensores de consulta (*Query*), clave (*Key*) y valor (*Value*), respectivamente, todos derivados de la entrada X mediante proyecciones lineales.
- QK^T es el producto matricial entre consultas y claves transpuestas, que mide la similitud entre elementos.
- d_k es la dimensión de las claves, utilizada para escalar la similitud y estabilizar los gradientes.
- $\frac{QK^T}{\sqrt{d_k}}$ es la puntuación de atención escalada.
- softmax$(\cdot)$ convierte las puntuaciones en una distribución de probabilidad, asignando pesos de atención.
- Attention(Q, K, V) es la salida del mecanismo de atención, una combinación ponderada de los valores V según los pesos calculados.

Atención Multi-Cabeza Para capturar múltiples patrones, la *atención multi-cabeza* aplica varias atenciones en paralelo y su fórmula se muestra en la ecuación 3:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(H_1, ..., H_h)W^O. \tag{3}$$

Donde:

- Q, K, V son los tensores de consulta (*Query*), clave (*Key*) y valor (*Value*), respectivamente, derivados de la entrada mediante proyecciones lineales.
- $H_1, ..., H_h$ representan las salidas de las h cabezas de atención, cada una operando de manera independiente con sus propias matrices de pesos.
- Concat$(H_1, ..., H_h)$ indica la concatenación de las salidas de todas las cabezas de atención.
- W^O es la matriz de pesos utilizada para proyectar la concatenación de las cabezas de vuelta al espacio de salida original.
- MultiHead(Q, K, V) es la salida final del mecanismo de atención multi-cabeza, que integra múltiples perspectivas de atención para una representación más rica y contextual.

3.2. Selección de hiperparámetros

La tabla 1 presenta los hiperparámetros utilizados en la evaluación final de la arquitectura. La selección de estos valores se basó en una revisión exhaustiva de literatura relacionada, en la que se identificaron configuraciones comunes y efectivas en estudios previos con tareas similares. A partir de esta base, se realizaron ajustes empíricos orientados a maximizar el rendimiento y la eficiencia computacional del modelo.

El entrenamiento de los modelos se realizó utilizando una **TPU v2-8** de Google colab gratuita pero limitada, compuesta por ocho núcleos especializados

Tabla 1. Configuración de hiperparámetros y su justificación.

Hiperparámetro	Valor	Justificación
Características continuas (`col_num`)	n	Depende del dataset, mantiene información relevante.
Características categóricas (`col_cat`)	n	Determina el número de embeddings.
Dimensión del embedding (`dim_embedding`)	128	Equilibrio entre capacidad de representación y sobreajuste.
Cabezas de atención (`num_heads`)	4	Captura dependencias sin aumentar excesivamente el costo computacional.
Capas del codificador (`num_layers`)	4	Profundidad suficiente sin riesgo significativo de sobreajuste.
Dropout (`dropout`)	0.2	Previene sobreajuste manteniendo la capacidad de aprendizaje.
Tamaño del batch (`batch_size`)	64	Balance entre estabilidad del gradiente y eficiencia computacional.
Épocas (`num_epochs`)	10	Pruebas empíricas indicaron convergencia sin sobreajuste.
Tasa de aprendizaje (`learning_rate`)	0.001	Valor estándar en Adam para convergencia estable.

para operaciones tensoriales de alto rendimiento, optimizada para cargas de trabajo en TensorFlow. Esta configuración permitió una aceleración drástica del proceso experimental y en el resultado final.

Gracias a esta capacidad de cómputo especializada, el tiempo de entrenamiento se redujo de forma significativa en comparación con una GPU convencional, completando múltiples iteraciones en cuestión de segundo por época, por lo cual no es factible mencionarlas. Esta mejora sustancial en rendimiento computacional no solo favoreció una exploración más amplia del espacio de hiperparámetros, sino que también ayuda la escalabilidad y reproducibilidad del estudio.

4. Resultados

De los varios grupos de variables creados por los selectores Lasso y RFE, en el análisis final se utilizó solamente el grupo de 20 variables numéricas/categóricas, seleccionadas por selector *Lasso* y *RFE*.

Cabe destacar que las variables seleccionadas a partir de la base de datos inicial contienen información proveniente de cinco grupos médicos principales: demográficos, medicamentos, laboratorios, medidas y diagnóstico médico. Esta diversidad de datos permitió una clasificación más precisa y robusta de los pacientes.

Tabla 2. Reporte de la clasificación usando 20 variables seleccionadas por Lasso.

Class	Precisión	Recall	F1-Score	Support
0	0.9964	0.9982	0.9973	547
1	0.9981	0.9962	0.9971	521
Accuracy			0.9972	1068

4.1. Análisis de variables obtenido por Lasso

Las primeras 11 variables corresponden a exámenes de laboratorio clave en la evaluación metabólica y renal. Las siguientes 2 variables, hemoglobina y albúmina, son biomarcadores importantes. Luego, se encuentran 2 parámetros físicos (peso y pie derecho, que pueden indicar neuropatía diabética), mientras que el grupo de medicamentos incluye una variable para el uso de antidiabéticos. Finalmente, los últimos 4 términos indican el diagnóstico de diabetes tipo 2 y sus complicaciones. En cuanto al desempeño del modelo son Test Loss = 0.0088, AUC = 99.82%, accuracy = 99.99% y specificity = 0.9982%. y además se obtuvo el reporte de la clasificación interna en la tabla 2.

La matriz de confusión:

$$\begin{bmatrix} 546 & 1 \\ 2 & 519 \end{bmatrix}$$

Donde:

- **546** → Verdaderos positivos (TP): Casos correctamente clasificados como sin nefropatía diabética (clase 0).
- **1** → Falsos positivos (FP): Casos incorrectamente clasificados como con nefropatía diabética (clase 1).
- **2** → Falsos negativos (FN): Casos incorrectamente clasificados como sin nefropatía diabética (clase 0).
- **519** → Verdaderos negativos (TN): Casos correctamente clasificados como con nefropatía diabética (clase 1).

Casi no se presentan falsos negativos, el modelo es 98% preciso para no pasar por alto a los pacientes con nefropatía diabética.

Evaluación de los pesos de atención extraídos del modelo

Se muestra en la figura 1, la interpretabilidad de las variables que el modelo prestó más atención para realizar su decisión en la clasificación. Los pesos de atención indican la importancia relativa que el modelo asigna a cada variable al realizar la clasificación.

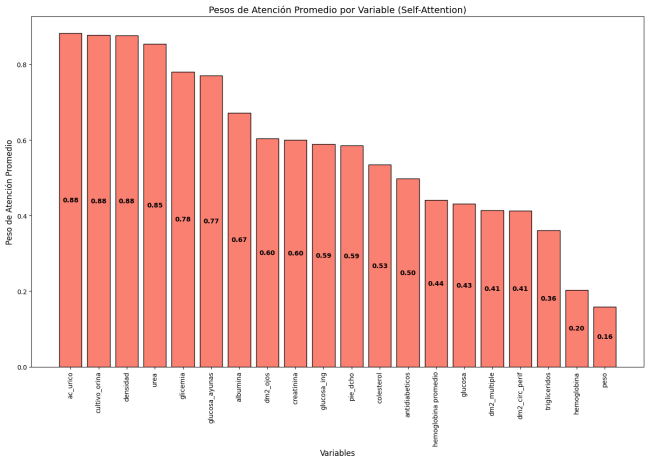


Fig. 1. Grupo de 20 Variables por Lasso y su peso de atención que impactan en los resultados del modelo.

Tabla 3. Reporte de la clasificación usando 20 variables por RFE.

Class	Precisión	Recall	F1-Score	Support
0	0.9982	0.9963	0.9973	547
1	0.9962	0.9981	0.9971	521
Accuracy			0.9972	1068

4.2. Análisis de variables obtenido por RFE

Las primeras variables analizadas, incluyendo colesterol, glucosa, hemoglobina, triglicéridos, urea, creatinina, ácido úrico y cultivo de orina, son fundamentales en la evaluación de riesgos metabólicos y renales, ya que reflejan tanto el control glucémico como la función renal. La densidad urinaria y el examen general de orina (EGO) desempeñan un papel crucial en la identificación de infecciones urinarias y disfunciones renales. Por otro lado, las variables relacionadas con los pies son decisivas para detectar neuropatía diabética y complicaciones circulatorias. Asimismo, biomarcadores como la albúmina y la hemoglobina son indicadores clave de la función renal y el estado hematológico. En cuanto al desempeño del modelo son Test Loss = 0.0077, AUC = 99.99%, accuracy = 99.63%, y specificity = 0.9972% y además se obtuvo el reporte de la clasificación interna en la tabla 3.

Los resultados muestran que el modelo tiene un rendimiento excelente, con una exactitud del 99.72%. La precisión y el recall para ambas clases (0 y 1) son muy altos, lo que indica que el modelo es eficiente en la clasificación correcta de ambas clases, con muy pocos falsos positivos y falsos negativos. En particular, el F1-score de 0.9972 refleja un buen equilibrio entre precisión y recall, lo que

sugiere que el modelo no solo acierta en la mayoría de las predicciones, sino que también es eficaz para identificar tanto las instancias positivas como las negativas.

La matriz de confusión, que muestra el rendimiento de un modelo de clasificación binaria. Aquí está la interpretación:

$$\begin{bmatrix} 545 & 2 \\ 1 & 520 \end{bmatrix}$$

Descripción de los valores:

- **545:** Número de verdaderos positivos (TP). El modelo clasificó correctamente 545 instancias de la clase 0 (negativa).
- **2:** Número de falsos positivos (FT). El modelo clasificó incorrectamente 2 instancias de la clase 0 como clase 1.
- **1:** Número de falsos negativos (FN). El modelo clasificó incorrectamente 1 instancia de la clase 1 como clase 0.
- **520:** Número de verdaderos negativos (TN). El modelo clasificó correctamente 520 instancias de la clase 1 (positiva).

Los resultados del modelo está bien ajustado para no pasar por alto las instancias de la clase 1, que probablemente se asocian con una condición médica importante, como la nefropatía diabética. La combinación de altos verdaderos positivos y bajos falsos negativos refleja un rendimiento excepcional en la clasificación, especialmente en contextos donde la detección de la condición positiva es crítica. Esta interpretación demuestra que el modelo tiene un buen equilibrio entre sensibilidad y especificidad, mejorando la clasificación de la DM2 y minimizando errores.

Evaluación de los pesos de atención extraídos del modelo

En la figura 2 se muestran los pesos de atención reflejan la importancia que el modelo asigna a cada variable al hacer predicciones. En este caso, algunas variables presentan pesos significativamente altos, lo que indica su relevancia en las decisiones del modelo.

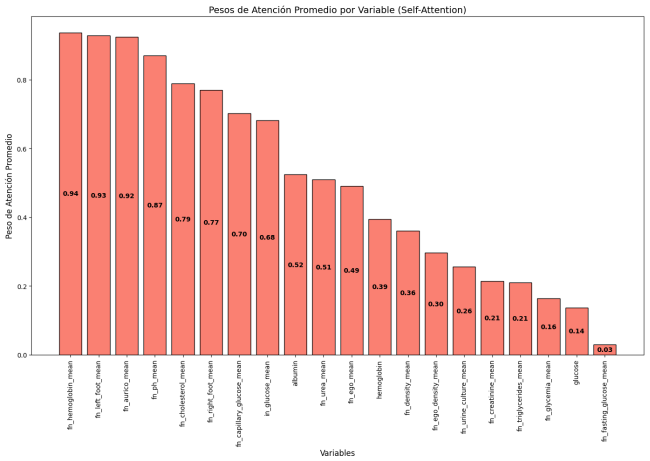


Fig. 2. Grupo de 20 Variables por RFE y su peso de atención para decidir sus resultados.

El modelo combina variables bioquímicas y clínicas, como **hemoglobina**, **urea** y **glucosa**, para tomar decisiones en la clasificación, especialmente en enfermedades renales y metabólicas. La atención elevada a **infección urinaria** resalta su relevancia en el diagnóstico.

4.3. Análisis de los pesos internos extraídos

Para comprender la lógica interna del modelo propuesto y validar su capacidad de interpretación clínica, se analizan los pesos asignados a cada variable durante el proceso de entrenamiento. Estos pesos reflejan la importancia relativa de cada característica en la predicción del diagnóstico de nefropatía diabética. A continuación, se presentan las variables con mayor peso, se muestran en la tabla 4, lo que permite identificar los biomarcadores más influyentes desde el punto de vista matemático y médico. Esta información es clave para establecer la concordancia entre el aprendizaje automático y el razonamiento clínico, y aporta evidencia sobre la robustez del modelo en la toma de decisiones diagnósticas.

Validación Cruzada Los resultados de validación cruzada utilizando 5 pliegues (KFold) muestra un rendimiento consistente del modelo, con una precisión promedio de 0.9947 y una pérdida promedio de 0.0179. Los resultados por pliegue se muestran en el tabla 5:

5. Conclusión y trabajo a futuro

Este estudio evaluó un modelo transformer de clasificación y se utilizo la técnica de validación cruzada para ayudar reducir el sobre ajuste , obteniendo

Tabla 4. Pesos internos de las variables más altas asignados por el modelo en los cálculos matemáticos.

Variable	Peso	Aportación médica
Ácido úrico	0.9369	Alta contribución a la predicción
Glucosa en ayunas	0.9369	Alta importancia en detección
Glicemia	0.9369	Relevancia diagnóstica significativa
Hemoglobina	0.9369	Indicador clave de condición general
Cultivo de orina	0.9279	Soporte en diagnóstico secundario
Urea	0.9245	Alta contribución
pH (orina)	0.8708	Contribución moderada

Tabla 5. Resultados de validación cruzada por pliegue.

Pliegue	Precisión	Pérdida
1	0.9986	0.0046
2	0.9951	0.0167
3	0.9951	0.0180
4	0.9909	0.0334
5	0.9937	0.0171

una precisión de 0.99% y una pérdida de 0.01%. Los resultados muestran su capacidad para generalizar en la clasificación de datos médicos complejos, como enfermedades metabólicas y renales.

El análisis de interpretabilidad del modelo, basado en los pesos de atención, resalta la importancia de variables metabólicas y renales en la clasificación de la nefropatía diabética. Se identificó que biomarcadores como el **ácido úrico**, **urea**, **densidad urinaria**, **glucosa en ayunas** y **glicemia** tienen un impacto significativo en la toma de decisiones del modelo, reflejando su enfoque en la evaluación del estado renal y el control glucémico. Además, variables como el **pH** y la **hemoglobina** demostraron ser determinantes en la predicción del deterioro renal. La alta atención asignada a la **infección urinaria** subraya su relevancia en la detección temprana de alteraciones renales.

Estos hallazgos indican que el modelo no solo alcanza una clasificación precisa, sino que también ofrece una interpretación alineada con el conocimiento clínico, fortaleciendo su aplicabilidad en la detección y monitoreo de la nefropatía diabética. Además de representar un avance metodológico en el uso de inteligencia artificial para el diagnóstico médico, este estudio sienta las bases para el desarrollo de sistemas de IA que optimicen la toma de decisiones en hospitales y unidades de salud pública en México, mejorando así la atención y el pronóstico de los pacientes.

