EDUCACIÓN
SECRETARÍA DE EDUCACIÓN PÚBLICA

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

Research in Computing Science

Vol. 154 No. 7
June 2025

Research in Computing Science

***RESEARCH IN COMPUTING SCIENCE**, Año 25, Volumen 154, No. 7, Julio 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 04 de Julio de 2025.*

***RESEARCH IN COMPUTING SCIENCE**, Year 25, Volume 154, No. 7, July, 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 July 4, 2025.*

Advances in Artificial Intelligence

Lourdes Martínez-Villaseñor (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

Sistema de reconocimiento de voz humana y sintética 5
Jibran Zaedt Rodriguez Garcia, Andrea Magadán Salazar

Gas Metal Arc Welding Dataset for Computer Vision Quality Assessment 19
José A. López-Islas, Oscar Camacho-Nieto, Yenny Villuendas-Rey

Segmentación automática de tumores cerebrales usando K-means 31
Kay García-Sánchez, Daniel Cantón-Enríquez, Hugo Jiménez-Hernández, Luis Antonio Díaz-Jiménez, Ana Marcela Herrera-Navarro, Jorge Luis Pérez-Ramos, Selene Ramírez-Rosales, Carlo Giovanni Cetina-Camacho

Algorithms and Approaches Used in Medical Image Segmentation for Cell Migration Tracking: A Systematic Literature Review 43
Mariela Judith Domínguez-Domínguez, Ángel J. Sánchez-García, María Yesenia Zavaleta-Sánchez, Carlos Adrián Alarcón Rojas

Enhancing Embryo Image Interpretability through Language Models 57
Alberto León, Isaac Aguilar, Omar Paredes

Revolutionizing the Fight against Child Cyberbullying: Using Holograms and Voice Recognition as Allies 69
María del Carmen Hidalgo Baeza, Alberto Ochoa Zezzatti, Víctor Manuel Casas Gómez, Irma Yazmín Hernández Baez

Clasificación de datos de cotización de la bolsa mexicana de valores usando aprendizaje automático 85
José Gonzalo Ramírez Rosas, Jorge de la Calleja, Araceli Ortiz Carranco, Martin Neri Suárez, Salvador Antonio Arroyo Diaz

Responsible Use of AI as a Transversal Theme in an Interaction Design Course: A Report on Participation in the Instructional Co-Design Process 97
Scarlett Itzel Xochicale Flores, Angelica Rodríguez Vallejo, Soraia Silva Prietch, Josefina Guerrero García, Juan Manuel González Calleros

Optimización de la arquitectura Pix2Pix: Un estudio de reducción de capas y calidad de imagen 113
Alexander Tapia Cortez, José Alejandro Tejeda Sánchez, Yolanda Moyao Martínez, David Eduardo Pinto Avendaño, José de Jesús Lavalle Martínez

Optimizing Best Response Dynamics-based Facility Location Games Using Reinforcement Learning 125
Andrés Burjand Torres Reyes, Rolando Menchaca-Méndez, Francisco Hiram Calvo-Castro

ISSN 1870-4069

Sistema de reconocimiento de voz humana y sintética

Jibran Zaedt Rodriguez Garcia, Andrea Magadán Salazar

TecNM/CENIDET
Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET),
México

{m23ce077, andrea.ms}@cenidet.tecnm.mx

Resumen. El advenimiento de la clonación de voz por inteligencia artificial ha revolucionado el campo de la síntesis de voz, ofreciendo una autenticidad y personalización sin precedentes. Las aplicaciones de esta tecnología son numerosas y diversas, abarcando el sector del entretenimiento, la accesibilidad, el marketing digital y enfoques pioneros para la creación de contenido y la comunicación digital. Este artículo analiza algunas de las principales plataformas de clonación de voz y sus aplicaciones prácticas. Examina la funcionalidad, características e impacto de cada plataforma en la comunicación. Luego, la discusión aborda las consideraciones éticas clave que rodean su uso, incluyendo el fraude, la privacidad, la desinformación y el impacto potencial en el empleo de los artistas de voz. El potencial de la clonación de voz para ofrecer experiencias completamente nuevas en muchas áreas es significativo, incluido el uso de voces de personajes en películas y la creación de audiolibros con voces históricas. Además, también proporciona información sobre los motores y algoritmos subyacentes que impulsan estas aplicaciones. También explica cómo la integración de redes neuronales con modelos de alto nivel permite la personalización de voces digitales. Finalmente, el artículo discute la necesidad de un marco ético y regulatorio que garantice prácticas responsables de clonación de voz que protejan los derechos individuales y el valor del trabajo humano en el entorno tecnológico en evolución.

Palabras clave: Herramientas de clonación de voz, algoritmos de clonación, aplicaciones de clonación, inteligencia artificial.

Human and Synthetic Speech Recognition System

Abstract. The advent of artificial intelligence voice cloning has revolutionized the field of speech synthesis, offering unparalleled authenticity and personalization. The applications of this technology are numerous and diverse applications, spanning the entertainment sector, accessibility, digital marketing, and pioneering approaches to content creation and digital communication. This paper analyses some of the main voice cloning platforms and their practical applications. It examines the functionality, features, and impact of

each platform on communication. The discussion then moves on to address the key ethical considerations surrounding their use, including fraud, privacy, misinformation, and the potential impact on the employment of voice artists. The potential for voice cloning to offer entirely new experiences in many areas is significant, including the use of character voices in films and the creation of audio books with historical voices. Furthermore, it also provides insights into the underlying engines and algorithms that power these applications. It also explains how the integration of neural networks with high-level models enables the customization of digital voices. Finally, the paper discusses the need for an ethical and regulatory framework to ensures responsible voice cloning practices that protect individual rights and the value of human labour in the evolving technological environment.

Keywords: Voice cloning tools, cloning algorithms, cloning applications, artificial intelligence.

1. Introducción

En la actualidad, los asistentes de voz o altavoces inteligentes están ganando protagonismo y ya forman parte de nuestra vida. Por ejemplo, se pueden controlar otros dispositivos como termostatos, aires acondicionados, luces, refrigeradores, televisores, entre otros, mediante la voz. Los teléfonos inteligentes han sido los precursores en estos desarrollos de la voz como elemento primordial. Las aplicaciones son variadas y van desde los asistentes personales para ejecutar órdenes, para reconocer dictados, leer textos, para añadir realismo en los videojuegos, etc.

Los sistemas de reconocimiento de voz y generación de voz que pueden procesar de forma natural un diálogo entre humano y máquina se encuentran bajo el término de sistemas o interfaces de voz natural que se destinan a aplicaciones de cliente-servidor en entornos conversacionales. La generación de voz es el proceso mediante el cual un dispositivo inteligente produce secuencias de habla artificial. El reconocimiento de voz es el proceso mediante el cual las computadoras interpretan y digitalizan las señales de voz, tanto para su análisis de contenido como para la interpretación de órdenes.

El proceso de conversión de texto a voz en los asistentes virtuales consta de tres etapas principales [1]:

1. **Entrada de texto y conversión fonémica:** El texto se transforma en una cadena de fonemas, incluyendo puntuación y límites de palabras. Esto permite al modelo capturar mejor la prosodia y los ritmos del habla.
2. **Creación de espectrogramas Mel con Tacotron:** Los fonemas se convierten en espectrogramas Mel utilizando una red basada en la atención secuencia a secuencia, Tacotron. Esta red emplea el enfoque secuencia a secuencia con capas semejantes a la red Long Short-Term Memory en la secuencia para procesar y generar muchos fotogramas del espectrograma al mismo tiempo, lo que lo hace más eficiente y de alta calidad.

3. **Conversión a audios con WaveRNN:** El espectrograma de Mel alimenta a una red neuronal autorregresiva llamada WaveRNN para generar audio, muestra por muestra. En ese nivel, la señal de audio se genera a partir del espectrograma mismo por la red neuronal, y el control de la velocidad y la calidad se realiza mediante una optimización adicional. Este proceso no utiliza los datos de voz exactos que uno podría desear simular. En cambio, utiliza una voz de archivo en el proceso de creación de los fonemas asociados para el texto dado y, por lo tanto, realmente pierde mucho en términos de fidelidad para simular la voz de una persona real.

En los últimos años ha surgido otra área de desarrollo conocida como clonación de voz. En términos sencillos, la clonación de voz es el proceso de copiar la voz de una persona para reproducirla o generarla en un contexto diferente al original. Es producir una voz artificial que tenga las mismas características (suene igual) como si la hubiera pronunciado una persona objetivo [1].

La clonación de voz no es nueva; sin embargo, las nuevas herramientas de inteligencia artificial logran mayores niveles de autenticidad y personalización. Este progreso se ha utilizado para replicar la voz humana de formas que ensalzan en sectores como el entretenimiento, la accesibilidad y el marketing digital. La clonación de voz está liderando actualmente la creación de contenido y las experiencias de comunicación digital [2]. Si bien algunas aplicaciones consideran texto para la clonación de voz, no se recomienda porque suele provocar pérdida de información en el proceso de transmisión oral de un mensaje.

El objetivo de este artículo es presentar las principales plataformas para la clonación de voz y sus aplicaciones en la vida real, así como analizar el uso de la frecuencia fundamental (f_0) como característica principal en el entrenamiento de un clasificador basado en Máquinas de Vectores de Soporte (SVM, por sus siglas en inglés) para la identificación de voces clonadas y naturales. Se exploran las capacidades de esta metodología en la detección de patrones distintivos entre voces generadas artificialmente y voces humanas reales, con el fin de evaluar su eficacia en la autenticación y verificación de identidad en entornos digitales.

Los sistemas revisados se basan en las tendencias de las comunidades activas como: Discord [3], GitHub [4] y similares, donde los desarrolladores y entusiastas comparten su experiencia de primera mano y sus preferencias por dichas técnicas. Se pretende examinar cómo funcionan las herramientas, en qué entornos y qué cambian en la comunicación digital. Esto hace que la revisión de las herramientas de clonación de voz sea relevante no solo para investigadores y desarrolladores, sino también para cualquier persona que esté interesada en los avances en la Inteligencia Artificial y sus consecuencias para las sociedades.

Para comprender mejor el potencial de la clonación de voz, a continuación, se presenta una lista de aplicaciones útiles de esta tecnología en diversos sectores.

2. Herramientas de clonación de voz

A diferencia de los avances en la detección de deepfakes en imágenes, se observa una escasez de trabajos dedicados a la detección de voces clonadas [5]. Esto subraya

la necesidad de investigaciones adicionales en este campo para abordar los desafíos específicos asociados con la manipulación de audio.