Para finalizar, futuras investigaciones podrían incorporar nuevos tipos de datos, para obtener una arquitectura multimodal, el desarrollo de modelos multimodales en salud ha cobrado interés, integrando datos tabulares, imágenes médicas y texto clínico. Sin embargo, la mayoría han sido entrenados con

datos de otras poblaciones, limitando su uso en México. Nuestra arquitectura transformer es un primer paso hacia modelos adaptados a esta población de interés, facilitando la integración de nuevos tipos de datos sin comprometer la precisión diagnóstica. Este trabajo no solo contribuye a la aplicación del AP en la medicina, sino que también abre oportunidades para mejorar la atención de enfermedades crónicas en México.

Referencias

1. Noor, M.H.M., Ige, A.O.: A Survey on State-of-the-art Deep Learning Applications and Challenges. arXiv print arXiv:2403.17561 (2024). https://arxiv.org/abs/2403.17561.
2. Vaswani, A.: Attention Is All You Need. arXiv print arXiv:1706.03762 (2017). https://arxiv.org/abs/1706.03762.
3. Ye, H.J. : A Closer Look at Deep Learning on Tabular Data. arXiv print arXiv:2407.00956 (2024). https://arxiv.org/abs/2407.00956.
4. Geshkovski, B. : A Mathematical Perspective on Transformers. arXiv print arXiv:2312.10794 (2023). https://arxiv.org/abs/2312.10794.
5. Somvanshi, S. : A Survey on Deep Tabular Learning. arXiv print arXiv:2410.12034 (2024). https://arxiv.org/abs/2410.12034.
6. Gorishniy, Y., Rubachev, I., Babenko, A.: On Embeddings for Numerical Features in Tabular Deep Learning. Adv. Neural Inf. Process. Syst. **35**, 24991–25004 (2022)
7. Zisser, M., Aran, D.: Transformer-based time-to-event prediction for chronic kidney disease deterioration. J. Am. Med. Inform. Assoc. **31**(4), 980–990 (2024).
8. Arafat, F. : A Deep Learning Approach to Predict Chronic Kidney Disease in Humans. In: Proc. IEEE IEMCON, pp. 1010–1015 (2021)
9. Singh, V., Asari, V.K., Rajasekaran, R.: A Deep Neural Network for Early Detection and Prediction of Chronic Kidney Disease. Diagnostics **12**(1), 116 (2022). https://doi.org/10.3390/diagnostics12010116
10. Cholakov, R., Kolev, T.: The GatedTabTransformer: An Enhanced Deep Learning Architecture for Tabular Modeling. arXiv print arXiv:2201.00199 (2022). https://arxiv.org/abs/2201.00199.
11. Somepalli, G. : SAINT: Improved Neural Networks for Tabular Data via Row Attention and Contrastive Pre-training. arXiv print arXiv:2106.01342 (2021). https://arxiv.org/abs/2106.01342.
12. Cai, S., : ARM-Net: Adaptive Relation Modeling Network for Structured Data. In: Proc. 2021 Int. Conf. Manage. Data, pp. 207–220 (2021)
13. Huang, X. : TabTransformer: Tabular Data Modeling Using Contextual Embeddings. arXiv print arXiv:2012.06678 (2020). https://arxiv.org/abs/2012.06678.
14. Arik, S.Ö., Pfister, T.: TabNet: Attentive Interpretable Tabular Learning. In: Proc. AAAI Conf. Artif. Intell. **35**(8), 6679–6687 (2021)
15. Tibshirani, R.: Regression Shrinkage and Selection via the Lasso. J. R. Stat. Soc. Ser. B **58**(1), 267–288 (1996)
16. Guyon, I., Weston, J., Barnhill, S., : Gene Selection for Cancer Classification using Support Vector Machines. Mach. Learn. **46**(1-3), 389–422 (2002)
17. Base de Datos DiabetIA. https://repositorio-salud.conacyt.mx/jspui/handle/1000/296.

18. Scikit-learn developers: InteractiveImputer — scikit-learn 0.24.0 documentation. `https://scikit-learn.org/stable/modules/generated/sklearn.impute.IterativeImputer.html`.
19. Rubin, D.B.: Multiple Imputation for Nonresponse in Surveys. Wiley (1987)

ISSN 1870-4069

Identification and Segmentation of Polyps in the Colon Applying Artificial Intelligence Techniques: A Systematic Literature Review

Valentina Cardenas Moreno[1], Luz Ivana Correa Hernández[2], Carlos Alberto López Herrera[3], Héctor Gabriel Acosta Mesa[3], Efrén Mezura Montés[3]

[1] Universidad Veracruzana, Facultad de Instrumentación Electrónica, Mexico

[2] Universidad Veracruzana, Facultad de Administración y Contaduría, Mexico

[3] Universidad Veracruzana, Instituto de Investigaciones en Inteligencia Artificial, Mexico

emezura@uv.mx

Abstract. Colorectal cancer, the fourth most common type of cancer worldwide, can be prevented through the early identification of colorectal polyps. In this regard, artificial intelligence has emerged as a promising tool to improve the identification, segmentation, and classification of polyps in medical images. The application of these techniques has been shown to aid medical diagnosis due to their accuracy and effectiveness in detecting polyps. This paper reviews the state of the art of existing artificial intelligence techniques, the most common architectures, the datasets used for training, and the metrics employed to evaluate the performance of the proposed models. Finally, opportunities for improvement are highlighted, and future research directions are proposed to optimize artificial intelligence-assisted diagnosis in gastrointestinal health.

Keywords: Colorectal polyps, artificial intelligence, datasets, performance metrics.

1 Introduction

Colon cancer is the fourth most common type of cancer worldwide and the third most common cause of cancer-related deaths [1]. This type of cancer typically develops from untreated colorectal polyps, which are caused by the growth of mucosal epithelial cells [2] and progress slowly. If not properly identified and treated within a period of 10 to 20 years, adenocarcinomatous polyps can lead to colorectal cancer [1].

The timely and early detection of colorectal polyps is essential for the prevention of colon cancer [3]. Therefore, it is crucial to minimize the occurrence of false negatives

during the analysis of a colonoscopy study, a medical procedure in which an instrument is used to film and examine the interior of the colon [4].

In the modern context of Artificial Intelligence (AI), there is no standardized or widely adopted technique in medical practice. Therefore, this systematic review analyzes proposals aimed at implementing computer-aided diagnosis by applying techniques that facilitate the identification, segmentation, and/or classification of polyps for timely detection, thereby exploring more accessible and efficient alternatives.

2 Method

For the methodology of the systematic literature review (SLR), the model proposed by Kitchenham and Charters [5] was used, as it is specialized in computer science and provides a relevant and precise approach. Additionally, being a strict and well-structured model, it helps minimize biases in the selection of papers.

2.1 Justification

The timely detection of colorectal polyps is key to the prevention of colorectal cancer. Therefore, the aim of this review is to understand the state of the art of AI techniques applied to the identification and segmentation of intestinal polyps and how these support tools improve medical diagnosis.

2.2 Research Questions

The following questions were used to guide the research and narrow down the topic.

RQ1. What artificial intelligence techniques have been proposed for the identification and segmentation of intestinal polyps?

RQ2. What are the characteristics (name, type of access, number, format, resolution, and modality) of the image datasets available in the literature for the classification of intestinal polyps?

RQ3. What are the most commonly used performance metrics to evaluate popular AI techniques in the identification and segmentation of polyps?

2.3 Objectives

General objective. To investigate and understand the state of the art in the application of AI techniques for the identification and segmentation of intestinal polyps.

Specific objectives. 1. Identify the existing AI techniques used for the classification of intestinal polyps. 2. Recognize the characteristics of the image datasets available in the literature for the classification of intestinal polyps. 3. Understand the most commonly used performance measures to evaluate the most popular AI techniques.

Table 1. Inclusion and exclusion criteria applied to the papers retrieved with the search strings.

Inclusion criteria	Exclusion criteria
IC1 Studies published from 2019 to September 2024.	**EC1** Studies that are in a language other than English.
IC2 Open Access studies.	**EC2** Studies that are not research papers.
IC3 Studies that contain at least two keywords in the title.	**EC3** Studies whose title refers to polyps outside of the colon.
IC4 Studies that contain at least three keywords in the abstract.	**EC4** Studies conducted with proprietary datasets (or, alternatively, not publicly accessible datasets).
IC5 Studies that, when reading their abstract and/or conclusion, mention the AI technique applied, the datasets used, and the performance achieved.	

2.4 Search

Keywords selection. Based on the research questions and objectives, the necessary keywords were defined, along with their synonyms, as follows: "intestinal polyps," "colon polyps," "polyps," "artificial intelligence," "AI," "machine learning," "performance," "techniques," "classification," and "dataset."

Search strings. The following academic databases were selected: (1) IEEE Xplore Digital Library, (2) Wiley Online Library, (3) SpringerLink, (4) ACM Digital Library, and (5) ScienceDirect, where search strings were tested to finally select two strings, with one being exclusive for ScienceDirect due to its limitation of only supporting eight boolean operators.

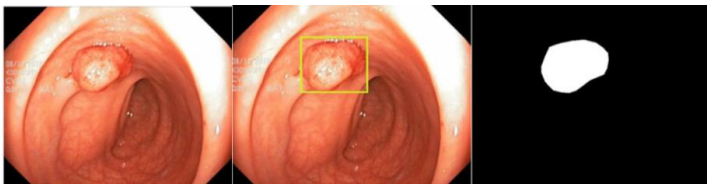
- Search string for IEEE, Wiley, Springer Link and ACM: ("intestinal polyps" OR "polyps" OR "colon polyps") AND ("artificial intelligence" OR "AI" OR "machine learning") AND ("classification" OR "techniques" OR "performance" OR "dataset").
- Search string for ScienceDirect: ("intestinal polyps" OR "polyps" OR "colon polyps") AND ("artificial intelligence" OR "AI") AND ("classification" OR "techniques" OR "performance" OR "dataset").

Inclusion and exclusion criteria. The established criteria are shown below in Table 1.

2.5 Data Selection and Extraction

A total of 1288 papers were identified in the five selected academic databases. After applying the criteria IC1, IC2, EC1, and EC2, the sample was reduced to 161 papers. Subsequently, using the criteria IC3, IC4, IC5, EC3, and EC4, 26 papers were selected for analysis in section 3.

For data extraction, key information was collected such as the paper title, publication date, authors, technique or method used, task type (identification, segmentation, and/or classification), datasets employed, technique description, performance metrics, and



..(a) (b) (c)

Fig. 1. (a) Original image from a colonoscopy; (b) identification with bounding box; (c) segmentation with the generated segmentation mask. [7].

areas of opportunity. Please refer to section 4. Discussion for the link to access the retrieved data from each of the selected papers.

3 Results

In this section, the results found within the 26 selected papers are presented, and the research questions are answered.

3.1 Selected studies and their characteristics

Of the 26 selected papers, 13 belong to IEEE, 7 to Wiley, and 6 to SpringerLink, with publications between 2019 and September 2024. Most of these studies focus on the use of artificial intelligence for medical image analysis to improve the detection of colorectal polyps. In particular, 20 papers explore methods related to the detection and delineation of polyps, while the remaining 6 focus on tasks related to the classification of gastrointestinal diseases.

All the proposed techniques employ artificial intelligence models trained with publicly or privately accessible datasets and are evaluated using various parameters known as performance metrics.

3.2 Identification, Segmentation and Classification

During the review, it was identified that the terms identification and segmentation can be used interchangeably in the medical field. However, in the context of AI techniques, these concepts have more specific differences that should be considered for a better understanding of their application. This was a key finding after the detailed analysis of the reviewed papers.

Identification of colorectal polyps refers to their detection within the image or video obtained during colonoscopy. In this process, the polyp is outlined with a rectangle indicating its location, also known as a "bounding box" [4]; as shown in Figure 1(b).

In contrast, segmentation aims to produce a mask (segmentation mask) that separates the areas of interest (polyps) from the healthy areas (colon), allowing the delineation of the polyp's body and edges. As a result, shown in Figure 1(c), an image is obtained with the areas of interest in white and the rest in black [6].

On the other hand, classification can have two meanings. In the context of medical evaluation, it refers to identifying the type of polyp being treated [8]. However, when focusing on the application of AI techniques, classification is associated with gastrointestinal diseases that may be identified, including polyps as one of the possible conditions of the intestinal tract [9].

3.3 Research Question 1. What Artificial Intelligence Techniques Have Been Proposed for the Identification and Segmentation of Intestinal Polyps?

The techniques reviewed in the 26 analyzed papers follow a common pattern, which is structured into four main components: (1) the architecture, which includes a backbone accompanied by additional processing; (2) preprocessing of the image dataset; (3) training using one or more datasets; and (4) evaluation using performance metrics.

Since there is no standard model and the characteristics of each proposal are so varied, it is only possible to examine the backbone in depth to relate the different methods to one another. Therefore, this research question, in order to be answered, focuses exclusively on the backbone component, as it is the only common characteristic that allows for comparison and evaluation across the methods proposed.

The backbone of the architectures constitutes the process through which features are extracted from the data, with pre-trained Artificial Neural Networks (ANNs) being a common prototype. In this review, three types of models were identified: CNN (Convolutional Neural Network), a type of neural network that uses convolution layers to detect local patterns (such as edges, textures, or shapes) [10]; Transformer, a type of neural network designed to handle data sequences, such as text or time series. It uses attention mechanisms for both global and local approaches [11]; and CNN + Transformer, the integration of both backbones. Figure 2(a) shows the recurrence of each of these backbones in the reviewed papers.

Another notable feature regarding the backbones is the frequency with which each one appears over the years. The most recurrent backbone is CNN, as it is the oldest; however, over time, it has been shown that the use of other architectures provides better performance [12], such as Transformer or the combination of CNN with Transformer, as seen in Figure 2(b). It is worth noting that no papers from 2020 were found in the selection for this review.

In addition to the base architecture, preprocessing is also performed in each of the techniques to process the images from the datasets and unify their characteristics. These modifications, shown in Figure 3, are mentioned in 17 out of the 26 papers, and the following were identified: data augmentation, image normalization, recoloring, pixel resizing, and CLAHE (contrast-limited adaptive histogram equalization). Moreover, in 9 papers, preprocessing is mentioned but not described, as shown in Figure 3 as "Not specified".

The data augmentation is worth mentioning, defined as a technique that involves rotation, scaling, horizontal and vertical flipping, and translation of each image to generate new samples and increase the dataset used for model training [9], in order to mitigate the overfitting problem; this occurs when the model performs exceptionally well on the training set but shows poor performance on the test set or unseen data [13].

(a) (b)

Fig. 2. (a) Recurrence of the identified backbones. (b) Incidence of the use of backbones over the years 2019-2024.

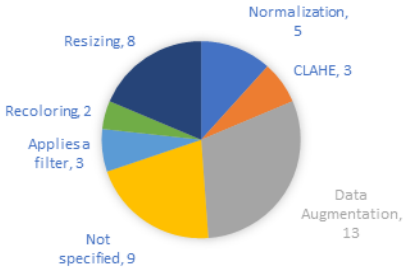


Fig. 3. Modifications made in the preprocessing of the dataset's images.

The datasets, although they will be analyzed in detail in section 3.4 of the review, have very varied characteristics, so each technique must apply this treatment to the images before passing them through the proposed model for training. An important relationship was found between image dimensions and model performance [14]. In Figure 4, we can observe the most common dimensions (pixels) and their frequency of use, as specified in 18 of the 26 selected papers.

3.4 Research Question 2. What are the Characteristics of the Image Datasets Available in the Literature for the Classification of Intestinal Polyps?

The image and video datasets used in the reviewed studies come from colonoscopies. In addition to the polyp image, these datasets include manual annotations made by specialists, who marked the location of the polyp with a bounding box or a segmentation mask, both called "Ground Truth" [10].

For the analysis, only those datasets that were used at least twice within the reviewed papers were considered. These include CVC-ClinicDB, CVC-ColonDB, ETIS-LaribPolypDB, Kvasir-SEG, Kvasir, Kvasir-Sessile, and Hyper Kvasir, with its frequency of appearance shown in Figure 5.

The most frequently used dataset was CVC-ClinicDB, followed by CVC-ColonDB and ETIS-LaribPolypDB. As shown in Table 2, all relevant datasets were created between 2012 and 2021 and are in JPG format, except for CVC-ClinicDB, which is in PNG/TIF format. Most datasets use the WLI (White Light Imaging) modality,

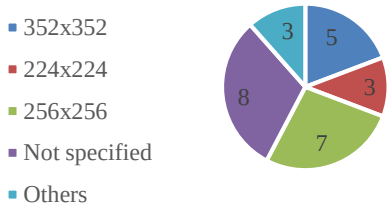


Fig. 4. Frequency of use of the selected dimension in image preprocessing**.**

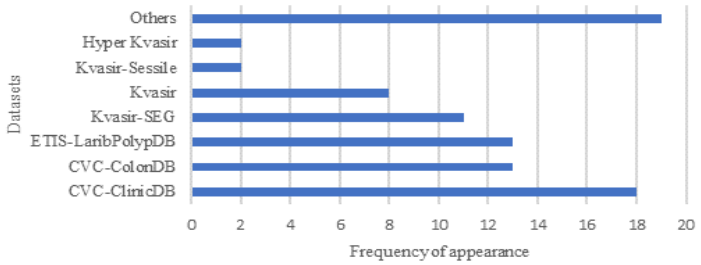


Fig. 5. Frequency of appearance of the most recurrent datasets.

including Kvasir and HyperKvasir, which also incorporate NBI (Narrow Band Imaging). However, CVC-ColonDB and Kvasir-Sessile do not specify the imaging modality.

WLI refers to capturing images using white light, a standard technique in colonoscopy [15]. In contrast, NBI is an advanced imaging technique that enhances the visualization of mucosal and vascular patterns by using narrow-band filters, improving the detection and characterization of lesions [11].

One particularity that can be highlighted about these datasets is the number of images they contain, as it is considered quite limited for the model training stage. This may explain why, in the image preprocessing, data augmentation is applied in 13 of the 26 papers analyzed.

In some studies, such as Pham et al. [16] with “seUNet-Trans” or Saad et al. [17] with “PolySeg Plus”, propose combining multiple datasets to increase the amount of data for both training and evaluation of the model. Additionally, an observable trend in the reviewed papers is the use of the “hold-out” validation method, which involves splitting the dataset into two parts: 80% for model training and 20% for evaluation. This allows for optimized learning and balanced performance measurement.

Since the image dimensions in the dataset do not match those used by the models, this explains the resizing process carried out in the studies during the preprocessing stage to unify their characteristics.

Table 2. Comparison of the most frequent datasets and their characteristics**.**

Name of the dataset	Year	# Images	Modality	Image size	Videos	Format
CVC-ClinicDB	2015	612	WLI	384x288	31	PNG / TIF
CVC-ColonDB	2012	380	-	574x500	15	JPG
ETIS-LaribPolypDB	2014	196	WLI	1226x996	34	JPG
Kvasir-SEG	2020	1000	WLI	332x487 - 1920x1072	-	JPG
Kvasir	2017	8000	WLI/NBI	720x576 - 1920x1072	-	JPG
Kvasir-Sessile	2021	196	-	-	-	JPG
Hyper Kvasir	2020	110,079	WLI/NBI	720x576 - 1920x1072	373	JPG

3.5 Research Question 3. What are the Most Commonly Used Performance Metrics to Evaluate Popular Artificial Intelligence Techniques in the Identification and Segmentation of Polyps?

To answer this research question, it is important to reiterate that the term "classification" refers to a different type of AI technique. While these models also identify polyps, their performance metrics are not directly comparable, as they evaluate the identification of various gastrointestinal diseases rather than solely polyp detection. For this reason, papers focused on classification will not be considered in this section, as they do not present enough common metrics to conduct a proper performance analysis. Consequently, the study will focus exclusively on the remaining 20 papers.

The identified metrics (with their frequency of occurrence in parentheses) from the 20 considered papers were: mDSC (15), mIoU (14), precision (14), recall (12), accuracy (5), F Score (5), F1 Score (2), F2 Score (1), MAE (3), S measure (2), specificity (2), and sensitivity (1). Among these, only the first five were considered to describe, analyze, relate, and evaluate the proposed models.

Each of these metrics evaluates different aspects of the model, making them unsuitable for direct comparison. Instead, they are selected based on the specific needs of each proposed technique.

The metric (a) mDSC (mean Dice Similarity Coefficient) refers to the relationship between the Ground Truth and the model's prediction overlap [10]; (b) mIoU (mean Intersection over Union) represents the average ratio between the predicted area and the actual area [10]; (c) precision measures the proportion of correctly predicted positive cases out of the total positive predictions made by the model [18]; (d) accuracy is the proportion of correct predictions, including both true positives and true negatives, over the total predictions made by the model [19]; and finally, (e) recall describes the proportion of actual positives that were correctly identified [20].

Below are the respective formulas for these performance metrics, where TP (True Positives) are correctly detected polyps, FP (False Positives) are regions mistakenly identified as polyps, TN (True Negatives) are correctly identified non-polyp areas, and FN (False Negatives) are polyps that the model failed to detect:

$$\text{mDSC} = \frac{2TP}{2TP+FP+FN}, \tag{a}$$

$$\text{mIoU} = \frac{TP}{TP+FP+FN}, \tag{b}$$

$$\text{Precision} = \frac{TP}{TP+FP}, \tag{c}$$

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}, \tag{d}$$

$$\text{Recall} = \frac{TP}{TP+FN}. \tag{e}$$

Due to the lack of standardization in model performance evaluation metrics, the obtained assessments for each proposed technique cannot be directly compared. A link is provided in Section 4 for a comprehensive review of each technique and its corresponding performance.

4 Discussion

Among all the proposals in the 26 reviewed papers, a strong connection was identified between the chosen backbone, its evolution over the years, and its performance. The trend shows a preference for Transformer-based models or the combination of CNN with Transformer due to their positive impact on evaluation metrics. For a better understanding of the techniques and the collected information, refer to the annex.

The datasets used in the reviewed studies share the characteristic of being small, as they consist of a limited number of videos and images. Most approaches opted to preprocess the datasets to homogenize their characteristics. The "hold-out" model was used for evaluation; however, given the small size of the datasets, the use of "cross-validation" is recommended, as it provides a more robust performance estimation.

Despite these limitations, the application of AI in endoscopy has gained interest from both academia and industry, leading to the development of commercial solutions that integrate computer-aided diagnosis into clinical practice. Some companies, such as Olympus with EndoBRAIN and EndoAid [21], as well as available systems like EndoMind [22] and CADEYE by Fujifilm [21], have developed and commercialized software for this purpose, though at a high cost. However, EndoMind is open-source software, making it an accessible alternative for researchers and developers. Additionally, various competitions focused on polyp detection [23] have been promoted, enabling the development of multiple datasets and the benchmarking of different methodological approaches.

Moreover, the metrics used to evaluate the proposed AI techniques vary widely, assessing different aspects in each approach, leading to a lack of standardization in truly determining each model's performance. As observed in the recurrence of each metric, not even the most popular ones (mDSC, mIoU, precision, accuracy, and recall) were reported in all the reviewed papers. Additionally, the high performance observed may be influenced by the limited number of images in the datasets; while this does not

constitute overfitting per se, the models might exhibit lower performance in real-world scenarios due to the insufficient number of training samples.

5 Conclusion

After reviewing 26 selected papers using the Kitchenham and Charters methodology, the proposed objectives were achieved, and the research questions were answered.

The proposed techniques, the architecture of each, their backbone, the corresponding preprocessing, the datasets used for training, and the performance metrics used to evaluate these techniques were identified.

From the findings, it is observed that significant challenges remain, particularly the lack of standardization in performance metrics for identification and segmentation, which makes it difficult to make objective comparisons between models performing these tasks. It is recommended to use widely accepted metrics, such as sensitivity and specificity, to complement traditional metrics and improve the validity of comparisons.

Another relevant aspect is the small size of the datasets used, which may influence the high performance reported in the studies. A possible solution for future research would be the creation of larger and more diverse datasets, in addition to implementing advanced data augmentation and transfer learning techniques to improve the robustness of the models.

As future work, considering that some companies already market AI-based software for polyp detection (such as Olympus), a comparative evaluation between these commercial systems and the academic models reviewed is recommended.

References

1. Al-Shamsi, H.O., Iqbal, F., Kourie, H.R.: Colorectal Cancer in the UAE. In: Al-Shamsi, H.O. (eds) Cancer Care in the United Arab Emirates, (2024) doi: 10.1007/978-981-99-6794-0_27.
2. Duc, N.T., Oanh, N.T., Thuy, N.T.: ColonFormer: An Efficient Transformer-based Method for Colon Polyp Segmentation. In: IEEE Access, 10, pp. 80575–80586 (2022) doi: 10.1109/ACCESS.2022.3195241.
3. Peng, Y., Feng, M., Zhai, Z.: PDLFBR-Net: Partial Decoder Localization and Foreground-Background Refinement Network for Polyp Segmentation. In: IEEE Access, 12, pp. 114280–114294 (2024) doi: 10.1109/ACCESS.2024.3445428.
4. Lima, A.C. d. M., Paiva, L.F. d., Bráz, G.: A Two-Stage Method for Polyp Detection in Colonoscopy Images Based on Saliency Object Extraction and Transformers. In: IEEE Access, 11, pp. 76108–76119 (2023) doi: 10.1109/ACCESS.2023.3297097.
5. Kitchenham, B., Charters, S.: Guidelines for Performing Systematic Literature Reviews in Software Engineering Version 2.3. EBSE Technical Report EBSE-2007-01. https://www.researchgate.net/publication/302924724 (2007)
6. Guo, Q., Fang, X., Wang, L.: Polyp Segmentation of Colonoscopy Images by Exploring the Uncertain Areas. In: IEEE Access, 10, pp. 52971–52981 (2022) doi: 10.1109/ACCESS.2022.3175858.
7. Lalinia, M., Sahafi, A.: Colorectal Polyp Detection in Colonoscopy Images Using YOLO-V8 Network. Signal, Image and Video Processing, 18, pp. 2047–2058 (2024) doi: 10.1007/s11760-023-02835-1.

8. Ahamed, M.F., Islam, M.R., Nahiduzzaman, M.: Automated Detection of Colorectal Polyp Utilizing Deep Learning Methods with Explainable AI. In: IEEE Access, 12, pp. 78074–78100 (2024) doi: 10.1109/ACCESS.2024.3402818.
9. Yogapriya, J., Chandran, V., Sumithra, M.G.: Gastrointestinal Tract Disease Classification from Wireless Endoscopy Images Using Pretrained Deep Learning Model. International Journal of Biomedical Imaging, (2021) doi: 10.1155/2021/5940433.
10. Yuan, J., Liu, G., Nam, H.: Polyp Segmentation Based on Multilevel Information Correction Transformer. In: IEEE Access, 12, pp. 91619–91633 (2024) doi: 10. 1109/ACCESS.2024.3421296.
11. Sikkandar, M.Y., Sundaram, S.G., Alassaf, A.: Utilizing Adaptive Deformable Convolution and Position Embedding for Colon Polyp Segmentation with a Visual Transformer. Scientific Reports, 14, Article 7318 (2024) doi: 10.1038/s41598-024-57993-0.
12. Pham, T. -H., Li, X., Nguyen, K. -D.: seUNet-Trans: A Simple Yet Effective UNet-Transformer Model for Medical Image Segmentation. In: IEEE Access, 12, pp. 122139–122154, doi: 10.1109/ACCESS.2024.3451304.
13. Obayya, M., Al-Wesabi, F.N., Maashi, M.: Modified Salp Swarm Algorithm with Deep Learning Based Gastrointestinal Tract Disease Classification on Endoscopic Images. In: IEEE Access, 11, pp. 25959–25967 (2023) doi: 10.1109/ACCESS.2023.3256084.
14. Elkarazle, K., Raman, V., Then, P.: Improved Colorectal Polyp Segmentation Using Enhanced MA-NET and Modified Mix-ViT Transformer. In: IEEE Access, 11, pp. 69295–69309 (2023) doi: 10.1109/ACCESS.2023.3291783.
15. Zhang, M., Sun, Q., Cai, F.: MHA-Net: A Multibranch Hybrid Attention Network for Medical Image Segmentation. Computational and Mathematical Methods in Medicine, (2022) doi: 10.1155/2022/8375981.
16. Hong, L.T.T., Thanh, N.C., Long, T.Q.: CRF-EfficientUNet: An Improved UNet Framework for Polyp Segmentation in Colonoscopy Images with Combined Asymmetric Loss Function and CRF-RNN Layer. In: IEEE Access, 9, pp. 156987–157001 (2021) doi: 10.1109/ACCESS.2021.3129480.
17. Saad, A.I., Maghraby, F.A., Badawy, O.: PolySeg Plus: Polyp Segmentation Using Deep Learning with Cost-Effective Active Learning. Journal of Ambient Intelligence and Humanized Computing, 16, Article 148, (2023) doi: 10.1007/s44196-023-00330-6.
18. Kader, R., Cid-Mejias, A., Brandao, P.: Polyp Characterization Using Deep Learning and a Publicly Accessible Polyp Video Database. Dentomaxillofacial Radiology, (2022) doi: 10.1111/den.14500.
19. Srinivasan, S., Durairaju, K., Deeba, K.: Multimodal Biomedical Image Segmentation Using Multi-Dimensional U-Convolutional Neural Network. BMC Medical Imaging, 24, Article 38, (2024) doi: 10.1186/s12880-024-01197-5.
20. Kang, J., Gwak, J.: Ensemble of Instance Segmentation Models for Polyp Segmentation in Colonoscopy Images. In: IEEE Access, 7, pp. 26440–26447 (2019) doi: 10.1109/ ACCESS.2019.2900672.
21. Kamitani, Y., Nonaka, K., Isomoto, H.: Current Status and Future Perspectives of Artificial Intelligence in Colonoscopy. Journal of Clinical Medicine, 11(10), 2923 (2022) doi: 10.3390/jcm11102923.
22. Lux, T.J., Banck, M., Saßmannshausen, Z.: Pilot Study of a New Freely Available Computer-aided Polyp Detection System in Clinical Practice. International Journal of Colorectal Disease, 37(6), pp. 1349–1354 (2022) doi: 10.1007/s00384-022-04178-8.
23. Ali, S., Ghatwary, N., Jha, D.: Assessing Generalisability of Deep Learning-Based Polyp Detection and Segmentation Methods Through a Computer Vision Challenge. Scientific Reports, 14(1), Artículo 2032, (2024) doi: 10.1038/s41598-024-52063-x.
24. Al-Mekhlafi, Z.G., Senan, E.M., Alshudukhi, J.S.: Hybrid Techniques for Diagnosing Endoscopy Images for Early Detection of Gastrointestinal Disease Based on Fusion Features. International Journal of Intelligent Systems, pp. 1–20 (2023) doi: 10.1155/2023/8616939.