La importancia de este campo radica en la necesidad de proporcionar al usuario final una mayor confianza en los sistemas de comunicación y verificación de identidad. A medida que las técnicas de falsificación de voces se vuelven más sofisticadas, se incrementa también la urgencia de desarrollar métodos de detección igualmente avanzados. Esto no solo contribuye a la seguridad personal y empresarial, sino que también juega un papel crucial en la preservación de la integridad de la información y la prevención del fraude.

Se llevó a cabo una investigación sobre varias herramientas, de preferencia gratuitas, que pueden generar audios de voces clonadas de manera convincente, evitando sonidos ”antinaturales”. Las siguientes herramientas de clonación de voz tienen distintas capacidades que se utilizan según las necesidades del usuario y/o de las aplicaciones:

1. **ElevenLabs** [6]: Es una herramienta avanzada de Inteligencia Artificial (IA) que ofrece tecnologías de Texto a Voz, Voz a Voz y Clonación de voz. Con esta aplicación es posible generar audio hablado de alta calidad en una variedad de voces, estilos e idiomas (actualmente 32). También permite ajustar géneros, edades, tonos y acentos según las preferencias del usuario.

   Su modelo de IA captura de manera excepcional la entonación y las inflexiones humanas, ofreciendo una experiencia de voz sumamente realista. Para evitar el uso de su tecnología en la creación de deepfakes, ElevenLabs ha adoptado controles, permitiendo que este producto esté disponible solo para usuarios verificados mediante suscripción.

   Utiliza algoritmos que maximizan la estabilidad y similitud de las voces, ajustables a través de su API. ElevenLabs ha proporcionado en GitHub la documentación y los ejemplos de código necesarios para integrarlo con herramientas como Python y Java. Los modelos disponibles en esta plataforma incluyen Multilingual v2, English v1, Turbo v2 y Turbo v2.5.

2. **VocaliD** [7]: Crea voces personalizadas para personas con discapacidades del habla. Para la generación de voz combina las características vocales de los usuarios con voces pregrabadas para generar una voz única. Esto lo logra mediante la combinación de su base de datos ”Human Voicebank”, que incluye más de 14.000 donantes en 110 países.

   La integración de estas voces en dispositivos de asistencia como Tobii Dynavox dice mucho sobre la personalidad y las emociones del usuario.

3. **Applio** [8]: Es una aplicación de clonación de voz de uso gratuito sin límites para crear su modelo de clonación de voz. Permite la síntesis de voz a texto y de voz a voz. Realiza la transformación de audio utilizando diferentes algoritmos de extracción de tono como Pitch Marking, Harvest, DIO, Rmvpe y Rmvpe_gpu.

   Permite varias opciones para ajustar el procesamiento de audio, lo cual brinda variedad entre ser una herramienta experimental o una aplicación más profesional para fines específicos. Esta aplicación es de uso público y gratuito, mediante la cual los usuarios pueden crear modelos de clonación de voz sin límite alguno. Además,

ofrece otras herramientas como la descarga de modelos de voces ya entrenados y listos para ser utilizados.

4. **RVC** [9]: Esta aplicación cuenta con una versión web y está disponible para el público en general de forma gratuita. A través de ella, los usuarios pueden crear modelos de clonación de voz sin restricciones.

   Además de esta función principal, la aplicación ofrece diversas opciones adicionales como cambiar el audio de la grabación seleccionando la frecuencia de muestreo (40k o 48k). También permite seleccionar el algoritmo de extracción, que puede ser Pitch Marking, Harvest, DIO o Rmvpe, con opciones de 0 a 8 subprocesos de CPU. Es una alternativa flexible para quienes necesitan personalizar los modelos de voz.

5. **Voice.ia** [10]: Esta aplicación cuenta con una amplia variedad de voces desarrolladas por la comunidad. Puede proporcionar un cambio de voz en tiempo real y da acceso a una gran cantidad de voces creadas y almacenadas.

   Es una aplicación paga, aunque incluye la opción de recibir una paga mínima diaria por iniciar sesión, que se puede usar para comprar más voces. Las voces se pueden usar sin restricciones una vez compradas. No se menciona ningún algoritmo de extracción, lo que puede afectar a los usuarios que buscan detalles técnicos específicos.

3. Algoritmos utilizados por las herramientas

Los algoritmos forman una parte esencial en la clonación de voces porque son la base de la construcción de las herramientas en el análisis, ajuste y reproducción de las voces con precisión. Las técnicas aplicadas incluyen:

– **Redes neuronales profundas**: Se utilizan para analizar y reproducir características vocales.

– **Pitch Marking**: El algoritmo se utiliza en el procesamiento del habla para detectar cambios en la frecuencia fundamental o el tono. Es útil para analizar la calidad de la voz y en la síntesis de voz.

– **Harvest**: Un algoritmo de extracción de tono que se utiliza para aplicar el tono de la voz original a la voz clonada.

– **DIO**: Es un método para estimar la frecuencia fundamental, son técnicas de procesamiento paralelo o distribuido en sistemas informáticos.

– **Modelo robusto para la estimación del tono vocal en música polifónica**: Se utiliza para estimar el tono vocal en música polifónica.

– **Modelos multilingües**: Estos modelos incluyen Multilingual v1 y Multilingual v2, que ofrecen estabilidad y soporte para 29 idiomas.

– **Turbo**: Turbo v2 y Turbo v2.5 son algoritmos de baja latencia optimizados para conversaciones en tiempo real, diseñados para quienes requieren hablar de manera rápida y sencilla.

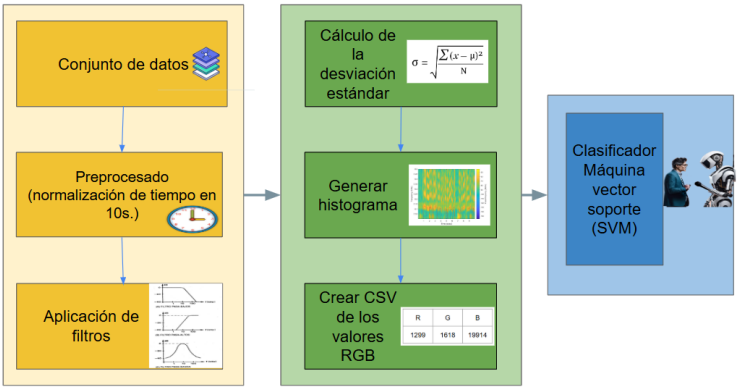


Fig. 1. Metodología propuesta.

– **CREPE**: Es un algoritmo de seguimiento de tono monofónico basado en una red neuronal simple para lograr la segmentación de notas monofónicas.

Conocer los algoritmos básicos detrás de la clonación de voz permite apreciar en profundidad cómo estas herramientas logran resultados precisos y de alta calidad.

3.1. Metodología

Como se puede apreciar en la figura 1, la metodología propuesta consta de tres etapas fundamentales:

1. **Procesamiento del conjunto de datos:**
   – Todos los audios fueron normalizados a una duración máxima de 10 segundos para garantizar uniformidad en el procesamiento.
   – Se verificó que todos los archivos estuvieran en formato .WAV; los que no cumplían con esta condición fueron convertidos utilizando scripts en Python.
   – Se aplicaron filtros digitales de tipo Butterworth de orden 5 con los siguientes parámetros:
     • Filtro pasa baja: Atenuación de frecuencias superiores a 500 Hz.
     • Filtro pasa alta: Atenuación de frecuencias inferiores a 2000 Hz.
     • Filtro pasa banda: Permite el paso de frecuencias entre 500 Hz y 2000 Hz.

     Estos filtros permiten resaltar componentes relevantes de la voz humana y reducir el ruido.
2. **Extracción de características:**
   – Se extrajo la frecuencia fundamental (F0) de cada audio utilizando un modelo convolucional especializado (convModel.mat), previamente entrenado para la estimación de tono en señales monofónicas [11].

- A partir de los valores de F0 extraídos, se construyeron histogramas que registran la distribución de frecuencias. Estos histogramas fueron representados mediante los canales de color Rojo (R), Verde (G) y Azul (B), lo cual permite organizar los datos en una estructura matricial.
- Como se menciona en el artículo [12], los histogramas de características acústicas pueden ser utilizados para el análisis de señales de voz. Sin embargo, a diferencia de dicho enfoque, que aplica una transformada de Fourier y una red neuronal para procesar la imagen del histograma, en el presente trabajo se utilizan directamente los valores de intensidad de píxeles en los canales RGB como vectores de entrada para el clasificador Máquina de Vectores de soporte (VSM por sus siglas en inglés), permitiendo una clasificación binaria entre voces naturales y clonadas sin recurrir a arquitecturas profundas.
- Esta estrategia permite capturar variaciones relevantes en la distribución de F0 de manera estructurada, facilitando al clasificador la detección de patrones distintivos entre ambas clases de voz.

3. **Clasificación:**
   - Los histogramas procesados fueron utilizados como vectores de características para entrenar un clasificador VSM.
   - Se probaron diferentes núcleos: lineal, polinomial, radial (RBF) y sigmoidal.
   - Se utilizó el 80 % del dataset para entrenamiento y el 20 % para validación.
   - La métrica principal de evaluación fue la exactitud, alcanzando un 95.54 % de exactitud utilizando filtro pasa baja y kernel sigmoidal.

4. Conjunto de datos

4.1. Conjunto de datos de voces naturales

CommonVoice [13] es un proyecto desarrollado por Mozilla que tiene como objetivo crear un conjunto de datos de voz abierto y diverso, destinado a mejorar la accesibilidad y la representación en las tecnologías de reconocimiento de voz. Este recurso se construye mediante las contribuciones de voluntarios, quienes participan grabando y validando frases en distintos idiomas.

Principales características:

- **Datos abiertos:** Las grabaciones recopiladas se distribuyen bajo la licencia CC0 (dominio público), lo que garantiza su disponibilidad sin restricciones. Esto permite que los datos sean utilizados en investigación, desarrollo de software y en la implementación de tecnologías de voz.
- **Idiomas y diversidad:** El proyecto incluye soporte para más de 100 idiomas y está diseñado para capturar una amplia gama de acentos y dialectos, lo que contribuye a reflejar la diversidad lingüística y cultural a nivel global.

Tabla 1. Conjunto de datos de voces clonadas.

Conjunto	Tiempo de audios	Número de audios
Voces Clonadas Español	149 minutos	796
Voces Clonadas Inglés	97 minutos	460

Tabla 2. RVC Dataset.

Idioma	Género	Cantidad
Español	Mujer	100
	Hombre	100
	Subtotal	200
Inglés	Mujer	60
	Hombre	100
	Subtotal	160
Total		360

Tabla 3. Applio Dataset.