25. Bhattacharya, D., Eggert, D., Betz, C.: Squeeze and Multi-Context Attention for Polyp Segmentation. International Journal of Imaging Systems and Technology, 33 (2022) doi: 10.1002/ima.22795.
26. Chen, L., Ge, H., Li, J.: CrossFormer: Multi-scale Cross-Attention for Polyp Segmentation. IET Image Processing, 17 (2023) doi: 10.1049/ipr2.12875.
27. Nguyen, N. -Q., Lee, S.-W.: Robust Boundary Segmentation in Medical Images Using a Consecutive Deep Encoder-Decoder Network. In: IEEE Access, 7, pp. 33795–33808 (2019) doi: 10.1109/ACCESS.2019.2904094.
28. Raju, A.S.N., Jayavel, K., Rajalakshmi, T.: ColoRectalCADx: Expeditious Recognition of Colorectal Cancer with Integrated Convolutional Neural Networks and Visual Explanations Using Mixed Datasets Evidence. Computational and Mathematical Methods in Medicine, (2022) doi: 10.1155/2022/8723957.
29. Yasmin, F., Hassan, M.M., Hasan, M.: GastroNet: Gastrointestinal Polyp and Abnormal Feature Detection and Classification with Deep Learning Approach. In: IEEE Access, 11, pp. 97605–97624 (2023) doi: 10.1109/ACCESS.2023.3312729.
30. Younas, F., Usman, M., Yan, W.Q.: A Deep Ensemble Learning Method for Colorectal Polyp Classification with Optimized Network Parameters. Pattern Analysis and Applications, 53, pp. 2410–2433 (2023) doi: 10.1007/s10489-022-03689-9.

Annex

				Architecture		Image processing		Evaluation				
Title	**date**	**Academic databases**	**Proposed AI technique**	**Backbone**	**Additional processing**	**Resizing**	**Extra modifications**	**mDSC**	**mIoU**	**Precision**	**Accuracy**	**Recall**
seUNet-Trans: A Simple Yet Effective UNet-Transformer Model for Medical Image Segmentation	sep-24	IEEEXplore	seUNet-Trans	CNN + Transformer	seU-Net + Head Transformer	128 x128	Dataset Combination	0.919	0.85	0.926	NS	0.912
PDLFBR-Net: Partial Decoder Localization and Foreground-Background Refinement Network for Polyp Segmentation	ago-24	IEEEXplore	PDLFBR-Net	Transformer	Cross-level Attention-enhanced Fusion Module (CAFM) + Position Recognition Module (PRM) + Foreground-Background Refinement Module (FBRM)	352 x352	Data augmentation	0.937	0.895	NS	NS	NS
Polyp Segmentation Based on Multilevel Information Correction Transformer	jul-24	IEEEXplore	MICT	CNN + Transformer	Multiscale Feature Extractor (MFE) + Multilevel Lesion Correction Module (MLC) + Feature Selection Fusion Module (FSF)	352 x352	NS	0.8176	0.7422	NS	NS	NS
Automated Detection of Colorectal Polyp Utilizing Deep Learning Methods With Explainable AI	may-24	IEEEXplore	TR-SE-Net	CNN + Transformer	ResNet50 + Squeeze-and-Excitation Blocks	256 x256 y 224 x224	Contrast Limited Adaptive Histogram Equalization (CLAHE) + data augmentation	0.875	0.7961	0.9027	NS	0.8879
Automated Colorectal Polyps Detection from Endoscopic Images using MultiResUNet Framework with Attention Guided Segmentation	abr-24	SpringerLink	MultiResUNet Framework	CNN	MultiResUNet Blocks + Guided Attention (GA)	256 x256	Normalization + Data Augmentation	0.8663	0.8277	0.9364	0.9593	0.806
Utilizing adaptive deformable convolution and position embedding for colon polyp segmentation with a visual transformer	mar-24	SpringerLink	Polyp-ViT	CNN + Transformer	ViT + Adaptive Deformable Convolutional Network (ADCN) + ResNet Blocks	256 x256	Images in JPG format	0.9871	0.9889	0.9827	0.9891	NS
Multimodal Biomedical Image Segmentation using Multi-Dimensional U-Convolutional Neural Network	feb-24	SpringerLink	MDU-CNN	CNN	U-Net + Multidimensional Convolutions + Skip Connections + Convolution Paths	256 x256	NS	NS	NS	NS	NS	NS

				Architecture		Image processing		Evaluation				
Colorectal polyp detection in colonoscopy images using YOLO-V8 network	dic-23	SprinGerLink	YOLO-V8 for polyp detection	CNN + Transformer	YOLOv8 + bounding boxes + Extremely Lightweight Adaptive Networks + Task-aligned One-stage Object Detection	NS	Normalization + Filters (Hue and Brightness) + Data Augmentation	NS	NS	0.956	NS	0.917
GastroNet: Gastrointestinal Polyp and Abnormal Feature Detection and Classification With Deep Learning Approach	sep-23	IEEEXplore	GastroNet	CNN	YOLOv5 + Cross Stage Partial Networks (CSPDarknet) + Neck: PANet (Path Aggregation Network) y SPPF (Spatial Pyramid Pooling - Fast)	416 x416	Normalization + Data (Mosaic) Augmentation	NS	NS	0.99 * (NA)	NS	1 * (NA)
PolySeg Plus: Polyp Segmentation Using Deep Learning with Cost Effective Active Learning	ago-23	SprinGerLink	PolySeg Plus	CNN	Unet/UNet++/ResUNet++ /ResUNet + Locally Shared Features (LSF) + grid search	224 x224	Data Augmentation + Gaussian Filters	0.9476	NS	0.8768	NS	0.9245
CrossFormer: Multi-scale cross-attention for polyp segmentation	jul-23	Wiley	CrossFormer	Transformer	PVTv2 (Pyramid Vision Transformer v2) + CEM (Channel Enhancement Module) + MSCAM (Multi-Scale Cross-Attention Module)	352 x352	NS	0.9249	0.8739	0.9259	NS	0.9437
Improved Colorectal Polyp Segmentation Using Enhanced MA-NET and Modified Mix-ViT Transformer	jul-23	IEEEXplore	Enhanced MA-NET	Transformer	Multi-Scale Attention Network (MA-NET) + Mix-ViT	256 x256	Normalization + CLAHE + CIELAB Color Conversion	0.983	0.973	0.989	NS	0.983
A Two-Stage Method for Polyp Detection in Colonoscopy Images Based on Saliency Object Extraction and Transformers	jul-23	IEEEXplore	SOE DETR	Transformer	Detection Transformer (DETR) + Dense Prediction Transformer (DPT) + Visual Saliency Transformer (VST)	640 x640	Images converted to SGB (Grayscale)	NS	NS	0.932	0.842	0.879
Hybrid Techniques for Diagnosing Endoscopy Images for Early Detection of Gastrointestinal Disease Based on Fusion Features	abr-23	Wiley	VGG-16 / DenseNet-121	CNN	VGG-16/DenseNet-121 + SVM (Support Vector Machines)/Random Forest + PCA	NS	Data augmentation	NS	NS	NS	NS	NS
Modified Salp Swarm Algorithm With Deep Learning Based Gastrointestinal Tract Disease Classification on Endoscopic Images	mar-23	IEEEXplore	MSSADL-GITDC	CNN	CapsNet (Capsule Network) + Class Attention Layer (CAL) + Modified Salp Swarm Algorithm (MSSA) + DBN (Deep Belief Network) + ELM (Extreme Learning Machine)	NS	Median Filter (MF)	NS	NS	0.9216 (NA)	0.9803 (NA)	0.9213 (NA)
Polyp characterization using deep learning and a publicly accessible polyp video database	dic-22	Wiley	CNN Classification	CNN	ResNet101 + Softmax	768 x576	NS	NS	NS	NS	NS	NS
ColoRectalCADx: Expeditious Recognition of Colorectal Cancer with Integrated Convolutional Neural Networks and Visual Explanations Using Mixed Dataset Evidence	nov-22	Wiley	ColoRectalCADx	CNN	ResNet-50/VGG-16/DenseNet-201 + Support Vector Machine (SVM)/Long Short-Term Memory (LSTM) + Grad-CAM	224 x224	Data augmentation	NS	NS	NS	NS	NS
MHA-Net: A Multibranch Hybrid Attention Network for Medical Image Segmentation	oct-22	Wiley	MHA-Net	CNN + Transformer	ResNet-34 + MA-NET + Mix-ViT (Mix Transformer) + Multiscale Attention Module (PSA)	256 x256	Normalization + CLAHE	0.7692	0.8311	0.8618	0.9656	0.8
ColonFormer: An Efficient Transformer Based Method for Colon Polyp Segmentation	ago-22	IEEEXplore	ColonFormer	CNN + Transformer	Mix Transformer (MiT) + UPerNet + Reverse Attention	352 x352	NS	0.927	0.877	NS	NS	NS
Squeeze and multi-context attention for polyp segmentation	jul-22	Wiley	Squeeze and Multi-Context Attention	CNN	U-Net/Attention U-Net/R2U-Net/ResUNet++/R2AU-Net + SMCA (Squeeze Multi-Context Attention)	196 x196, 256 x256 (test) y 512 x512 (training)	NS	0.58	0.47	0.57	NS	0.78
Polyp Segmentation of Colonoscopy Images by Exploring the Uncertain Areas	may-22	IEEEXplore	UnX + FeE	Transformer	Uncertainty eXploration (UnX) + Feature Enhancement (FeE)	352 x352	NS	0.912	0.859	NE	NS	NS
A deep ensemble learning method for colorectal polyp classification with optimized network parameters	may-22	SprinGerLink	Deep ensemble learning method	CNN	GoogLeNet/ResNet-50/Inception-v3/Xception/DenseNet-201/SqueezeNet + transfer learning	NS	Data Augmentation + Data Oversampling	NS	NS	NS	NS	NS
CRF-EfficientUNet: An Improved UNet Framework for Polyp Segmentation in Colon-	nov-21	IEEEXplore	CRF-EfficientUNet	CNN	Unet + CRF-RNN (Conditional Random Field as a Recurrent Neural Network) + EfficientNet B7	NS	Data Augmentation	0.9272	0.8769	0.9492	NE	0.9702

				Architecture		Image processing		Evaluation				
oscopy Images With Combined Asymmetric Loss Function and CRF-RNN Layer												
Gastrointestinal Tract Disease Classification from Wireless Endoscopy Images Using Pretrained Deep Learning Model	sep-21	Wiley	Wireless Endoscopy	CNN	VGG16/ResNet-18/GoogleNet	NS	Data Augmentation	NS	NS	NS	NS	NS
Robust Boundary Segmentation in Medical Images Using a Consecutive Deep Encoder-Decoder Network	mar-19	IEEEXplore	CDED-net	CNN	Deep Encoder-Decoder Networks (DEDNs) + Dicoss-loss	NS	Data Augmentation	0.891	NS	0.95	0.987	NS
Ensemble of Instance Segmentation Models for Polyp Segmentation in Colonoscopy Images	feb-19	IEEEXplore	Ensemble Mask R-CNN	CNN	ResNet-50/ResNet-101 + Two R-CNN Masks	NS	Data Augmentation	NS	0.6946	0.7792	NS	0.7625
								Average of the obtained evaluation				
								0.88168 4615	0.80899 0909	0.89555	0.94435	0.87761 8182

Notes:

- Marked in blue: Refers to the articles specialized in gastrointestinal disease classification.
- Marked in yellow: highest scores achieved on each evaluation metric.
- NS: Refers to “Not Specified”.
- NA: Refers to “Not Applicable”.

ISSN 1870-4069

Phonocardiogram Classification Using Neural Networks for Anomaly Heart Detection

Juan Eduardo Tovar Díaz, Said Polanco Martagón, Marco Aurelio Nuño Maganda, Yahir Hernández Mier, Mario Enrique García Luna

Universidad Politécnica de Victoria,
Mexico

{1830013,spolancom,mnunom,yhernandezm,2439004}@upv.edu.mx

Abstract. Heart diseases have been one of the leading health problems over time, being one of the leading causes of death among the population, affecting people of all ages and social classes. Socioeconomic status, lifestyle, and lack of awareness about symptoms have contributed to its prevalence. In Mexico, in 2022, out of approximately 650,000 deaths, 200,000 were caused by heart disease. In response to this issue, the present research demonstrates the development of a Feedforward Neural Network (FNN) model for detecting cardiac anomalies by analyzing time-domain and frequency-domain features extracted from pre-processed heart sound recordings from the PhysioNet Challenge 2016: Heart Sound Classification dataset. The model demonstrated its effectiveness by achieving a classification accuracy of 99.66% for heart sounds. Amplitude change, pitch shifting, noise addition, noise removal were applied as data augmentation techniques. In addition, SMOTE data balancing techniques were applied to increase the diversity of the dataset and improve the model's performance. Results show that the proposed model can identify complex patterns in heart sound recordings through the extracted features, with an accuracy of almost 99.66%. Although the model performed well, it is important to acknowledge the potential influence of bias on the results, even after applying data augmentation techniques and data balancing techniques.

Keywords: Feedforward Neural Network (FNN), heart sound, data augmentation.

1 Introduction

Heart diseases have been one of the significant problems that medicine has faced over time, as they have been a leading cause of death among the population, affecting people of all ages and social conditions. Factors such as socioeconomic status, lifestyle, and lack of knowledge about symptoms contribute to the proliferation of this problem. In recent years, especially in Mexico, it has been observed that one of the leading causes of mortality in the population is heart disease [4]. Out of about 650,000 deaths in 2022, heart disease was responsible

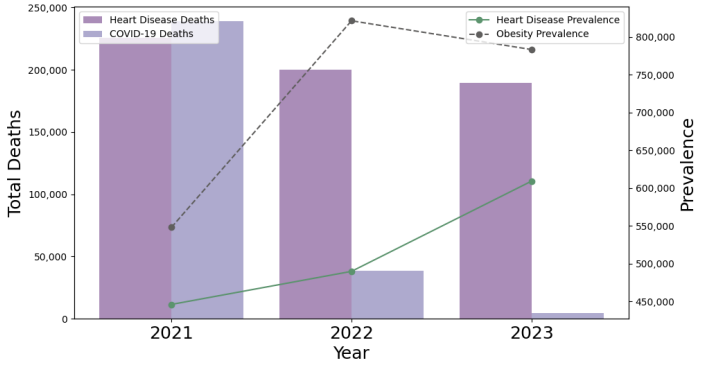


Fig. 1. Comparison of Deaths and Disease prevalence (2021-2023).

for 200,000 of them. Heart diseases significantly impact individuals across all age groups, with a higher incidence observed from the age of 45. To further emphasize the fact that heart diseases are a serious problem, in 2021, at the height of the COVID-19 pandemic, the difference in causes of death between COVID-19 and heart diseases was barely an estimated 13,000 deaths, with COVID-19 totaling nearly 239,000 deaths compared to 226,000 deaths caused by heart diseases [4]. This trend continues in the years 2022 and 2023.

In addition to this issue, the number of people suffering from heart disease has only continued to rise. In Mexico, the number of diagnosed individuals with hypertension increased from 445,993 in 2021 to 548,045 in 2022 [10]. Moreover, the incidence of obesity, a condition linked to hypertension [12], has shown a significant increase. In 2021, 489,731 people were diagnosed with obesity [10], but in 2022, this figure nearly doubled compared to the previous year, reaching 821,255 patients diagnosed with obesity [10]. Although it can be observed that in 2023, the number of patients with obesity decreased to a total of 783,207, those with high blood pressure increased to 609,070 [10]. See figure 1 for a summary of the data.

Various ways have been presented to provide a diagnosis. The most commonly used method is the analysis of electrocardiogram, which interprets the heart's electrical activity [2]. There are other more invasive methods, such as magnetic resonance imaging (MRI) and echocardiograms. However, auscultation, a technique involving using a stethoscope to listen to the sounds produced by the heartbeats [14], is a fundamental and accessible method in diagnosing heart conditions. AI has emerged as a promising tool, showcasing its versatility in various medical applications such as disease diagnosis, treatment optimization, and patient care management. Among the most notable models are neural networks, particularly feedforward neural networks (FNN), which offer an effective method for extracting and classifying information. This research uses feedforward neural networks (FNN) to analyze and classify heart sounds

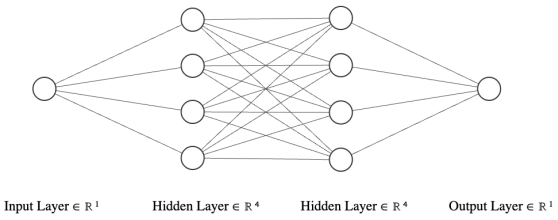


Fig. 2. Structure of a Feedforward Neural Network.

based on extracted features from the PhysioNet Challenge 2016: Heart Sound Classification database. Aiming to enhance the early detection of cardiac problems, the study provides an efficient and accurate tool for an early diagnosis.

2 Theoretical Framework

2.1 Neural Network

A neural network [13] is a network of interconnected nodes designed to generate an output. Biological studies of the nervous system heavily inspire these elements. In short, neural networks are a model of artificial intelligence designed to mimic the functioning of the human brain, assisting machines in learning patterns and making decisions. The elements are represented as "neurons"(N), which, when grouped, form layers to which a numerical value is assigned.

2.2 Feedforward Neural Network

They are artificial neural networks (ANN) in which the connections between neurons do not form a cycle. Feedforward neural networks were the first type of artificial neural network invented. They are called feedforward because information travels only forward through the network (without loops), first through the Input Nodes/Input Layer(IL), then through the Hidden Nodes/Hidden Layers(HL), and finally through the Output Nodes/Output Layer(OL) [16]. Figure 2 shows the structure of a Feedforward Neural Network (FNN).

2.3 Dataset

In this investigation, the PhysioNet/CinC Classifying Heart Sounds Challenge 2016 training dataset was implemented and has been extensively employed in numerous related studies. This dataset tries to represent real-life cases containing phonocardiograms in which the heartbeat is clear and recordings where noise is present. The recordings have a duration of 5 to 120 seconds and were obtained from both healthy and sick patients, including children and adults. This dataset resulted from a combination of nine different databases, and the equipment used to obtain the samples varied across the different databases, as did the environment for each one [8].

2.4 Data Augmentation

It is a set of techniques that generate new information based on limited information. It provides a large amount of data to machine learning models for their training. It also reduces the likelihood that the model will suffer from overfitting while training with the obtained data. Additionally, it helps improve the accuracy and performance of the created models [9].

2.5 Data Balancing

Data balancing encompasses strategies aimed at alleviating class imbalance between different categories. The most commonly employed SMOTE approach is among the many ways to alleviate the issue. This method generates synthetic data from the existing data of the minority class [1].

2.6 Time Domain Features

The simplest way to analyze an audio signal is through time, as it is a time series. Every audio signal evolves, and by visualizing them, we can observe certain key features that help predict and analyze similar signals [17].

Figure 3 shows an example of the main features extracted from the audio signal a0007.wav.

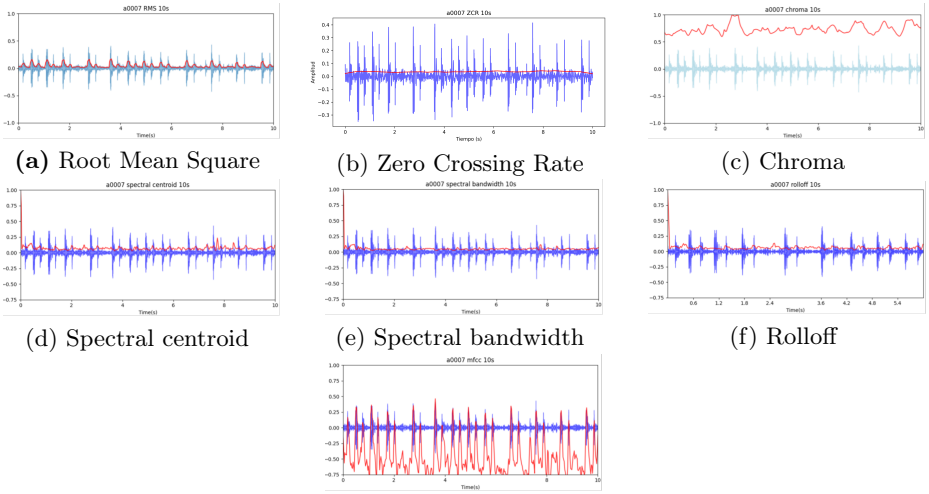
Among these key features, the following will be utilized in this research.

- **RMS (Root Mean Square).** The volume or intensity of a sound is one of the most important characteristics of the human auditory system. Mathematically, volume is defined as the signal's magnitude's root mean square (RMS) value. This feature is used for speech discrimination and music, speech segmentation, and classification of acoustic scenes [9]. See figure 3.a.
- **ZCR (Zero Crossing Rate).** The ZCR can be defined as the zero-crossing rate of an audio signal over a specific time. Mathematically, it is the number of times a signal changes from positive to negative and vice versa, the number of times a signal crosses zero [17]. See figure 3.b.

2.7 Frequency Domain Features

Temporal domain graphs show the variations of the signal over time. The characteristics of the signal are extracted using the Fourier Transform [9] to analyze the signals in frequency.

- **Chroma.** Chromatic characteristics are an audio representation where the entire frequency spectrum of the audio signal is grouped into 12 bins representing the musical octave's 12 semitones (Chroma). This mapping can be obtained through the audio signal's short-time Fourier transform (STFT). These features are particularly useful for identifying the similarity between different interpretations of a musical piece or audio signal [17]. See figure 3.c.



(a) Root Mean Square (b) Zero Crossing Rate (c) Chroma

(d) Spectral centroid (e) Spectral bandwidth (f) Rolloff

(g) MFCC

Fig. 3. Features extracted from the audio a0007.wav. In each subplot, the blue waveform corresponds to the original audio signal, while the red curve represents the extracted feature over time: (a) Root Mean Square (RMS), (b) Zero Crossing Rate (ZCR), (c) Chroma, (d) Spectral Centroid, (e) Spectral Bandwidth, (f) Rolloff, and (g) Mel Frequency Cepstral Coefficients (MFCCs).

- **Spectral centroid.** Indicates where the centroid of the spectrum is located. It describes the "brightness" of an audio signal, calculated by considering the spectrum as a distribution where the values are the frequencies and the probabilities of observing them are the normalized amplitudes. It is used to measure music's timbre and for musical classification [17]. See figure 3.d.
- **Spectral bandwidth.** It is a second-order statistical value that determines low-bandwidth sounds compared to high-frequency sounds [17]. See figure 3.e.
- **Rolloff.** It is the frequency below a certain percentage of the total frequencies of the audio signal, for example, 95% [17]. See figure 3.f.
- **Mel Frequency Cepstral Coefficients (MFCC).** They are the central representation of an audio signal. The MFCCs represent an audio clip's short-term power spectrum based on the power spectrum's discrete cosine transform on a non-linear mel scale. In the MFCCs, frequency bands are equally spaced on the mel scale, closely mimicking the human auditory system [15]. See figure 3.g.

3 State of the Art

The literature review on heart sound classification using feedforward neural networks (FNN) reveals various methodologies proposed for the same purpose.

Krishnan, Balasubramanian, and Umapathy [6] propose the creation of four neural network models: three one-dimensional convolutional neural networks (1D-CNN) and one feedforward neural network (FNN) aimed at classifying unsegmented sounds from a portion of the PhysioNet Challenge 2016: Heart Sound Classification dataset. These sounds are subsequently divided into 6-second sections and used for training the models, with the results considered for the feedforward neural network (FNN), which achieved an accuracy of 82.52%.

On the other hand, Chowdhury, Poudel, and Hu [3] propose a more sophisticated approach than the previous one, using a feedforward neural network to classify sounds from the same dataset. However, they added a smaller dataset composed of 18 audio files. These audio files are segmented using the zero-crossing rate and the Shannon energy envelope, allowing the detection of cardiac cycle beats and segmenting them by each cycle.

After that, the features constituting the Mel spectrogram and Mel-frequency cepstral coefficients are extracted from each audio file and used for training a feedforward neural network (FNN), which achieved an accuracy of 97.10%. Khan, Abid, and Khan [5] use the PhysioNet Challenge 2016: Heart Sound Classification dataset in two different ways: first, with unsegmented audio, and second, with segmented audio. They use the zero-crossing rate for segmentation to determine when a cardiac cycle starts and ends. Subsequently, they use Hidden Markov Models (HMM) and a modified HMM to improve the accuracy when determining the duration of a cycle.

With both unsegmented and segmented information, Mel-frequency cepstral coefficients are extracted, and training is conducted using different types of classifiers, including a Feedforward Neural Network (FNN), which achieved a precision of 79.3% with segmented audio and 80.9% with unsegmented audio. In some studies, although modifications to the neural network's architecture are introduced, the purpose and the data used remain constant.

Li et al. [7] used a Convolutional Neural Network to analyze and classify information from the segmented audio of the PhysioNet Challenge 2016: Heart Sound Classification dataset. Features from the Mel-frequency cepstral coefficients were extracted from the segmented audio, which was then used to train the model, achieving an accuracy of 96.48%. Similarly, Norman, Ting, Salleh, and Ombao [11] propose three models of convolutional neural networks: a one-dimensional Convolutional Neural Network (1D-CNN), which receives the raw audio directly and learns its features; a two-dimensional Convolutional Neural Network (2D-CNN), which learns from features extracted through Mel-frequency cepstral coefficients; and a combination of both models (TF-ECNN).

Focusing on the 1DCNN model, it achieved results already present in the current theoretical framework, reaching 87.23% in accuracy, 87.57% in sensitivity, 85.84% in specificity, and 86.7% in modified accuracy. In Table 1, a summary of the reviewed studies from the current state of the art can be observed.

Table 1. Comparison of the studies reviewed in the state of the art.

Art	Model used	Dataset	Features	Segmentation	Accuracy
[6]	FNN	Part of the PhysioNet Challenge 2016: Heart Sound Classification	Raw audio signal	Sections of 6 seconds	82.52%
[3]	FNN	PhysioNet Challenge 2016 + additional dataset of 18 audio recordings	MFCC	Segmentation by heartbeats cycles using ZCR and Shannon Envelope	97.10%
[5]	FNN	PhysioNet Challenge2016	MFCC	Segmentation by heartbeats cycles using ZCR and HMMS.	79.3% (segmented) 80.9% (Unsegmented)
[7]	CNN	PhysioNet Challenge 2016 + additional dataset of 45 audio recordings	MFCC	Segmentation by 5-second cycles.	96.48%
[11]	1D-CNN	PhysioNet Challenge2016	Raw audio signal	Unsegmented	87.23%

Compared to the studies present in the state of the art, this work differs by the wide range of features extracted (ZCR, RMS, Chroma, Spectral Centroid, Spectral Bandwidth, Rolloff, and MFCC) from each of the audio files in the dataset, making the obtained information more significant and representative for each audio. Furthermore, using a Feedforward Neural Network (FNN) demonstrates that it is possible to classify complex and delicate information without needing a more complex and heavy architecture such as a Convolutional Neural Network. This implements a tool based on this type of model (FNN) that is feasible for use in real-life cases, as a lightweight model allows it to be implemented on various devices.

4 Methodology

4.1 Dataset Modification

The audio files from the dataset are divided into six directories, labeled from a to f (6 folders), these folders were simplified in 2 categorized into "normal" and "abnormal" with a total of 665 abnormal audios and 2575 normal audios respectively. As can be observed, audio files exhibit a significant imbalance, considering the total number of audio files per class, with the "normal" labeled audio files being more dominant.

4.2 Data Augmentation

Data augmentation was performed on both classes for better data balancing. The following techniques were used to generate more audio from existing audio.

- **Amplitude change.** This amplitude change is a multiplication of the audio signal by a random value between the intervals of 0.1 and 10, with a step of 0.1.
- **Pitch shifting.** The pitch change is performed by randomly selecting from a maximum number of semitone steps of 2. The pitch can be altered in both negative and positive directions. For example, if the maximum number of steps is 2, the selected value can range between -2 and 2.
- **Noise Addition.** Gaussian noise is added to the audio at the same length as the original signal. This is done by calculating the audio's energy and normalizing the noise to the desired noise level relative to the audio's energy. The desired noise level is a random value ranging from 0 to 0.5, with a step of 0.001. This approach ensures that the noise level does not exceed appropriate limits, preventing the creation of unhelpful audio samples.
- **Noise Removal.** Bandpass filter with a low cutoff frequency of 25 Hz and a high cutoff frequency of 400 Hz, applied with a second-order filter design.