Idioma	Género	Cantidad
Español	Mujer	56
	Hombre	100
	Subtotal	156
Inglés	Mujer	64
	Hombre	140
	Subtotal	204
Total		360

– **Contribuciones voluntarias:** La plataforma permite que cualquier persona participe leyendo frases para grabar su voz o revisando las grabaciones de otros para validar su calidad. Este enfoque colaborativo es fundamental para la construcción del conjunto de datos.

– **Propósito del proyecto:** La iniciativa busca democratizar el acceso a las tecnologías de voz, reduciendo sesgos en los sistemas existentes y promoviendo herramientas inclusivas que representen a una mayor variedad de usuarios.

4.2. Conjuntos de datos de voces generados artificialmente

Se plantea la necesidad de abordar la limitación encontrada en los conjuntos de datos existentes de voces clonadas, los cuales, como se mencionó anteriormente, no son óptimos, ya que son distinguibles [14]. Por consiguiente, se realizó la generación de un nuevo conjunto de datos que abarca aproximadamente 1300 audios clonados, como se muestra en la tabla 1. Para este propósito, se utilizan 14 audios por voz, lo que resulta en un estimado de 100 voces diferentes, tanto masculinas como femeninas.

Los audios presentes en este dataset son de tiempos variables, desde los 11 segundos hasta los 2 minutos. En las tablas 2, 3, 4, y 5 se muestra información de los conjuntos de datos.

Tabla 4. Conjunto de datos Elevenlabs.

Idioma	Género	Cantidad
Español	Mujer	60
	Hombre	40
	Subtotal	100
Inglés	Mujer	40
	Hombre	60
	Subtotal	100
Total		200

Tabla 5. Conjunto de datos Voice.ia.

Idioma	Género	Cantidad
Español	Mujer	40
	Hombre	60
	Subtotal	100
Inglés	Mujer	200
	Hombre	36
	Subtotal	236
Total		336

4.3. Procesado de los audios

Se verifica que los archivos de audio estén en formato ".WAV", por lo que el primer paso consiste en transformar aquellos que no cumplen con este requisito al formato adecuado. Para ello, se emplea un código en Python que analiza todos los archivos de audio en la carpeta y, en caso de ser necesario, realiza la conversión correspondiente.

A continuación, se aplican los filtros de "pasa baja", "pasa alta" y "pasa banda" [15] a los audios en formato correcto. Posteriormente, se calcula la frecuencia fundamental y se generan los histogramas, los cuales se almacenan en un archivo de texto.

Los datos de los histogramas se dividen en los canales "R", "G" y "B", registrando la frecuencia de los valores en un rango de 0 a 255 dentro de un archivo de Excel. Final- mente, estos valores son utilizados en una máquina de soporte vectorial, creando un conjunto de datos para cada filtro por separado.

5. Etapa 2: Extracción de características

En [16] se señala que existe una diferencia significativa entre la frecuencia fundamental (F0) promedio de hombres y mujeres. En términos generales, la F0 promedio es notablemente más alta en mujeres y el rango de F0 en Hz es más amplio en comparación con el de los hombres. A pesar de que las diferencias en el rango de F0 entre sexos desaparecen cuando se expresan en semitonos o como un factor de modulación, los valores promedio de F0 continúan siendo distintos. En resumen, la F0 promedio de los hombres no es equivalente a la de las mujeres.

El factor de modulación, en el contexto de la voz, se refiere a la variación de la frecuencia fundamental (F0) de la voz a lo largo de un período de tiempo. Esta medida indica cuánto fluctúa la frecuencia fundamental alrededor de su valor

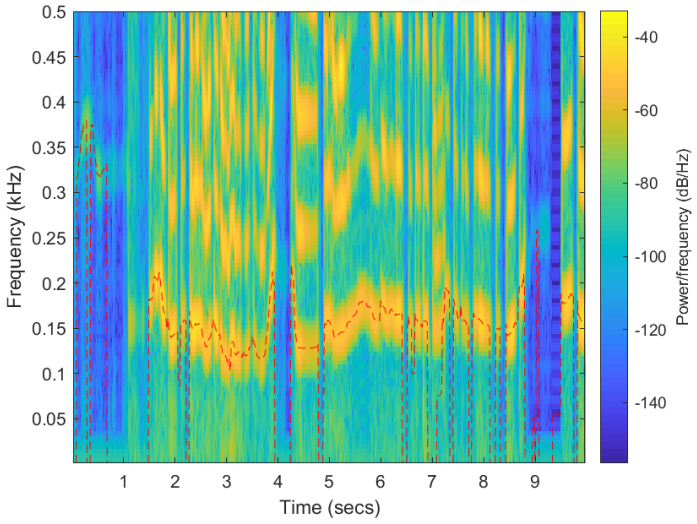


Fig. 2. Histograma de voz natural.

promedio. La modulación de frecuencia puede ser cuantificada en porcentaje o en unidades de frecuencia (como Hz). Un alto factor de modulación sugiere que la voz presenta numerosas variaciones en tono, lo cual puede interpretarse como expresividad o variabilidad vocal.

Teniendo en cuenta lo anterior, se usó el modelo de `Crepe-tiny` [11] para el procesamiento de $hf0$, el cual es un rastreador de tono monofónico basado en una red neuronal convolucional poco profunda que opera sobre la función de autocorrelación normalizada en el dominio del tiempo. Dicho código ha sido modificado para ser capaz de procesar diversos datos, ya que su versión base trabajaba únicamente con un audio a la vez y no se guardaban los histogramas generados ni el vector de datos.

En la figura 2 se puede observar cómo el histograma muestra en el eje X el tiempo en segundos, desde los 0 hasta los 10, en color se muestra la potencia y en el eje Y se muestra la frecuencia de la voz.

6. Experimentación

6.1. Especificaciones de hardware, versión de Python y librerías utilizadas

El equipo utilizado para desarrollar el proyecto consta de un CPU Intel i7 -600K, 32Gb de memoria RAM DDR4 y una GPU NVIDIA GeForce RTX 3080 Ti de 12Gb de VRAM. Se trabajó con Python 3.9.13, las librerías utilizadas junto a su versión y una descripción corta se encuentran en la tabla 6.

Se seleccionó el uso de Máquinas de Vectores de Soporte (SVM) debido a su alta capacidad para resolver problemas de clasificación binaria en espacios de alta dimensión, como el caso de la representación mediante histogramas de F0 [17], con los kernels “lineal, polinomial, gaussiano y sigmoide”, haciendo uso de los hiper

Tabla 6. Librerías utilizadas en el proyecto.

Librería	Versión	Descripción
matplotlib	3.9.2	Librería de visualización para crear gráficos en 2D.
numpy	1.23.5	Librería para cálculos numéricos.
pandas	1.5.3	Librería para análisis y manipulación de datos.
librosa	0.10.2.post1	Librería para el procesamiento de audios.
CUDA	11.2.67	Librería para hacer uso de GPU.
crepe	0.0.16	Librería para calcular f0 de los audios.
sklearn	1.6.1	Librería para el uso de máquina de soporte vectorial.

Tabla 7. Resultados con la exactitud.

Características	Kernel	Exactitud
Pasa banda	Lineal	0.9018
Pasa banda	Polinomial	0.8661
Pasa banda	RBF	0.9018
Pasa banda	Sigmoideo	0.8304
Pasa alto	Lineal	0.8571
Pasa alto	Polinomial	0.8750
Pasa alto	RBF	0.8571
Pasa alto	Sigmoideo	0.8393
Pasa baja	Lineal	0.9018
Pasa baja	Polinomial	0.8482
Pasa baja	RBF	0.8571
Pasa baja	Sigmoideo	0.9554
Sin filtro	Lineal	0.9018
Sin filtro	Polinomial	0.8839
Sin filtro	RBF	0.9018
Sin filtro	Sigmoideo	0.9196

parámetros por defecto en la librería sklearn. Del conjunto de datos se utilizó el 80 % para entrenamiento y el 20 % para validación. En la tabla 7, se observan los resultados de los filtros con todos los kernels utilizados.

El análisis de los resultados obtenidos mediante la aplicación de diferentes filtros y kernels de SVM revela patrones significativos que pueden guiar la selección del método más adecuado para tareas específicas de clasificación de señales.

– **Filtro Pasa Banda:** El filtro pasa banda muestra un rendimiento consistente con los kernels lineal y RBF, ambos alcanzando una exactitud de 0.9018. Sin embargo, el kernel RBF logra una sensibilidad perfecta (1.0000), lo que sugiere una capacidad superior para identificar correctamente las señales positivas, aunque con una especificidad moderada. Este kernel podría ser preferido en aplicaciones donde es crucial minimizar los falsos negativos.

– **Filtro Pasa Alta:** En el caso del filtro pasa alta, el kernel polinomial presenta un buen equilibrio entre exactitud (0.8750) y alta sensibilidad (0.9286), pero con una especificidad constante de 0.8214. Esto indica que el kernel polinomial es eficiente en la detección de verdaderos positivos, aunque su capacidad para discriminar los

verdaderos negativos es limitada. Este filtro y kernel pueden ser útiles en escenarios donde la detección correcta de señales es prioritaria sobre la especificidad.

– **Filtro Pasa Baja:** El filtro pasa baja combinado con el kernel sigmoide destaca al lograr la mayor exactitud (0.9554) y alta especificidad (0.9286), lo que sugiere una excelente capacidad tanto para identificar correctamente las señales positivas como para minimizar los falsos positivos. Este método sería ideal para aplicaciones que requieren una alta exactitud general y una excelente discriminación entre señales positivas y negativas.
– **Sin Filtro:** Sin la aplicación de un filtro, el kernel sigmoide también muestra un rendimiento notable, con una exactitud de 0.9196 y una alta especificidad (0.9464). Esto sugiere que este método es altamente efectivo para aplicaciones donde se necesita un análisis detallado de las características de frecuencia de las señales.

7. Conclusión

Los resultados obtenidos en este estudio demuestran que las características extraídas del histograma de la frecuencia fundamental ($f0$) son una herramienta eficaz para diferenciar entre voces clonadas y naturales. El uso de un clasificador basado en Máquinas de Vectores de Soporte (SVM) permitió alcanzar una precisión del 95 %, lo que indica un alto nivel de acierto en la identificación de voces generadas artificialmente. Estos hallazgos validan la utilidad de la frecuencia fundamental como una característica distintiva en el análisis de señales de voz y refuerzan su potencial en aplicaciones de autenticación y detección de fraudes en entornos digitales. Con estos resultados, se concluye que el objetivo del artículo se ha cumplido satisfactoriamente, destacando la relevancia de esta metodología en el estudio y la seguridad de los sistemas de clonación de voz.

Actualmente, el conjunto de datos empleado no ha sido publicado debido a que forma parte de un trabajo de investigación en curso dentro de un programa de maestría. Aún se están evaluando distintas metodologías y realizando experimentaciones complementarias. Una vez concluido el proceso académico y obtenido el grado correspondiente, se planea poner a disposición de la comunidad científica el dataset completo junto con los scripts utilizados, con el fin de favorecer la reproducibilidad y futuras investigaciones en este campo.

Agradecimientos. Se agradece a SECIHTI por el apoyo económico brindado mediante la beca para los estudios de maestría.

Referencias

1. S. Achanta., R. Maas., R. Clark.: On-Device Neural Speech Synthesis. arXiv, pp. 1–7 (2021), doi: 10.48550/arXiv.2109.08710.
2. Extracta.ai: Exploring the Impact of AI Voice Cloning: Transforming Digital Storytelling. URL: https://extracta.ai/exploring-the-impact-of-ai-voice-cloning-transforming-digital-storytelling (2024)

3. Discord: Discord. URL: https://discord.com (2022)
4. GitHub: GitHub. URL: https://github.com (2024)
5. Meta AI: Deepfake Detection Challenge Results: An Open Initiative to Advance AI. URL: https://ai.meta.com/blog/deepfake-detection-challenge-results-an-open-initiative-to-advance-ai (2023)
6. Eleven Labs: Eleven Labs. URL: https://elevenlabs.io (2024)
7. VocaliD: Vocalid. URL: https://vocalid.ai (2024)
8. GitHub: Iahispano/applio-rvc-fork. URL: https://github.com/IAHispano/Applio-RVC-Fork (2023)
9. GitHub: RVC-Project/Retrieval-based-Voice-Conversion-WebUI. URL: https://github.com/RVC-Project/Retrieval-based-Voice-Conversion-WebUI (2023)
10. Voice.ai: Voiceai. https://voice.ai (2023)
11. Pradeepiit: Hf0. URL: https://github.com/Pradeepiit/hf0 (2024)
12. Lim, S.-Y., Chae, D.-K., Lee, S.-C.: Detecting Deepfake Voice Using Explainable Deep Learning Techniques. Applied Sciences, 12(8), pp. 1–15 (2022) doi: 10.3390/app12083926.
13. Mozilla: Common Voice. URL: https://commonvoice.mozilla.org/en/datasets (2024)
14. Blue, L., Warren, K., Abdullah, H.: Who Are You (I Really Wanna Know)? Detecting Audio DeepFakes Through Vocal Tract Reconstruction. In: 31st USENIX Security Symposium, pp. 2691–2708 (2022) URL: https://www.usenix.org/conference/usenixsecurity22/presentation/blue
15. Corchete, V.: High-Pass, Low-Pass and Band-Pass Filtering. Universidad de Almería, pp.1–5 (2019) doi: 10.13140/RG.2.2.25817.67686.
16. Traunmüller, H., Eriksson, A.: The Frequency Range of the Voice Fundamental in the Speech of Male and Female Adults. URL: https://www.researchgate.net/publication/240312210 (1995)
17. Carmona, E. J.: Tutorial sobre Máquinas de Vectores Soporte (SVM). pp. 1–27. URL: https://www.researchgate.net/publication/263817587.

ISSN 1870-4069

Gas Metal Arc Welding Dataset for Computer Vision Quality Assessment

José A. López-Islas[1,2], Oscar Camacho-Nieto[2], Yenny Villuendas-Rey[2]

[1] Instituto Politécnico Nacional,
Escuela Superior de Ingeniería Mecánica y Eléctrica, Unidad Azcapotzalco,
Mexico

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

{jlopezi;ocamacho;yvilluendasr}@ipn.mx

Abstract. Robotic Gas Metal Arc Welding is commonly used for several industrial purposes and requires expedited quality assurance feedback due to the nature of the robotic welding. Usually, in small and medium enterprises, this feedback comes from a human expert. There is an interest in digitalizing the experts' knowledge regarding welding quality to train computer vision systems for automatic quality assessment. Having the correct data is crucial for this task. In this paper, we introduce a novel dataset of robotic Gas Metal Arc Welding images belonging to four categories: good welding, welding that can be reworked by the robot itself, welding that can be reworked by a human expert, and welding with unsalvable fails (usually named as scrap). The proposed dataset includes the experts' knowledge and annotates the individual welding seam quality. It is publicly available and can be used to train computer vision systems for welding quality assurance.

Keywords: Robotic gas metal arc welding, quality assessment, computer vision dataset.

1 Introduction

Robotic welding is widely used in the metalworking industry and is most often supervised by qualified and trained personnel for weld seam verification. With increasing technological development worldwide, artificial intelligence is being integrated into industrial processes. It seeks to digitize the knowledge of specialized supervisors, both to verify process quality and for its optimization.

Among the various types of welding, robotic gas metal arc welding (GMAW) stands out for its wide-ranging applications in diverse industries such as automotive and manufacturing, among others. GMAW consists of a wire-shaped welding electrode, which, through the application of a controlled electrical charge by means of the welding

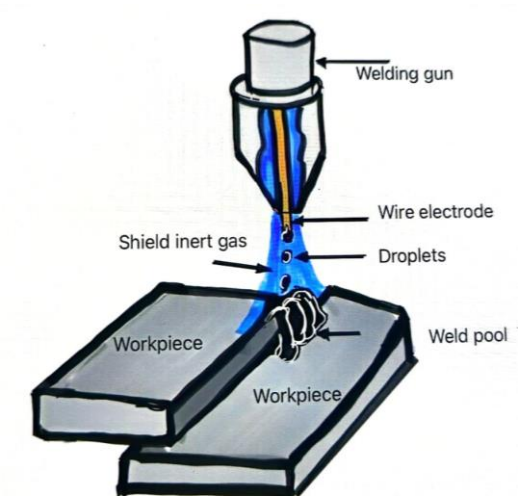


Fig. 1. Main elements of GMAW.

gun [1], causing it to reach its melting point and drip continuously towards the material (weld pool). The process is carried out progressively, so it requires an automatic wire feeder. When this welding process is carried out, an active or inert gas (shield of inert gas) is used to protects it from atmospheric contamination [2], as shown in Figure 1.

The application of robotic GMAW welding has undergone significant development in terms of automation and supervision, seeking to create intelligent welding systems. This approach to applying artificial intelligence models is known as "Intelligent Welding Manufacturing (IWM)" [3]. These systems can include sensing technology, followed by an image processing method based on prior knowledge, and acquire weld characteristics by measuring pixels obtained from the images in real time.

Some research has also focused on collecting electrical and mechanical data from the welding process in addition to capturing images for the design of an intelligent system [4].

However, most research is conducted with process data from different environments, which are not made public for research. Therefore, it is difficult to find images of robotic GMAW weld seams to train computer vision systems.

To address this limitation, this article proposes a robotic GMAW image dataset. This dataset stands out for its low acquisition cost and for its ability to include human expert labeling of weld seam characteristics.

The remainder of the paper is as follows: Section 2 covers some of the existing works in image-based automatic quality assurance for robotic GMAW. Section 3 provides an explanation of the robotic process and the characteristics of the welding parts. Section 4 describes the proposed dataset, and Section 5 covers the conclusions of the paper.

Fig. 2. Metallic parts to be welded.

Table 1. Chemical properties of the metallic parts.

Chemical element	Composition
C	0.100
Mn	0.600
P	0.030
S	0.035
Cu	0.200
Ni	0.200
Cr	0.150
Mo	0.060
V	0.008
Cb	0.008
Ti	0.008

2 Related Works

Several researchers have addressed the topic of GMAW quality assurance by computer vision. Li et al. used Convolutional Neural Networks (CNNs) for defect prediction [5]. They recorded the welding process with a welding camera. Then, they used the video frames as a molten pool of images to construct the dataset. Unfortunately, in [5] is stated that "The datasets generated and analyzed during the current study are not publicly available due the confidentiality of the data".

Kim et al. also created a dataset for classifying penetration conditions in GMAW processes by CNNs. They also stated that "The data are not publicly available due to

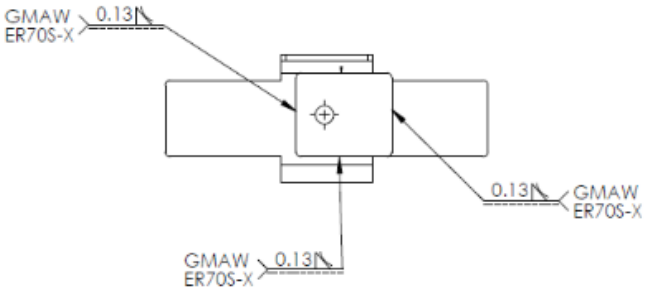


Fig. 3. Welding specifications.

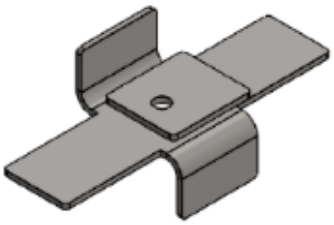


Fig. 4. 3D isometric view of the parts.

privacy", although it can be provided by the corresponding author upon reasonable request [6].

Diaz-Cano et al. [7] created a set of images of weld seams considered acceptable, as well as seams with defects such as lack of penetration or undercuts. Their images correspond to FCAW and GMAW welding processes and were captured with a high-resolution camera (Ensenso model N35) positioned on the robotic arm. The camera was moved in seven poses for each weld seam, each with different luminosities. The images captured are publicly available at the following links:

- https://universe.roboflow.com/weldingpic/good-op-lop-under/dataset/1
- https://universe.roboflow.com/weldingpic/weld_fcaw_gmaw/dataset/1
- https://universe.roboflow.com/weldingpic/weld_fcaw_gmaw/dataset/2

However, in their work, the images correspond to straight lines of welding beads, with the purpose of measuring the thickness and height (in the best of cases) in a controlled, smooth relief without any distortion in its trajectory, therefore, the deposition of the liquid metal is only dispersion.