Each audio file underwent an amplitude change and a combination of amplitude changes with one of the remaining modifications. The combinations were as follows:

- Amplitude change (AC)
- Amplitude change and pitch shifting (ACPS)
- Amplitude change and noise addition (ACNA)
- Amplitude change and noise removal (ACNR)

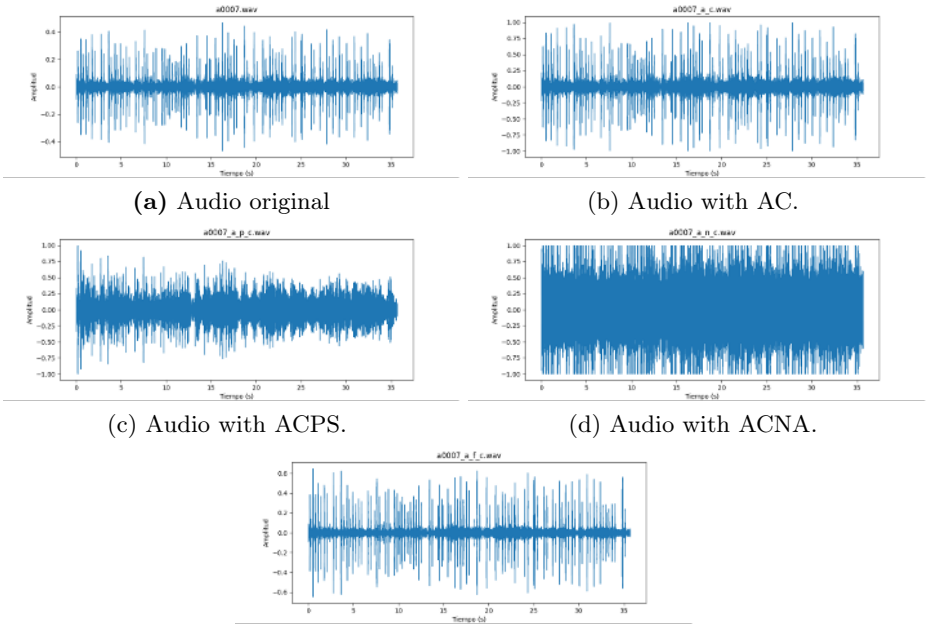
The audio classified as "normal" underwent these data augmentation techniques four times, while the "abnormal" audio was applied eight times. Figure 4.a shows the original audio signal, while figure 4.b shows the signal with the amplitude change. Figure 4.c displays the signal with both amplitude and pitch-shifting techniques. Figure 4.d shows the amplitude change with noise addition, and Figure 4.e illustrates the amplitude change with noise removal.

The newly generated audio files initiated feature extraction from the time and frequency domains. The extracted features were as follows, RMS, ZCR, Chroma, centroid, bandwidth, roll-off, MFCC

This process was carried out using the Librosa library. The mean and variance were calculated for each extracted feature, and the data was stored and categorized in a CSV file. This CSV file will be used to train the feedforward neural network.

4.3 Data Balancing

SMOTE data balancing technique is applied with a random seed of 42 to balance the amount of data labeled as "abnormal" with those labeled as "normal." A



(a) Audio original

(b) Audio with AC.

(c) Audio with ACPS.

(d) Audio with ACNA.

(e) Audio with ACNR.

Fig. 4. Data augmentation of audio a0007.wav.

Table 2. PhysioNet Challenge2016: Heart sound classification Dataset

	Class					
	Original dataset		After data augmenting		After data balancing	
	Normal	Abnormal	Normal	Abnormal	Normal	Abnormal
Total	2575	665	12,441	5,985	12,441	12,441

new CSV file is then created where the data balancing technique has already been applied, and this file is used to train the feedforward neural network.

Table 2 shows the number of samples of each class before and after data augmentation and the bedore and after data balancing.

The test data is located within a specific folder of the PhysioNet Challenge 2016: Heart sound classification Dataset called "validation", which contains 150 normal audios and 151 abnormal audios. The same features were extracted from each of the audios as in the training data for the subsequent creation of a test CSV file.

5 Experimental Results

Four tests were conducted, training different structures of a feedforward neural network, where each test was repeated 10 times through 250 epochs with a learning rate of 0.0001.

Table 3. Neural Network Architectures and optimizers.

Test	Network Structure	Optimizer	Avg training time (min)
1	Sequential (Hidden Layer: 64 N, ReLU) → (Hidden Layer: 32 Neurons, ReLU) → (Output Layer: 2 Neurons, Softmax)	Adam	5 min.
2	Sequential (Hidden Layer 128 N, ReLU) → (Hidden Layer: 64 Neurons, ReLU) → (Output Layer: 2 Neurons, Softmax)	Adam	5 min. 30 sec.
3	Sequential (Hidden Layer: 256 N, ReLU) → (Hidden Layer: 128 Neurons, ReLU) → (Output Layer: 2 Neurons, Softmax)	RMSprop	7 min.
4	Sequential (Hidden Layer: 256 Neurons, ReLU) → (BN) → (D: 0.5) → (Hidden Layer: 128 Neurons, ReLU) → (BN) → (Dropout: 0.5) → (Hidden Layer: 64 N, ReLU) → (Output Layer: 2 Neurons, Softmax)	SGD	8 min.

Table 4. Metrics of the models trained

Test	Accuracy (Acc)	Normal Acc	Abnormal Acc	Min Acc	Max Acc	Mean Acc	F1-score	Spec	Sens
1	.93	.93	.93	.897	.93	.919	.93	.93	.927
2	.98	.99	.97	.93	.98	.969	.983	.97	.99
3	**.9966**	**1**	**.9933**	**.93**	**1**	**.989**	**.996**	**.993**	**1**
4	.906	.966	.846	.88	.906	.894	.915	.84	.96

The following equipment specifications used in the training of the models:

- **CPU:** Intel i3-12100
- **RAM:** 16 GB 3200 MHZ DDR4
- **GPU:** Nvidia GeForce RTX 3060 12GB VRAM

Table 3 shows the structure and optimizers of the models trained. Table 4 details the results obtained from each configuration. As shown, the neural network in test 3 achieves an accuracy of 0.9966 in the test subset, with a mean of 0.989. This contrasts with the current state of the art, showcasing a broader range of extracted features from each audio file, enhancing the depth and representativeness of the information obtained. Figure 6 shows the confusion matrices of the best result of each configuration.

Figure 5 shows the accuracy distribution of all the models trained for each test performed in this study.

6 Conclusions and Future Work

In this work, a feedforward neural network is used to classify the unsegmented heart sounds from the PhysioNet challenge, i.e., each heart cycle of each

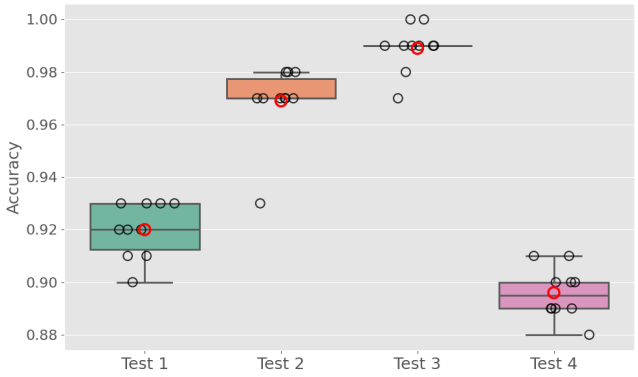
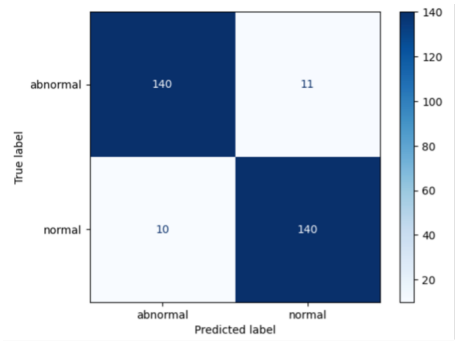
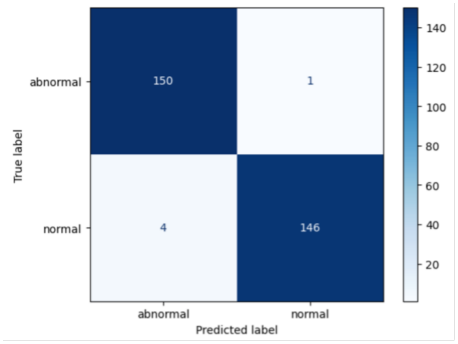


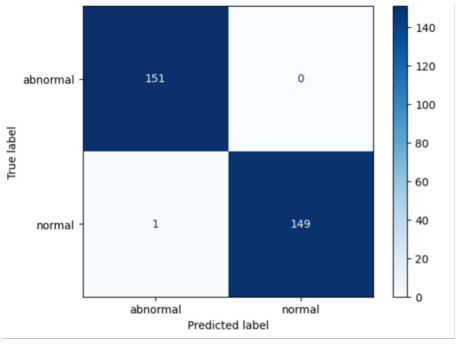
Fig. 5. Distribution of Accuracy Obtained by the Four Neural Network Models Across the 10 Performed Tests.



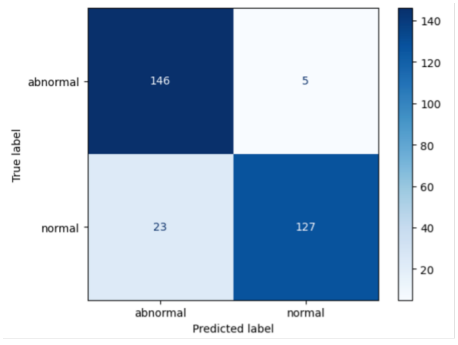
(a) Matrix Confusion of test 1.



(b) Matrix Confusion of test 2.



(c) Matrix Confusion of test 3.



(d) Matrix Confusion of test 4.

Fig. 6. Confusion matrices of the tests performed.

record has not been divided to train the model. Then, according to experiments, feedforward neural networks have demonstrated efficacy in identifying abnormalities within these sounds.

Furthermore, it is shown that feedforward neural networks are nearly on par with more complex and heavier models, such as convolutional neural networks (CNN), making their implementation on smaller and more accessible devices feasible for real-world applications.

The results obtained in the study demonstrate the potential of feedforward neural networks (FNN) for heart sound classification, as well as their ability to classify complex patterns, specially when using data augmentation techniques and extracting a wide range of features extracted from each phonocardiogram in the dataset.

Data augmentation techniques allowed us to enrich samples from the dataset, while data balancing techniques allowed us to balance the minority class results in higher accuracy for the proposed model, exhibiting the capabilities inherent in feedforward neural networks.

This result could help healthcare practitioners to properly access and use non-complex, lightweight models, for efficient and extensive application in practical circumstances such as electronic stethoscope applications, medical mobile apps, or Internet of Things medical applications.

Currently, the proposed technique has some limitations when used in a real environment, since the model could yield false positives because in this case, the input audio is not segmented into single duration chunks.

Acknowledgments. This study was funded by the Secretaria de Ciencia, Humanidades, Tecnología e Innovación (SECIHTI) through a master's degree scholarship.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Ahsan, M. M., Siddique, Z.: Machine learning-based heart disease diagnosis: A systematic literature review (2021), https://arxiv.org/abs/2112.06459
2. Brady, W. J., Lipinski, M. J., Darby, A. E., Bond, M. C., Charlton, N. P., Hudson, K. B., Williamson, K.: Electrocardiogram in Clinical Medicine. Wiley (2020), https://books.google.es/books?id=MCP5DwAAQBAJ
3. Chowdhury, T. H., Poudel, K. N., Hu, Y.: Time-frequency analysis, denoising, compression, segmentation, and classification of pcg signals. IEEE Access, vol. 8, pp. 160882–160890 (2020) doi: 10.1109/ACCESS.2020.3020806
4. INEGI: Estadísticas de defunciones registradas (edr) (2023), https://www.inegi.org.mx/programas/edr/, Último acceso: 30 de Junio del 2024
5. Khan, F. A., Abid, A., Khan, M. S.: Automatic heart sound classification from segmented/unsegmented phonocardiogram signals using time and frequency features. Physiological Measurement, vol. 41, pp. 055006 (2020) doi: 10.1088/1361-6579/ab8770
6. Krishnan, P. T., Balasubramanian, P., Umapathy, S.: Automated heart sound classification system from unsegmented phonocardiogram (pcg) using deep neural

network. Physical and Engineering Sciences in Medicine, vol. 43, pp. 505–515 (2020) doi: 10.1007/s13246-020-00851-w
7. Li, F., Liu, M., Zhao, Y., Kong, L., Dong, L., Liu, X., Hui, M.: Feature extraction and classification of heart sound using 1d convolutional neural networks. EURASIP Journal on Advances in Signal Processing, vol. 2019, pp. 59 (2019) doi: 10.1186/s13634-019-0651-3
8. Liu, C., Springer, D., Li, Q., Moody, B., Juan, R. A., Chorro, F. J., Castells, F., Roig, J. M., Silva, I., Johnson, A. E. W., Syed, Z., Schmidt, S. E., Papadaniil, C. D., Hadjileontiadis, L., Naseri, H., Moukadem, A., Dieterlen, A., Brandt, C., Tang, H., Samieinasab, M., Samieinasab, M. R., Sameni, R., Mark, R. G., Clifford, G. D.: An open access database for the evaluation of heart sound algorithms. Physiological Measurement, vol. 37, pp. 2181 (2016) doi: 10.1088/0967-3334/37/12/2181
9. Mushtaq, Z., Su, S.-F., Tran, Q.-V.: Spectral images based environmental sound classification using cnn with meaningful data augmentation. Applied Acoustics, vol. 172, pp. 107581 (2021) doi: https://doi.org/10.1016/j.apacoust.2020.107581
10. de México, G.: 20 principales causas de enfermedad nacional (2023), https://epidemiologia.salud.gob.mx/anuario/html/principales_nacional.html, Último acceso: 1 de Agosto de 2024
11. Noman, F., Ting, C. M., Salleh, S. H., Ombao, H.: Short-segment heart sound classification using an ensemble of deep convolutional neural networks. In: ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1318–1322 (2019) doi: 10.1109/ICASSP.2019.8682668
12. Organization, W. H.: Hipertensión (2023), https://www.who.int/es/news-room/fact-sheets/detail/hypertension, accedido: 2 de Agosto de 2024
13. Picton, P.: What is a Neural Network?, pp. 1–12. Macmillan Education UK (1994), https://doi.org/10.1007/978-1-349-13530-1_1
14. Reyna, M. A., Kiarashi, Y., Elola, A., Oliveira, J., Renna, F., Gu, A., Alday, E. A. P., Sadr, N., Sharma, A., Kpodonu, J., Mattos, S., Coimbra, M. T., Sameni, R., Rad, A. B., Clifford, G. D.: Heart murmur detection from phonocardiogram recordings: The george b. moody physionet challenge 2022. PLOS Digital Health, vol. 2, pp. e0000324– (9 2023)
15. Sandhya, P., Spoorthy, V., Koolagudi, S. G., Sobhana, N. V.: Spectral features for emotional speaker recognition. In: 2020 third international conference on advances in electronics, computers and communications (ICAECC). pp. 1–6. IEEE (2020)
16. Sharkawy, A.-N.: Principle of neural network and its main types. Journal of Advances in Applied Computational Mathematics, vol. 7, pp. 8–19 (2020)
17. Sharma, G., Umapathy, K., Krishnan, S.: Trends in audio signal feature extraction methods. Applied Acoustics, vol. 158, pp. 107020 (2020)

ISSN 1870-4069

Automatic Detection of Diabetic Retinopathy Using Classification Techniques and Computer Vision

Jorge Antonio Hernández Magallanes

Universidad Autónoma de Aguascalientes,
Centro de Ciencias Básicas,
Departamento de Ciencias de la Computación,
Mexico

jhernandez.dev00@gmail.com

Abstract. One of the most important challenges in modern medicine is the timely diagnosis of chronic diseases, as early detection can make a significant difference in the patient's quality of life and better management of the health system. A clear example of this problem is diabetic retinopathy (DR), an ocular complication of diabetes that constitutes one of the main causes of blindness worldwide. In this sense, early detection is key to avoid irreversible vision damage. This paper explores the use of imaging processing techniques and artificial intelligence as tools to address this challenge, focusing specifically on the automatic analysis of retinal scans for the early detection of diabetic retinopathy. Among the techniques used to preprocess the retinographies, resampling to standardize the number of images, noise elimination, cropping of the area of interest, as well as brightness and contrast adjustment stand out. In addition, contrast Limited Adaptative Histogram Equalization (CLAHE) and gramma correction were applied to improve image quality. For classification tasks, a convolution neuronal network (CNN) based on the DenseNet121 model is employed, which has shown promising results in initial tests. Although it is still under development, the aim is to improve its accuracy and efficiency to make it a practical and reliable tool in the early diagnosis of diabetic retinopathy.