Research has taken into account different cameras for the welding process and certain additional elements such as the distance, width, area, and angle of the welding electrode. When using low-cost cameras, such as a webcam, it is necessary to apply

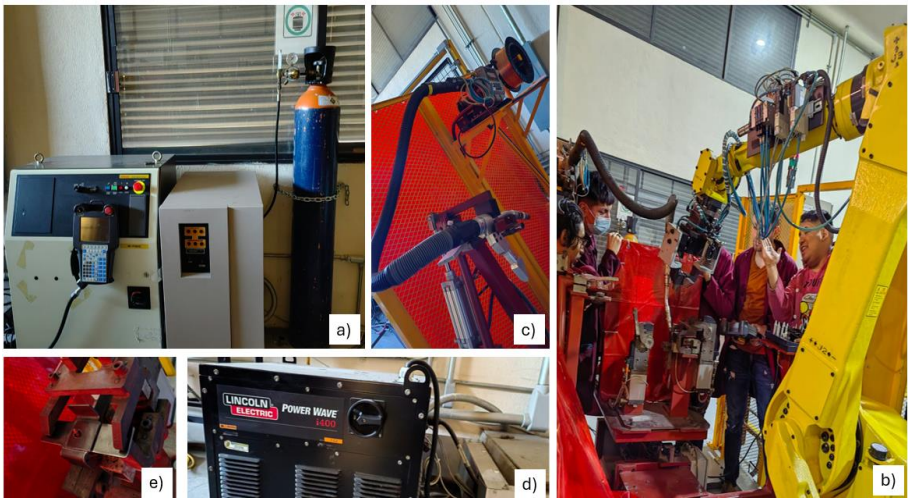


Fig. 5. GMAW Welding Station. (a) Fanuc R30ia Controller with Ipendant, Vogar Voltage Regulator and Gas Tank, (b) Fanuc M70iC/50 Robot, (c) AutoDrive 4R90 Automatic Microwire Feeder and Lincoln Electric Welding Torch, (d) Lincoln Electric Power Wave i400 Welding Controller, and (e) Automatic Automotive Workholding System.

more processing methods such as YOLOv5 (You Only Look Once), PAN (Path Aggregation Network), CNN (Convolutional Neural Network), and FPN (Feature Pyramid Networks) [6].

In Spain, one of the investigations used two different robotic welding processes: Flux-Cored Arc Welding (FCAW) and Gas Metal Arc Welding (GMAW). The inspection technique used in this case was to train a neural network with a series of 2D images of welding beads, taken by special equipment to offer different shades of light, and a high-resolution industrial camera, placed on the end effector on another robotic arm, the technique, follows a deep learning geometric model in CNN and YOLOv8, the results were achieved in binary and multiclass classification [8]. In India, the focus of research is on a machine vision-based algorithm for robotic weld path detection, gap measurement, and weld length calculation in the GMAW process. The camera used is a low-cost webcam located on the robotic arm, and the captured images are converted to grayscale, and YOLOv5, PAN, CNN, and Feature Pyramid Network (FPN) are used for variance scaling in object detection [9]. In this research, we focused on the use of low-cost cameras.

In addition, our proposal corresponds to a higher level of complexity. It is derived from the fact that it is not only the deposition of the electrode material, but it also entails penetration into both materials for the purpose of a permanent union. During the welding path, there will always be an imaginary line between both joining materials that will not be homogeneous. In addition, the technique in cutting the material will make the edges rough, not smooth. This causes the electric arc that is generated to not be the same throughout its path, and the movement of the materials due to their thermal expansion generates different welding patterns along their path.

The next section presents the details regarding the materials and methods used in our research.

3 Materials and Methods

This section details the elements necessary to obtain the welding seams, the capture medium, the number of images, and the parameters used for the classification of the process quality.

3.1 Metallic Parts

The parts to be welded are two and are made of 0.25-inch thick black sheet, whose classification is hot-rolled steel sheet, designated as Commercial Steel Type A (CS Type A), as illustrated in Figure 2. These pieces were machined with a water jet. The mechanical properties of the parts are "Yield Strength" (Ksi): 30 to 50, and Elongation in 2 in. [50 mm] % ≥25. The chemical properties of the parts are provided in Table 1.

The welding implemented was by Pulse Spray Transfer, whose characteristics are: Wire feed: 145, Trim: 1.05, Voltage: 29V, Travel speed: 36, Pulse: 0.045, Feedback Current: 199.4, and Air mix: Ar + CO2.

The welding specifications are detailed in Figure 3, and how both parts must be superimposed is shown with a 3D isometric view in Figure 4.

3.2 Welding Equipment

We used the welding equipment of the Automotive Cell laboratory at the School of Mechanical and Electrical Engineering, Azcapotzalco Unit, of the Instituto Politécnico Nacional. It includes a robotic welding station (Figure 5). It contains an Industrial robotic system. Figure 5a shows a model r30iA controller commanded by an IPendant, as well as its Vogar voltage regulator; Figure 5b shows a Fanuc robot model M710iC/50 with a Schunk brand gripper end effector to hold the welding torch.

In addition, the station has an automatic welding system, including a GMAW welding torch with an AutoDrive 4R90 automatic feeder (Figure 5c), a Power Wave i400 controller from Lincoln Electric (Figure 5d), and the active gas tank (MAG) with a combination of Ar + CO2 (Figure 5a). Finally, the station also includes an automatic clamping system, as shown in Figure 5e), is used to hold the automotive part while it is being welded.

3.3 Image Acquisition

For welding, both parts must be aligned using a locating pin in the center, as shown in Figure 6a. Both ends of the lower piece, which is cross-shaped, are held by an automatic clamping device designed for the piece. This prevents movement of the piece at the time of welding, which may occur due to thermal expansion caused by the process.

(a) (b)

Fig. 6. Metallic parts to be welded. (a) In clamps and (b) With welding seams.

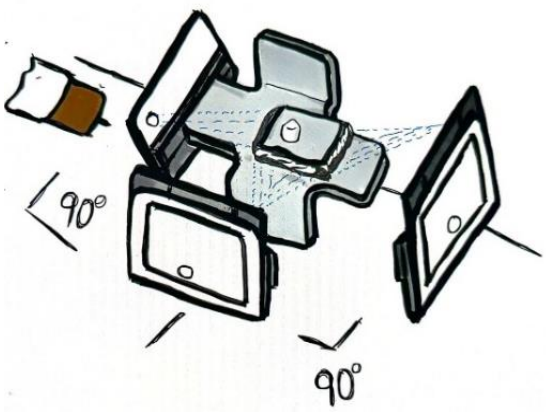
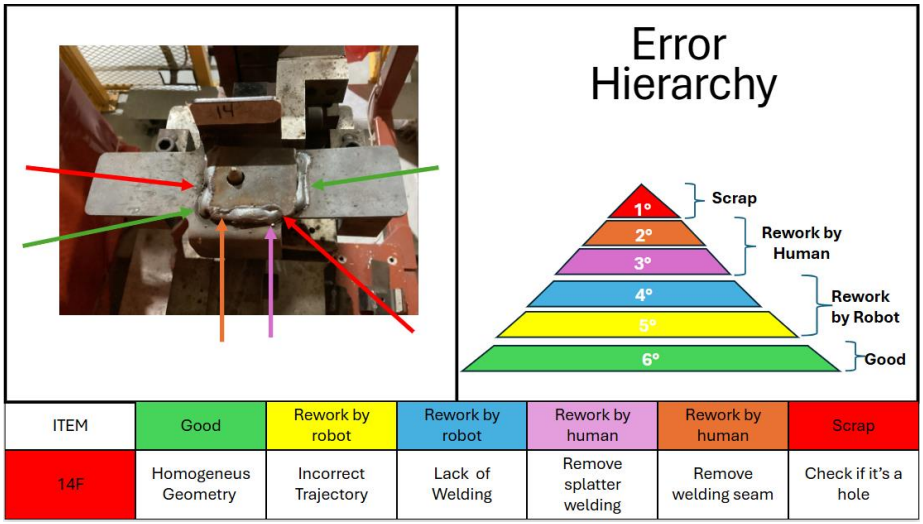


Fig. 7. Image acquisition of the welded parts. Note how the capturing device is arranged in three different positions to acquire a good view of the weld seams.

The first weld bead must be made from the front, in order to be able to remove the part's holding element and continue with the side beads, ensuring that the part will not move again, leaving only the central locating pin responsible for the correct position of the part when it is released, as shown in Figure 6b.

The images were captured after applying the three welding paths, ensuring a parallelism with the weld bead to avoid shadows that could be reflected between the piece to be welded and the weld bead (Figure 6). The device used to capture the images was a 9th Generation iPad, with an 8MP wide-angle lens with an f/2.4 aperture and 5x digital zoom, and HDR for photos.



ITEM	Good	Rework by robot	Rework by robot	Rework by human	Rework by human	Scrap
14F	Homogeneus Geometry	Incorrect Trajectory	Lack of Welding	Remove splatter welding	Remove welding seam	Check if it's a hole

Fig. 8. Example of welded parts (image 14F) with the annotated characteristics (left) and the proposed error hierarchy for the quality assessment of the weld seams (right) in a colorimetric scale.

4 Results

A crucial aspect of this proposal is the human expert's identification of potential flaws in the weld seam. To this end, a hierarchy of errors is proposed (Figure 7). These errors range from the most serious (top of the pyramid) to the good weld (base of the pyramid), including the less serious errors. Thus, six elements are considered that can be visually recognized by the weld seam geometry. As can be seen, if there is at least one first-degree error in any of the weld seams of the parts to be welded, the entire part must be scrapped, as it would no longer be suitable for rework.

Similarly, if there is at least one error that requires the part to be worked by a human operator, it would be labeled as such. Something similar happens, for example, with a part that has two good weld seams and one that requires robotic rework. In this case, the entire part is labeled as "robotic rework," not as good. Even if most weld seams were correct, the presence of an error leads to labeling the part according to the highest-ranking error.

These elements represent a significant contribution to the state of the art, as the proposed dataset not only offers image labeling but also includes the knowledge of the human expert, identifying, for each weld seam, its geometric characteristics and corresponding errors on a colorimetric scale.

Depending on the characteristics of the weld seams, the images were labeled into four categories, as in Figure 7: Good, Rework by Robot (Robot), Rework by Human (Human), and Scrap.

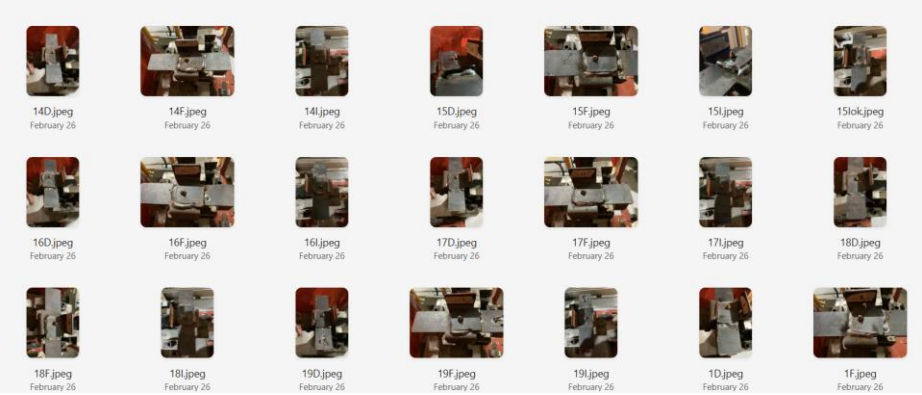


Fig. 9. Example of images in the proposed dataset.

The developed dataset included the analysis of weld seams from 30 automotive parts. Three weld seams were created on each part, for a total of 90 images. Figure 8 shows some of the images obtained. In addition to the images labeled into the four aforementioned categories, the database includes, for each image, labeling of the weld seam characteristics, according to the proposed error hierarchy.

The dataset will be donated to the Machine Learning Repository of the University of California at Irvine, to be publicly available worldwide. In the meantime, interested readers can find it in the following institutional link: https://correoipn-my.sharepoint.com/:f:/g/personal/yvilluendasr_ipn_mx/EmD7pChS4zpLkIZRsc6ldE0BgIzsgX9F4sErXxgCFxswlQ?e=NLZDo9

In the following, we summarize the criteria considered in the labeling process. First, A Good weld is homogeneous, without any porosity or black shadow that could indicate a perforation or spatter when welding (Figure 10a).

The weld that can be corrected (reworked) by the robot (Robot label) appears to be homogeneous but contains spatter in the form of small spheres around the weld bead (Figure 10b). These are generally visible as part of the unwelded side of the part, so the robot can re-pass through them, correcting the path.

As shown in Figure 10c, the weld bead is not homogeneous; it is thin in some places, bulged in the middle, and has a lot of spatter. However, the color is light, meaning it was not overheated, so the weld can be removed using a chisel and hammer. Thus, a human operator can rework the weld bead and not discard the part. Such seams are labelled as Human.

Finally, as shown in Figure 10d, round-shaped shadows appear on the weld bead, so there is a risk of overheating of the part and possible perforation of the part, so it would not withstand a second welding application. Therefore, they are labeled as Scrap.

Fig. 10. Examples of weld seams belonging to the four categories. (a) Good, (b) Robot, (c) Human, and (d) Scrap.

As shown in Figure 10c, the weld bead is not homogeneous; it is thin in some places, bulged in the middle, and has a lot of spatter. However, the color is light, meaning it was not overheated, so the weld can be removed using a chisel and hammer. Thus, a human operator can rework the weld bead and not discard the part. Such seams are labelled as Human. Finally, as shown in Figure 10d, round-shaped shadows appear on the weld bead, so there is a risk of overheating of the part and possible perforation of the part, so it would not withstand a second welding application. Therefore, they are labeled as Scrap.

Of the 90 images analyzed, we have a classification of 36 Good, 19 manual rework (Human), 30 robotic rework (Robot), and five bad seams for discard (Scrap). The images were divided into train and test sets by a human expert. It allows the replicability of the experiments researchers can make in the future and guarantees the representativity of the weld seams.

5 Conclusions

This paper focused on quality assessment of GMAW in automotive parts. This process entails high complexity derived from the fact that it is not only the deposition of the electrode material but also the penetration into both materials for the purpose of a permanent union. In addition, the technique in cutting the material will make the edges rough, not smooth. This causes the electric arc that is generated to not be the same throughout its path, and the movement of the materials due to their thermal expansion generates different welding patterns along their path.

We proposed a novel image dataset from real-world welding images, that includes the analysis of weld seams from 30 automotive parts. In addition, the dataset not only offers labelled images but also includes the knowledge of the human expert, identifying, for each weld seam, its geometric characteristics and corresponding errors on a hierarchical colorimetric scale. The images were labeled into four categories: Good, Rework by Robot (Robot), Rework by Human (Human), and Scrap. In addition to the images labeled into the four aforementioned categories, the database includes, for each

image, labeling of the weld seam characteristics according to the proposed error hierarchy.

In future work, we want to increase the number of images. We also want to add a camera to the robotic arm to have more degrees of liberty for image capturing.

Acknowledgments. The authors would like to thank the Instituto Politécnico Nacional (Secretaría Académica, Secretaría de Investigación y Posgrado, ESIME Azcapotzalco, and Centro de Inno-vación y Desarrollo Tecnológico en Cómputo), the Secretaría de Ciencia, Humanidades, Tecnolo-gía e Innovación, and Sistema Nacional de Investigadores for their economic support to develop this work.

References

1. url: https://www.keyence.com.mx/ss/products/measure/welding/
2. url: https://www.millerwelds.com/resources/article-library/the-history-of-welding
3. Hou, Z., Xu, Y., Xiao, R.: A Teaching-Free Welding Method Based On Laser Visual Sensing System in Robotic GMAW. The International Journal of Advanced Manufacturing Technology, 109, pp. 1755–1774 (2020).
4. Wang, X., Chen, Q., Sun, H.: GMAW Welding Procedure Expert System Based On Machine Learning. Intelligence & Robotics 3, pp. 56–75 (2023).
5. Li, H., Ma, Y., Duan, M.: Defects Detection of GMAW Process Based On Convolutional Neural Network Algorithm. Scientific Reports 13, 212–219 (2023).
6. Kim, D.-Y., Lee, H.W., Yu, J.: Application of Convolutional Neural Networks for Classifying Penetration Conditions in GMAW Processes Using STFT of Welding Data. Applied Sciences 14, 4883 (2024).
7. Diaz-Cano, I., Morgado-Estevez, A., Rodríguez Corral, J.M.: Automated Fillet Weld Inspection Based on Deep Learning From 2D Images. Applied Sciences 15, 899 (2025).
8. Pico, L.E.A., Marroquín, O.J.A., Lozano, P.A.D.: Application of Deep Learning for the Identification of Surface Defects Used in Manufacturing Quality Control and Industrial Production: A Literature Review. Ingeniería, 28, pp. 1–20 (2023).
9. Reddy, J., Dutta, A., Mukherjee, A.: A Low-Cost Vision-Based Weld-Line Detection and Measurement Technique for Robotic Welding. International Journal of Computer Integrated Manufacturing, pp. 1–21 (2024).

ISSN 1870-4069

Segmentación automática de tumores cerebrales usando K-means

Kay García-Sánchez, Daniel Cantón-Enríquez, Hugo Jiménez-Hernández, Luis Antonio Díaz-Jiménez, Ana Marcela Herrera-Navarro, Jorge Luis Pérez-Ramos, Selene Ramírez-Rosales, Carlo Giovanni Cetina-Camacho

Universidad Autónoma de Querétaro, Facultad de Informática, México

kgarcia@uaq.mx

Resumen. La segmentación automática de tumores cerebrales en imágenes médicas es un desafío debido a la variabilidad en la morfología y el contraste de los tumores. En este trabajo, se propone un marco de referencia basado en detección de contornos mediante el operador de Sobel y segmentación con el algoritmo K-means, refinando los resultados con filtrado morfológico. La metodología implementada permite diferenciar tejido sano de masa tumoral, adaptándose a variaciones de intensidad y morfología. Los resultados muestran que la técnica propuesta es capaz de segmentar tumores en casos complejos, incluso en casos con bordes difusos. Este enfoque proporciona una alternativa reproducible para la segmentación automática de tumores, facilitando la toma de decisiones clínicas y promoviendo el uso de inteligencia artificial en salud.

Palabras clave: Segmentación automática, tumores cerebrales, clustering, aprendizaje automático, inteligencia artificial médica.

Automatic Brain Tumor Segmentation Using K-means

Abstract. Automatic segmentation of brain tumors in medical images is a challenge due to the variability in tumor morphology and contrast. In this work, we propose a reference framework based on contour detection using the Sobel operator and segmentation with the K-means algorithm, refining the results with morphological filtering. The implemented methodology allows differentiating healthy tissue from tumor mass, adapting to variations in intensity and morphology. Results show that the proposed technique is capable of segmenting tumors in complex

cases, even in those with diffuse borders. This approach provides a reproducible alternative for automatic tumor segmentation, facilitating clinical decision-making and promoting the use of artificial intelligence in healthcare.

Keywords: Automatic segmentation, brain tumors, clustering, machine learning, medical artificial intelligence.

1. Introducción

Los tumores cerebrales presentan una amplia heterogeneidad en su gravedad determinada por su grado de malignidad según la OMS [1]. Este espectro clínico abarca desde gliomas de alto grado como el glioblastoma (grado IV), con un pronóstico desfavorable; hasta meningiomas generalmente benignos (grado I), que a pesar de su baja agresividad, pueden resurgir localmente [2]. Asimismo, los adenomas pituitarios, aunque mayoritariamente indolentes, tienen el potencial de afectar funciones endocrinas críticas [3].

La detección de tumores cerebrales enfrenta importantes desafíos técnicos, entre ellos la dificultad para diferenciar radiológicamente la progresión tumoral de los efectos postratamiento en diversas afecciones oncológicas [4], así como la alta variabilidad interobservador en tumores con bordes difusos [5]. En este contexto, la segmentación automática surge como una solución prometedora para reducir la inconsistencia en las anotaciones manuales [6]. Además, de cerrar la brecha entre los datos de investigación y las condiciones clínicas reales [7].

A nivel global, el Objetivo de Desarrollo Sostenible (ODS) 3 de las Naciones Unidas, centrado en "Salud y Bienestar", enfatiza la importancia de fortalecer los sistemas sanitarios mediante la adopción de tecnologías médicas innovadoras y accesibles [8]. Este objetivo está alineado con los esfuerzos por reducir la brecha en el diagnóstico temprano de tumores cerebrales en países de bajos ingresos, donde la falta de equipos de neuroimagen avanzada incrementa la mortalidad por gliomas no detectados [9]. Además, como señala Gupta et al. [10], la implementación de herramientas de inteligencia artificial de código abierto para la segmentación automática de tumores cerebrales podría democratizar el acceso a diagnósticos precisos, contribuyendo al cumplimiento de la meta 3.8 del ODS la cual busca cobertura sanitaria universal mediante al acceso equitativo a servicios de salud.

Por otro lado, la detección temprana de tumores cerebrales no solo mejora el pronóstico, sino que también preserva la calidad de vida de los pacientes. En el caso de los glioblastomas diagnosticados en estadios iniciales, la supervivencia media se duplica y las resecciones completas son más factibles [11]. Por otro lado, en meningiomas asintomáticos de alto riesgo, el seguimiento mediante MRI permite evitar complicaciones neurológicas graves [12]. Estas ventajas contrastan con las marcadas diferencias en tasas de supervivencia entre los distintos tipos tumorales: los pacientes con gliomas de alto grado tienen una supervivencia a 5 años inferior al 20%, mientras que en adenomas pituitarios no invasivos esta cifra alcanza el 95% [13,14].

La calidad de las imágenes de resonancia magnética es un factor clave en la precisión de la segmentación automatizada de tumores cerebrales. Existen protocolos con alta resolución espacial (≥ 1 mm^3) y secuencias como T2-FLAIR, que suprimen la señal del líquido cefalorraquídeo, mejorando así la delimitación de tumores infiltrativos [15]. No obstante, parámetros de adquisición subóptimos, como tiempos de eco prolongados, pueden reducir la precisión de los algoritmos de inteligencia artificial hasta en un 30% [16]. Además, artefactos comunes, como el movimiento del paciente o una baja relación señal-ruido ($SNR < 20$ dB), generan errores críticos en la segmentación de tumores pequeños, lo que hace necesario el uso de técnicas de corrección basadas en Deep Learning para mitigar su impacto [17,18].