Keywords: Diabetic retinopathy, retinography, automatic detection, computer vision, convolutional neural networks.

1 Introduction

Medicine is one of mankind’s oldest branches of science, which over the centuries has been transformed for the purpose of improving people's quality of life. Thanks to advances in technology, especially in the computer science field, medicine has undergone a revolution in the way diseases are diagnosed and treated. Currently, one of the biggest challenges facing medicine is the timely diagnosis of chronic diseases, as the timing of such diagnoses is crucial to prevent a person’s health from being severely impacted.

A good example of this is diabetic retinopathy (DR), a common complication of diabetes that, if not detected early, can lead to irreversible vision loss, severely affecting

Fig. 1. Examples of types of retinopathy a) NO DR, b) MILD, c) MODERATE, d) SEVERE, d) DR PROLIFERATIVE.

the patient's quality of life. This disease, associated with diabetes mellitus [1], does not present significant symptoms in the early stages, making it difficult to detect in a timely manner.

According to the study carried out in 2021, entitled "Global Prevalence of Diabetic Retinopathy and Projection of Burden through 2045" estimates that currently, the worldwide prevalence of patients with this medical complication is situated between 22.27% among people with diabetes, and it is estimated that year 2045 this figure will increase exponentially [2]. These data are more alarming when we consider that according to another study published in the journal springer entitled "Diabetic retinopathy as the leading cause of blindness and early predictor of cascading complications-risks and mitigation" diabetic retinopathy (DR) is considered one of the leading causes of irreversible blindness in the world, especially in developing countries [3], As for Mexico, the prospect is not encouraging, since, according to a Mexican government fact sheet issued on July 22, 2018, the prevalence of diabetic retinopathy among people with diabetes is close to 31.5% [4].

Due to the nature of this disease, timely diagnosis is crucial to prevent serous complications and irreversible damage. However, traditional diagnosis based on manual inspection of retinal images is a time – consuming process and, above all, requires experience, which limits its accessibility in many regions. In this sense, image processing and the use of advanced artificial intelligence algorithms represent an opportunity to address and try to provide a solution to this problem.

By automating the retinal scan analysis process, early signs of diabetic retinopathy can be detected and help healthcare professionals make faster and more effective diagnostics, even in resource-limited settings, and enable more continuous and accessible monitoring for a wider range of patients.

With this scenario, the objective of this research is to develop an automated system that using image processing techniques and artificial intelligence models can detect early any of the stages of diabetic retinopathy, which can be seen in the image (See Fig. 1), thanks to the combination of image processing and the use of convolutional neural network can significantly improve the accuracy and efficiency of automated diagnosis compared to traditional approaches, this contributing to a useful tool especially in context with limited resources. However, the study continues to evaluate the performance of other neural networks models to identify the most suitable architecture of this task.

2 Related Work

During the last few years, many number of studies have tried to explore and propose various approaches and techniques to try to solve this problem, such as the study entitled "Automatic Detection of Diabetic Retinopathy Applying Computer Vision and Convolution Neural Network" [5], in which, by means of transformation of retinography to grayscale, edge detection and clipping, attempts to perform early identification of signs of retinopathy, with a degree of certainty of 91%, the model is proposed as a stable model for the detection of retinopathy. Another relevant study is "Deep learning based binary classification of diabetic retinopathy images using transfer learning approach" [6], in this work, pretrained network are used with the use of three different datasets, which are: DRD-EyePACS, IDRiD and APTOS-2019, divided into training, testing and validation sets, and with the use of preprocessing and data augmentation techniques, the most outstanding model reaches 97.33%.

In addition, a study entitled "Binary Classification of Diabetic Retinopathy Using CNN Architecture" [7] using several pretrained networks such as EfficientNet, VGG16 and MobileNet, among others, applies processing techniques such as the application of Gaussian filters, image resizing and data augmentation, to achieve an accuracy of 94.55%. However, all of them focus on binary disease classification, i.e., they determine whether an image shows signs of retinopathy.

This binary approach, although useful, limits the ability of the models to capture the complexity of diabetic retinopathy, as it does not consider the different stages of the disease or the severity of damage, which could be crucial for early diagnosis and appropriate intervention.

Other studies address this problem from a multiclass classification approach, which seeks to identify the degree of severity of diabetic retinopathy according to the different clinical stages of the disease (mild, moderate, severe and proliferative). This type of approach provides a more detailed analysis, such is the case of the study entitled "Improved Automatic Diabetic Retinopathy Severity Classification Using Deep Multimodal Fusion of UWF-CFP and OCTA Images" [8] which uses ultra-widefield fundus images (UWF-CFP) and optical coherence tomography angiography (OCTA) using ResNet50 and 3D-ResNet50 models with attention blocks, Finally, the study entitled "Transfer-Ensemble Learning based Deep Convolutional Neural Networks for Diabetic Retinopathy Classification" [9] proposes an ensemble model that combines VGG16 and Inception V3 pre-trained networks to classify DR images into five classes. Using the APTOS dataset, the model achieved an accuracy of 96.4%. However, for further studies, some papers propose the idea of using different deep learning techniques with the aim of improving the accuracy in multiclass classification of diabetic retinopathy and addressing challenges such as class imbalance and variability in image quality.

The main contribution of this article is the use of different techniques to improve the quality of images that may present unwanted variations, trying with image processing techniques that have not been used in the literature of the problem.

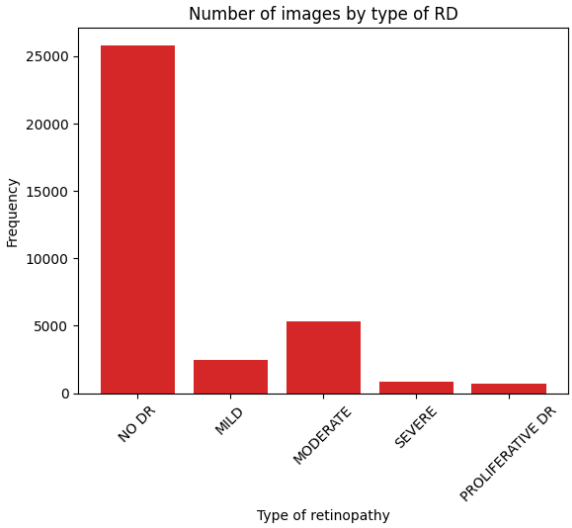


Fig. 2. Number of images per type of retinopathy are present in this dataset.

3 Methodology

The methodology of this study consists of several phases, from the acquisition of the database and the preparation of the data to the training and evaluation of the preliminary results, the process is as follows:

3.1 Image Acquisition

Preprocessing is the stage in which it is intended to repair the images obtained from any defects that may be acquired during the capture of the image, solving flaws produced or generated by the capture hardware.

To achieve the objective of building a computer vision system, it is necessary to carry out the first stage of the process, which is to obtain the dataset with which to work. In this case, the dataset used will be the "Diabetic Retinopathy Detection" [10], which was extracted from Kaggle. All the images contained in this dataset are in JPEG format with dimensions of 4752 x 3168 px. The dataset includes images covering the five stages of diabetic retinopathy, which can be seen in the image above (See Fig. 1), allowing a classification according to their severity.

Before starting to apply the technique to improve the quality of the image, it is necessary to verify that the dataset is balanced.

If there are many more images of one class than another, the model may become biased and fail to correctly learn the necessary patterns, through an exploration data analysis (EDA), to understand the nature and state of the dataset, the results derived from this process can be seen in the graph above (See Fig. 2), where the distribution in the number of images between classes is evident, so a resampling technique should be applied.

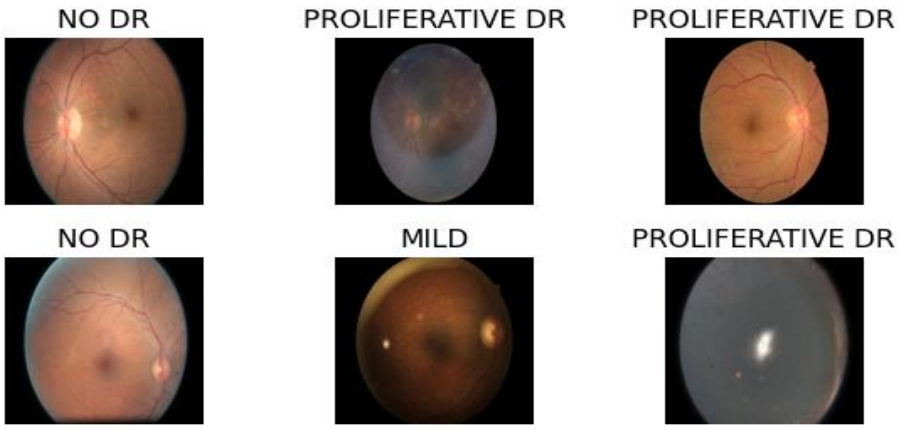


Fig. 2. Examples of retinography without preprocessing techniques

3.2 Resampling

Resampling is a technique that consists of creating a new data sample from a set of original data with some kind of bias [11], To address this problem, there are several resampling techniques, the most common being oversampling (which consists of increasing the number of samples of minority classes), under sampling (which reduces the number of samples of majority classes) and the use of class weights (which assigns a higher weight to the less represented classes during the training of the model, without the need to modify the size of the data set). The use of class weights could have led to a longer convergence time, precisely because of the large disparity in the number of images between classes.

We will then resample 700 images per class to avoid duplication of data and randomize the order of the samples, resulting in a balanced dataset of 3,500 retinography. Furthermore, recent studies in automatic medical image diagnosis, such as "Self-Supervision for Medical Image Classification: State-of-the-Art Performance with ~100 Labeled Training Samples per Class [12]", have shown that datasets with between 500 and 1000 images per class, combined with data augmentation techniques, are sufficient to achieve robust performance metrics.

3.3 Image Preprocessing Techniques

The quality and consistency of the input images for a classification model play a crucial role in the performance of machine learning models for diabetic retinopathy detection.

In the image above (See Fig. 3), we can see the retinography of the dataset without any modifications, where we can see problems in visualizing the key factors that could help to predict the type of retinopathy in which the patient is. To address this problem, we will apply an enchantment process to optimize the quality of the images and highlight important details, allowing for a more accurate analysis, which is described below.

Denoising filter. When capturing an image, depending on multiple factors, unwanted variations may be introduced in the pixels. This is known as noise and affects the image, since it may acquire different brightness or color characteristics than the original ones [13]. They're capturing an image, depending on multiple factors, unwanted variations may be introduced in the pixels. This is known as noise and affects the image, since it may acquire different brightness or color characteristics from the original ones.

There are several types of digital noise, each with specific characteristics, which necessitates the use of different filters designed to improve image quality. One of these filters is bilateral filtering, which, unlike other types of filters, is particularly effective in reducing noise in images while keeping the edges sharp. Because of this capability, bilateral filtering is especially useful in applications where preserving fine image detail is critical, such as in medical image analysis.

Cropping. The process of cropping involves trimming the outer edges of an image to remove unnecessary or irrelevant areas, focusing the analysis on the most important central features. This technique can help eliminate noise or artifacts that may be present in the peripheral parts of the image, improving the overall quality and accuracy of the analysis. To achieve edge clipping, the following process is carried out: first, the thresholding technique is applied, which is used in image processing to generate binary images from a grayscale image. This is achieved by setting a threshold value: all pixels that exceed this value become white, while those that do not are transformed into black. Subsequently, the function find Contours of OpenCV is used, which allows to detect the internal and external contours of a binary image, this function will be used to detect the main contour, to achieve this, the results of this function will be taken, the largest contour is selected, which will serve to obtain the delimiting rectangle of the contour and finally the image is cropped.

Brightness and contrast adjustment. In image processing, brightness and contrast are two fundamental characteristics that affect the perception and quality of images. Brightness is related to the overall light intensity, while contrast measures the difference between light and dark areas, allowing them to highlight important details. To improve these parameters, a technique based on the use of the mean and the standard deviation present in the image is applied.

Contrast adjustment was performed using the inverse of the standard deviation, while brightness adjustment was centered using the median of the pixel values (50th percentile), allowing for more stable and robust corrections. Both factors were constrained with limit values to avoid overexposure or underexposure of the images.

CLAHE adjustment. CLAHE (Contrast - Limited Adaptive Histogram Equalization) is an advanced method used to enhance the contrast of images, especially those with non-uniform illumination variation.