En este artículo, se presenta un marco de referencia basado en K-means para segmentar tumores cerebrales. Este algoritmo, combinado con técnicas de preprocesamiento, permite una binarización eficiente de imágenes en escala de grises, facilitando la identificación de regiones tumorales. Su robustez ante variaciones de contraste lo hace una alternativa competitiva frente a otros algoritmos de agrupamiento como DBSCAN, que debido a su sensibilidad a parámetros como ϵ, convirtiéndolo en un método menos robusto en imágenes con variaciones locales de contraste, típicas en imágenes de resonancias magnéticas de tumores cerebrales [16].

El resto del artículo está organizado de la siguiente manera: se presentan los materiales y métodos utilizados para la segmentación automática de los tumores cerebrales; después, se muestran los resultados obtenidos de la experimentación; posteriormente, la discusión de los resultados; por último, las conclusiones y trabajos a futuro.

2. Materiales y métodos

2.1. Descripción del conjunto de datos

El dataset utilizado en este estudio fue obtenido del repositorio público *Brain Tumor MRI Dataset* [19] , el cual contiene 7,023 imágenes de resonancia magnética en formato JPG, clasificadas en cuatro tipos; véase Tabla 1. Las imágenes fueron preprocesadas para garantizar uniformidad en dimensiones (512 x 512 px) y contraste. Se incluyeron imágenes donde la región tumoral era claramente distinguible y se excluyeron aquellas con artefactos de movimiento severos (SNR $<$ 20 dB). Para mejorar la calidad del dataset, se realizaron pruebas de normalización de intensidades y eliminación de ruido utilizando filtros adaptativos.

2.2. Metodología

El proceso realizado para la segmentación automática de tumores cerebrales a partir de imágenes de resonancia magnética (MRI) consta de tres etapas principales: 1) preprocesamiento, donde, se buscan los contornos de las

Table 1. Distribución del dataset Brain Tumor MRI Dataset [19].

Tipo de tumor	Descripción	Cantidad
Meningioma	Crecimiento en meninges, ejerciendo presión sobre el cerebro.	1,621
Glioma	Proliferación de células anormales en una región cerebral.	1,645
Pituitario	Tumores en la glándula pituitaria, con afectaciones endocrinas.	1,757
Sin tumor	Imágenes sin evidencia de tumor.	2,000

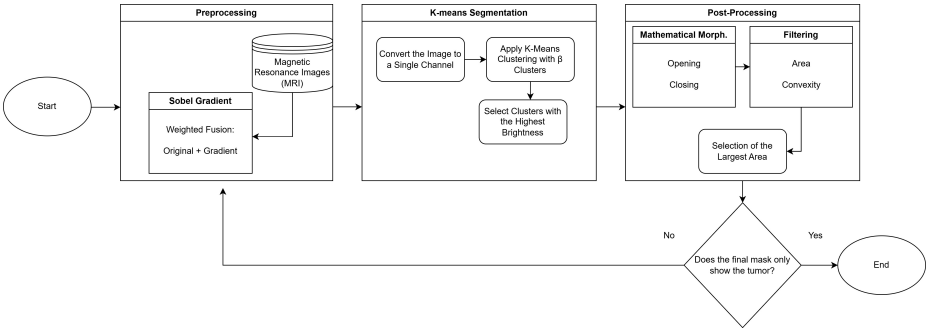


Fig. 1. Diagrama de flujo del marco de referencia propuesto.

imagénes MRI; 2) segmentación con K-means, donde, se agruparn los píxeles en dos clústeres (tejido sano y tumoral); y 3) postprocesamiento, donde, se seleciona la región tumoral dominante usando filtros de convexidad y área para eliminar artefactos.

La metodología empleada en este trabajo se muestra de forma general en el diagrama de flujo de la figura 1.

Detección de contornos: Primero, se aplicó el operador de Sobel, un filtro de convolución discreto ampliamente utilizado en visión por computadora para la detección de bordes. Este operador calcula una aproximación del gradiente de la imagen en dos direcciones: horizontal (eje X) y vertical (eje Y), lo que permite resaltar los contornos de los objetos presentes en la imagen [20]. Matemáticamente, se define mediante dos kernels de convolución de tamaño 33, como se muestra en la Ecuación 1:

$$G_x = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}, \quad G_y = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix}. \tag{1}$$

El núcleo convolucional G_x resalta variaciones abruptas de intensidad a lo largo del eje X, facilitando la detección de bordes verticales. De manera análoga, G_y permite identificar bordes horizontales al calcular los cambios de intensidad

en el eje Y. Por otro lado, la magnitud del gradiente proporciona una medida cuantitativa de la intensidad total de los bordes en la imagen, y se calcula mediante la Ecuación 2:

$$|G| = \sqrt{G_x^2 + G_y^2}. \tag{2}$$

Detección del Umbral Autómatico con K-means: El algoritmo K-means es un método de aprendizaje no supervisado utilizado para agrupar datos en K clústeres en función de sus similitudes. Su objetivo es minimizar la varianza intra-clúster, asignando cada dato al clúster cuyo centroide se encuentre más cercano. Este proceso se describe matemáticamente en la Ecuación 3 [21]:

$$J = \sum_{i=1}^{K} \sum_{\mathbf{x} \in C_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2, \tag{3}$$

donde K es el número de clusters, C_i es el conjunto de puntos en el agrupamiento i y $\boldsymbol{\mu}_k$ es el centroide del cluster i-ésimo. Para este algoritmo se han de seguir los siguientes pasos:

1. Seleccionar K centroides iniciales (aleatorios o heurísticos).
2. Cada píxel $\mathbf{x}$ se asigna al cluster C_k cuyo centroide $\boldsymbol{\mu}_k$ está más cerca siguiendo la ecuación 4:

$$C_k = \arg\min_k \|\mathbf{x} - \boldsymbol{\mu}_k\|^2. \tag{4}$$

3. Recalcular los centroides $\boldsymbol{\mu}_k$ como la media de los puntos en C_k.
4. Repetir hasta que los centroides no cambien, véase Ecuación 5:

$$\Delta\boldsymbol{\mu}_k < \text{tolerancia}. \tag{5}$$

El algoritmo K-means puede utilizarse para la segmentación de imágenes agrupando píxeles que comparten características similares, como color o intensidad, en dos clústeres: fondo y objeto de interés [23]. Dada una imagen en escala de grises $I \in \mathbb{R}^{m \times n}$, de tamaño $m \times n$, esta se modela como un conjunto de $N = m \times n$ puntos 1D, donde cada punto corresponde a una intensidad de píxel $x_p \in [0, 255]$). La binarización con $K = \alpha$ se define en la Ecuación 6:

$$\text{Cluster 1 (Fondo)}: \ \mu_1 \approx 0, \quad \text{Cluster 2 (Tumor)}: \ \mu_2 \approx 255. \tag{6}$$

Post-procesamiento: La morfología matemática es una técnica de procesamiento de imágenes fundamentada en la teoría de conjuntos y la geometría, que se utiliza para modificar la estructura de objetos en imágenes binarias o en escala de grises. Sus operaciones básicas, como la apertura y la cerradura, permiten eliminar ruido, rellenar huecos y preservar la forma de las regiones de interés [25].

En este contexto, un elemento estructurante es una pequeña matriz de píxeles con forma y tamaño definidos, utilizada para modificar la estructura de los objetos en la imagen durante las operaciones morfológicas. Define cómo se modifican los píxeles en erosión y dilatación, y es fundamental en la morfología matemática, ya que controla el tipo de modificación aplicada y, por ende, los efectos de apertura y cerradura.

En esta propuesta, la morfología se aplicó en dos etapas: apertura (erosión seguida de dilatación) y cerradura (dilatación seguida de erosión), utilizando un elemento estructurante cuadrado de 5×5 píxeles.

Primero, la erosión elimina regiones pequeñas y ruido, conservando únicamente los píxeles z donde el elemento estructurante B cabe completamente dentro de la máscara binaria A. La fórmula correspondiente se presenta en la Ecuación 7.

$$A \ominus B = \{z \in \mathbb{Z}^2 \mid B \subseteq A\}. \tag{7}$$

Por otro lado, la dilatación, por su parte, recupera áreas útiles erosionadas, expandiendo los bordes de las regiones restantes donde B_z intersecta con la máscara binaria (A). La estructura de la dilatación se describe en la Ecuación 8:

$$A \oplus B = \{z \in \mathbb{Z}^2 \mid B_z \cap A \neq \emptyset\}. \tag{8}$$

La operación de cierre (dilatación seguida de erosión) tiene como objetivo consolidar los agujeros microscópicos que puedan quedar después del proceso de apertura, logrando esto mediante la conexión de regiones cercanas y asegurando la continuidad en la máscara binaria.

El elemento estructurante empleado fue cuadrado de 5×5, lo que equilibra eficazmente la eliminación de ruido y la preservación de las estructuras. Matrices pequeñas podrían retener artefactos, mientras que elementos estructurantes mayores podrían erosionar zonas tumorales válidas o fusionar regiones no relacionadas.

Finalmente, se aplicó un filtro de área mínima, que es una técnica que elimina regiones conectadas en una imagen binaria cuyo tamaño es inferior a un umbral predefinido. Su objetivo principal es descartar el ruido residual o artefactos morfológicos no relevantes [22]. Este filtro se puede definir de manera que, dado un contorno C, su área se calcula mediante integración discreta, véase Ecuación 9:

$$C = \sum_{(x,y) \in C} 1. \tag{9}$$

El filtro de área mínima se complementa con el filtro de convexidad, que evalúa la "compactidad" de una región al comparar su área con la de su envolvente convexa mínima o casco convexo (*hull*). Los tumores cerebrales suelen mostrar alta convexidad debido a su crecimiento expansivo [26]. La convexidad se puede determinar como una fracción, cuya fórmula se presenta en la Ecuación 10:

$$\tau = \frac{\text{Área}(C)}{\text{Área}(\text{CascoConvexo}(C))}. \tag{10}$$

De este modo, se pueden definir que valores cercanos a 1 indican formas convexas, como los meningiomas, mientras que valores bajos (< 0.7) sugieren bordes irregulares, característicos de los gliomas infiltrativos. Además, se realiza un aislamiento del tumor seleccionando el clúster de mayor área, lo que garantiza que solo se retenga la masa tumoral primaria. Este proceso puede expresarse mediante la ecuación 11:

$$C_{\max} = C_j = \underset{i \in \{1,2,\ldots,k\}}{\arg\max} |C_i|. \tag{11}$$

2.3. Complejidad computacional

La complejidad computacional para el proceso de segmentación automática propuesto se determina como una función $O : \mathbb{R}^2 \to O : \mathbb{R}^2$. Por tanto, se tiene $O(I, k, B)$, donde I es la imagen que se analiza en escala de grises; k es el número de clusteres utilizados por K-means; y B es el elemento estructurante utilizado.