Unlike traditional histogram equalization, which uses the range of image intensities globally, CLAHE employs an adaptive strategy, which consists of dividing the image into smaller blocks or regions, over which a local histogram is calculated for each section. The local histograms are then used to adjust the brightness and contrast of each of these regions individually. [14]

One of the main benefits of CLAHE is the ability to avoid overexposure in areas of highlight intensity, a common problem in global histogram equalization. By limiting

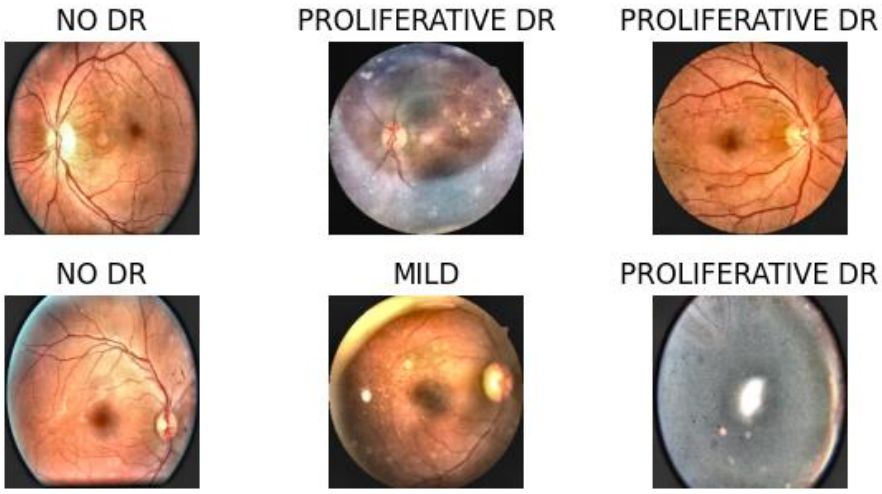


Fig. 4. Examples of retinography with preprocessing techniques applied.

the contrast of each local region by means of a threshold (which is adjusted for each image), the risk of generating unwanted changes in the image is reduced.

With these changes, image detail is more easily perceived, while preserving more subtle information in areas of low contrast.

GAMMA correction. Gamma is the relationship between the numerical value of a pixel and its actual luminance. Without this value, the tones captured by any device could not be represented in the way a person can visualize it, this correction allows them to compensate for these differences.

As a result of this process, the following image (See Fig. 4) shows a visible improvement of the details present in the image, with more balanced tones and highlighting of anatomical features of the eye necessary to perform the classification process.

To evaluate the improvement in the images two parameters will be used to evaluate them the Signal-to-Noise Ratio (SNR) and Peak Signal-to-Noise Ratio (PSNR), the SNR is a measure that compares the power of the useful signal (relevant information) with the power of the noise (unwanted interference), a high value indicates that the image is not affected by noise, while a low value indicates the presence of significant noise in the image, while the PSNR is a metric used to evaluate the quality of the processed images compared to the original ones. It quantifies the ratio between the maximum possible value of a signal (the highest pixel value) and the noise present in the image. A higher PSNR indicates that the processed image is of high quality, with little distortion compared to the original image. PSNR values above 40 dB are considered to be of excellent quality. [15]

The quantitative results suggest that the applied preprocessing has significantly improved the quality of the retinal images.

```
Average Original SNR: 0.284702357254216O5
Average Processed SNR: 4.766901181346978
Average Contraste: 0.9885407340956234
Average PSNR: 58.73854062813858
```

Fig. 5. Results of the evaluation of changes in image quality.

The values for this can be seen in the image above (See Fig. 5). Overall, it can be concluded that the preprocessing has significantly reduced the noise present in the images, as evidenced by the increase in the signal-to-noise ratio (SNR) value, which indicates an improvement in the clarity of the processed images and the elevated PSNR reveals that the processed images have maintained a quality very close to the originals.

For more details on the image enhancement process, as well as access to the code and examples used, please visit the repository: https://github.com/ jhernandezdev00/RDVis_DX

4 Construction of the Classifier

Classification is defined as the process by which a set of objects or elements can be grouped into different categories or classes, depending on some characteristic or factor that they share [16].

In machine learning, the choice of classification method depends largely on the type of learning used. There are three main approaches: supervised, unsupervised and semi-supervised learning, differing in the way the model works with the data.

In supervised learning, the data are labeled, that is, they have a predefined class, while in unsupervised learning the data do not have a label, so the model must group them according to patterns and similarities detected in the dataset, finally, semi-supervised learning combines the two previous ones, thus avoiding the need for exhaustive labeling [17].

4.1 Construction of the Convolutional Neural Network

A convolutional neural network (CNN) is a type of network specialized in deep pattern learning by using filters to extract relevant features from data. These networks are used to perform tasks that require object recognition or pattern identification [18].

For the structure of the model, we have chosen to use a pre-trained model as the basis for the classification model, although initially we had chosen to use EfficientNetB0 due to its optimal performance and not being a network with too much computational load, we have decided to switch to DenseNet121. This decision was based on performance comparisons reported in the literature, where DenseNet121 shows better feature extraction and recognition capability in classification tasks.

In addition, in preliminary tests it shows a better accuracy rate on the dataset that has been used, although it is important to note that, at the time of writing this paper, other techniques are still being investigated and tested to find the one that gives the best performance and results.

4.2 Network Architecture

To design the network architecture for the classification task, the input images are first resized to 256 x 256 pixels. The model uses DenseNet121, pre-trained on ImageNet, as the base. This base model is configured without its top classification layer and with its weights initially frozen to retain previously learned features. On top of the base, a Global Average Pooling 2D layer is added, followed by a sequence of dense blocks to refine the learned features. The first block includes a dense layer with 512 units and ReLU activation, followed by Batch Normalization and a 30% Dropout layer. Next, a second block includes a dense layer with 256 units and Batch Normalization. This is followed by a third block with a 128-unit dense layer, again with ReLU activation, Batch Normalization, and a 30% Dropout layer. Finally, the output layer is a dense layer with 5 units and SoftMax activation, providing the final predictions for multiclass classification.

In addition, the fine-tuning technique was applied to optimize performance. Initially, a pre-trained model was used on ImageNet, so that the base layers are frozen and only the upper layers are trained. After some epochs, some of the deeper layers are unfrozen to adjust the weights gradually on our specific dataset. In future phases of the project, other types of more comprehensive methods are planned.

4.3 Model Training

The model uses the Adam optimizer with batch of 32 images, during the first training stage, corresponding to the fine-tuning process. To dynamically adjust the learning rate, ReduceLROnPlateau is used, which reduces the value of the learning rate that allows using the validation metrics to dynamically adjust the learning value, together with the use of Early Stopping that stops the training if it does not improve over a certain number of epochs to avoid over-fitting, this first stage is executed for about 15 epochs.

For the second stage that will perform the final model adjustment, where the last 15 layers of the base model are unfrozen, the learning rate is adjusted to 0.00001, this to prevent the model from forgetting everything it learned in the first stage and Early Stopping is used again, for the final stage it is run for 30 epochs.

4.4 Model Evaluation

To test the accuracy of the training, tests were performed with random images that were not part of the training or validation, and to record the success cases of the prediction, as well as the degree of accuracy of each of the tests, and the accuracy and loss generated from the set of tests were considered as main factors.

5 Results

Before going into detail on the results, it is important to note again that at the time of writing this article (April, 2024), the model building and testing phase is still in an early stage of development.

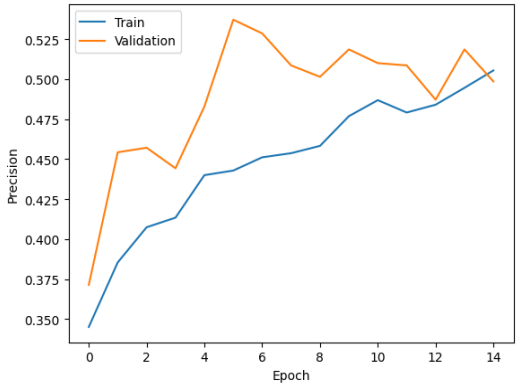


Fig. 6. Results of the first stage of model training.

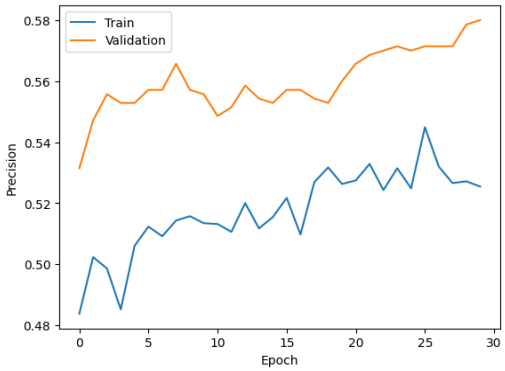


Fig. 7. Results of the second stage of model training.

```
22/22 ━━━━━━━━━━━━━━━━━━━━ 4s 173ms/step - accuracy: 0.5770 - loss: 1.0290
Loss: 1.0238395929336548
Accuracy: 0.5785714387893677
```

Fig. 8. Accuracy values obtained in tests.

To check the accuracy of the training, tests were performed with random images that were not part of the training and validation, and to record the success cases of the prediction, as well as the degree of accuracy of each of the tests. During the training it is observed that in the first phase of the fine-tuning, the training starts between 20% ~ 25%, with a loss value close to 2.15, during the 15 epochs to which the model was subjected, the value of the accuracy ended with a value close to 50% and reducing the loss value to 1.12 in the tests.

Passing these epochs in the second and last stage the model will be trained for 30 epochs or until the loss value does not improve, in order to avoid overfitting, the model starts with an accuracy value of 53% with a loss of 1.08, when completing its training the accuracy value improves to 57% and ends up reducing the loss to a value close to 0.91, these results can be seen in the graph below (See Fig 6 and 7).

As can be seen, fine tuning was a good strategy to finish adjusting and completing the model, even so, the values obtained do not reach the expected accuracy value, as can be seen in the image above, where the tests resulted in an accuracy of 57% (See Fig. 8).

Even so, these results are encouraging, since the detection of each type of retinopathy represents a very complex process, even for experienced specialists. The ability of model to differentiate between the different categories, although still at an early stage, suggests that convolutional neural networks may become a valuable diagnostic support tool in clinical settings.

6 Conclusions

Preliminary results suggest that the use of convolutional neural networks as a tool for the diagnosis of diabetic retinopathy represents a significant breakthrough in the growing integration and use of artificial intelligence in the field of medicine. Although the results do not yet reach the expected levels of accuracy, they show favorable progress in the ability of a single model to extract and recognize the most relevant patterns present in a retinography, which is essential for accurate and automated detection of diabetic retinopathy.

Furthermore, this initial breakthrough demonstrates the critical role that preprocessing plays in image standardization and enhancement, as it allows the network to learn from a more uniform and optimized dataset, which helps to improve its interpretation of the anatomical features of the eye. This stage is even more relevant considering that some studies tend to work with the dataset without applying any enhancement or modification, which can limit the performance of the model and its generalization capability.

In addition, this first phase of research has identified limitations inherent to shallow models, which usually present greater complications when generalizing the knowledge extracted from the features or biased to some of them in specific to give their prediction.

Therefore, it is recommended for these cases the use of models robust enough for this task or that integrate advanced feature extraction mechanisms, such as deep convolutional layers or attention techniques, and in cases where the number of images is reduced or the hardware capabilities are limited, the use of state-of-the-art pre-trained models, in order to significantly improve the accuracy in the detection of retinopathy, especially in its early stages where identification is more complex.

As a next step for the research, the use of regularization techniques and strategies to increase the robustness of the model is proposed, as well as the expansion of the data set to expose the algorithm to a greater diversity of clinical cases. These actions aim to improve the generalization capacity of the system and help mitigate possible biases that the model may generate during its learning process, with the ultimate goal of moving towards a more accurate and reliable model that can be used as an effective support tool for health professionals in the early detection and timely treatment of diabetic retinopathy.

References

1. Centro de Oftalmología Barraquer 2024, Retinopatía diabética. https://www.barraquer.com/patologia/retinopatia-diabetica
2. Teo, Z.L. *et al.*: Global Prevalence of Diabetic Retinopathy and Projection of Burden Through 2045: Systematic Review and Meta-analysis, Ophthalmology, 128(11), pp. 1580–1591 (2021) doi: 10.1016/j.ophtha.2021.04.027.
3. Kropp, M. *et al.*: Diabetic Retinopathy as the Leading Cause of Blindness and Early Predictor of Cascading Complications—Risks and Mitigation. Springer Science and Business Media Deutschland GmbH, (2023) doi: 10.1007/s13167-023-00314-8.
4. Gobierno de México, Retinopatía diabética o ceguera irreversible por inadecuado control de la diabetes. https://www.gob.mx/salud/documentos/retinopatia-diabetica-o-ceguera-irreversible-por-inadecuado-control-de-la-diabetes
5. Bernal Catalán, E., De La Cruz Gámez, E., Montero V., J.A., H. Reyna, M.T.I.: Detección automática de retinopatía diabética aplicando visión artificial y redes neuronales convolucionales.
6. Saproo, D., Mahajan, A.N., Narwal, S.: Deep Learning Based Binary Classification of Diabetic Retinopathy Images Using Transfer Learning Approach. Journal of Diabetes & Metabolic Disorders, 23(2), pp. 2289–2314 (2024) doi: 10.1007/s40200-024-01497-1.
7. Khudaier, A.H., Radhi, A.M.: Binary Classification of Diabetic Retinopathy Using CNN Architecture. Iraqi Journal of Science, 65(2), pp. 963–978 (2024) doi: 10.24996/ijs.2024.65.2.31.
8. Daho, M.E.H. *et al.*: Improved Automatic Diabetic Retinopathy Severity Classification Using Deep Multimodal Fusion of UWF-CFP and OCTA Images, (2023) doi: 10.1007/978-3-031-44013-7_2.
9. Ghosh, S., Chatterjee, A.: Transfer-Ensemble Learning Based Deep Convolutional Neural Networks for Diabetic Retinopathy Classification, doi: 10.48550/arXiv.2308.00525.
10. Dugas, E., Jared, J., Cukierski, W.: Diabetic Retinopathy Detection. https://kaggle.com/competitions/diabetic-retinopathy-detection
11. Data Science. Resampling: A Method for Balancing Data. https://datascientest.com/en/resampling-a-method-for-balancing-data
12. Nielsen, M., Wenderoth, L., Sentker, T.: Self-Supervision for Medical Image Classification: State-of-the-Art Performance with ~100 Labeled Training Samples per Class, 10(8), pp. 895 (2023) doi: 10.3390/bioengineering10080895.
13. Fotolarios: El ruido digital - Qué es, por qué aparece, cómo evitarlo. https://www.fotolarios.es/2018/11/el-ruido-digital.html
14. Free-Astro Team: Contrast-Limited Adaptive Histogram Equalization (CLAHE). https://siril.readthedocs.io/es/stable/processing/clahe.html
15. Gonzalez, R.C., Woods, R.E.: Digital Image Processing, Global Edition. 2017 https://dl.icdst.org/pdfs/files4/01c56e081202b62bd7d3b4f8545775fb.pdf
16. IBM, Tipos de clasificación. https://www.ibm.com/docs/es/msp/7.6.3?topic=classifications-classification-types
17. Zoumana, K.: Classification in Machine Learning: An Introduction. https://www.datacamp.com/blog/classification-machine-learning
18. Zoumana Keita: ¿Qué es una red neuronal convolucional (CNN)?. https://www.datacamp.com/es/tutorial/introduction-to-convolutional-neural-networks-cnns

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

http://rcs.cic.ipn.mx

Centro de Investigación
en Computación