La complejidad del proceso paso a paso conlleva el siguiente análisis. Primero, la detección de contornos utiliza un convolución con un kernel 3×3 sobre una imagen de tamaño $m \times n$. Cada píxel se procesa en $O(1)$ y se recorren los $m \times n$ píxeles, por lo que la complejidad es $O(mn)$. Luego, al realizar la segmentación automática con K-means tanto la asignación de clústeres y actualización toma $O(kmn)$ en cada iteración. Además, si el algoritmo tarda t iteraciones en converger, la complejidad de K-means es $O(tkmn)$, donde, t y k son constantes en la práctica, se aproxima a $O(mn)$. Posteriormente, en las operaciones de convoluciones con el elemento estructurante B de tamaño $b \times b$, por lo que la complejidad es $O(b^2mn)$. Aunque tomando en cuenta que b es constante, la complejidad es $O(mn)$. Por ultimo, en el cálculo de área y convexidad, se recorren los contornos que depende del número de regiones segmentadas r. En el peor caso, la complejidad es $O(mn)$ si hay regiones pequeñas.

En resumen, la complejidad del proceso de segmentación automática es:

$$\underbrace{O(mn)}_{\text{Sobel Gradient}} + \underbrace{O(mn)}_{\text{K-means}} + \underbrace{O(mn)}_{\text{Morphology}} + \underbrace{O(mn)}_{\text{Convexity Filter}} = O(mn).$$

La segmentación automática de tumores cerebrales es un desafío debido a la variabilidad en la apariencia del tumor y la similitud de intensidades con el tejido sano. Matemáticamente, se busca particionar la imagen I en una región tumoral T y el fondo B, con $I = T \cup B$. Tradicionalmente, esto se realiza manualmente por especialistas o mediante umbralización global, donde un valor T_{umbral} separa las regiones de interés. También se emplean contornos activos para ajustar curvas a los bordes de tumor. Sin embargo, estos métodos son sensibles a la calidad de la imagen y poco eficaces en tumores con bordes difusos. En contraste, el Algoritmo 1, presenta un pseudocódigo de la segmentación automática usando K-means.

Los parámetros de entrada del algoritmo son los siguientes: I es la imagen en escala de grises, k es el número de clústeres, y B es el elemento estructurante. Como parámetro de salida se tiene M_{tumor} es la mascara de la región tumoral.

Algorithm 1 Segmentación automática de tumores cerebrales con K-means

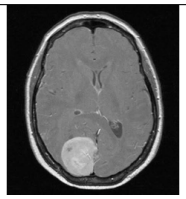
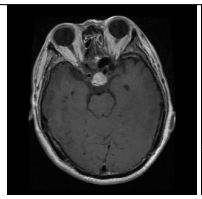
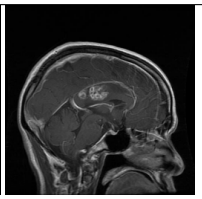
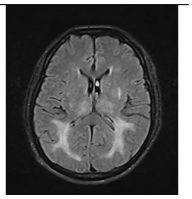
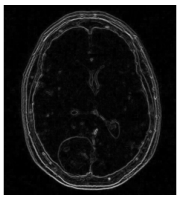
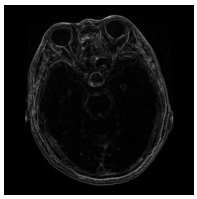
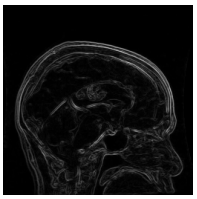
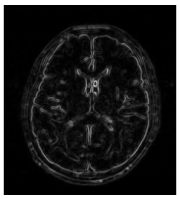
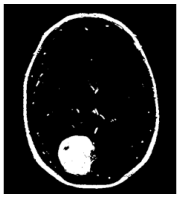
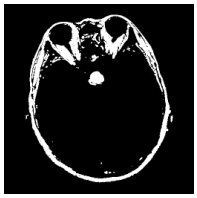
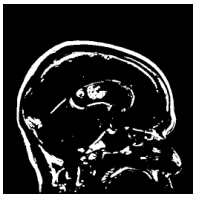
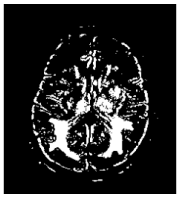
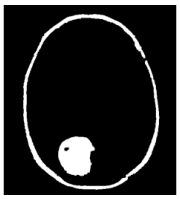
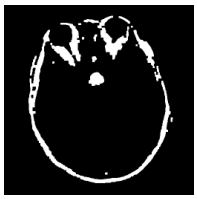
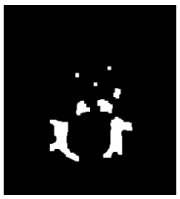
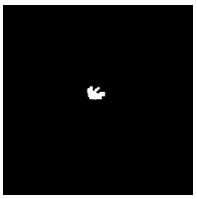
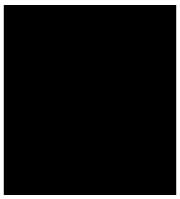
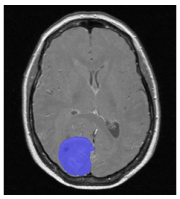
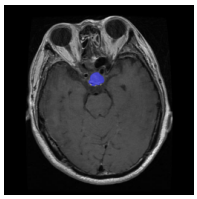
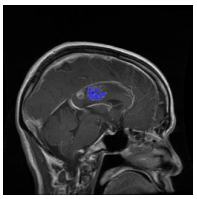
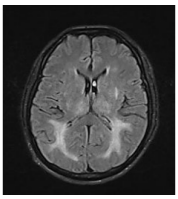
procedure SEGMENTATIONPROCESS(I,k,B)
  $G_x \leftarrow I * S_x$, $G_y \leftarrow I * S_y$ ▷ Calcular gradientes en x y y
  $G \leftarrow \sqrt{G_x^2 + G_y^2}$ ▷ Magnitud del gradiente
  $G_{norm} \leftarrow \frac{G - min(G)}{max(G) - min(G)}$ ▷ Normalizar gradiente
  $\mu_k \leftarrow \mathcal{U}(a, b)$ ▷ Inicializar k centroides
  for cada píxel $x \in I$ **do**
    $C_k \leftarrow \arg\min_k \|x - \mu_k\|^2$ ▷ Asignar x al clúster más cercano
    $\|\mu_k^{(t)} - \mu_k^{(t-1)}\| < \epsilon$ ▷ Recalcular centroides hasta convergencia
  end for
  $M_{Kmeans} \leftarrow \{x \in I \mid C(x) = C_{\text{tumor}}\}$ ▷ Generar máscara inicial
  $M_{apertura} \leftarrow (M_{Kmeans} \ominus B) \oplus B$ ▷ Mascara de apertura morfológica
  $M_{cerradura} \leftarrow (M_{apertura} \oplus B) \ominus B$ ▷ Mascara de cerradura morfológica
  $C \leftarrow \{C_1, C_2, ..., C_n\} \in M_{cerradura}$ ▷ Obtener contornos
  for cada contorno C_i **do**
    $A(C_i) \leftarrow \sum_{(x,y) \in C_i} 1$ ▷ Calcular áreas
    $\tau(C_i) \leftarrow \frac{A(C_i)}{A(CascoConvexo(C_i))}$ ▷ Calcular convexidades
  end for
  $M_{tumor} \leftarrow \{C_i \mid A(C_i) > A_{min} \wedge \tau(C_i) > \tau_{min}\}$ ▷ Obtener mascara tumoral
  return M_{tumor}
end procedure

3. Resultados y su discusión

El marco de trabajo propuesto mostró una capacidad robusta para segmentar tumores cerebrales en imágenes de resonancia magnética. A diferencia de los métodos de umbralización fija, la binarización adaptativa mediante K-means permitió ajustarse a variaciones de contraste tumoral, logrando una segmentación más precisa incluso en gliomas heterogéneos, tumores con bordes difusos y diferentes perspectivas. La combinación con operadores morfológicos y el filtrado posterior eliminó eficazmente el ruido residual, preservando la morfología tumoral y superando las limitaciones de técnicas basadas únicamente en contornos o umbrales globales, como se muestra en la Tabla 2.

Si bien el modelo ha demostrado ser efectivo, presenta algunas limitaciones. La elección fija de clústeres puede llevar a una subsegmentación en tumores complejos, como los glioblastomas con regiones necróticas, o en imágenes con baja relación señal-ruido ($SNR < 20dB$). Además, el uso de parámetros morfológicos estáticos, como un kernel de (5×5), restringe la adaptabilidad a diferentes resoluciones espaciales, lo que podría afectar su generalización a

Table 2. Etapas del marco de trabajo propuesto para la segmentación automática.

Categoría	Meningioma	Pituitaria	Glioma	Normal
Original				
Sobel Gradient				
K-means				
Morphology				
Convexity Filter				
Region of Interest				

otros protocolos de adquisición de imágenes. Como líneas de trabajo futuro, se plantean las siguientes mejoras:

1. Evaluar el proceso de segmentación mediante métricas como Dice Score e IoU, contrastándolo con el *Ground Truth*.
2. Comparar con otros algoritmos de segmentación, como U-Net o Gaussian Mixture Models, para analizar su desempeño relativo.
3. Vectorizar características de las imágenes para entrenar un modelo predictivo capaz de clasificar distintos tipos de tumores cerebrales.
4. Extender el marco de referencia a imágenes volumétricas en 3D, permitiendo una segmentación más robusta y una mejor integración con datos clínicos reales.

4. Conclusiones

En este trabajo, se desarrolló un marco de referencia para la segmentación automática de tumores cerebrales en imágenes de resonancia magnética, basado en tres pilares fundamentales: *i*) realce de bordes mediante el gradiente de Sobel, *ii*) segmentación con K-means para identificar las áreas de interés y *iii*) refinamiento de la máscara tumoral mediante filtros de convexidad y área.

Los resultados mostraron que este enfoque es capaz de segmentar con éxito las regiones tumorales, incluso en casos desafiantes (véase Tabla 2). En comparación con los métodos de umbralización fija, la capacidad adaptativa de K-means representa una mejora significativa, permitiendo manejar variaciones en la intensidad y morfología de los tumores.

Desde una perspectiva práctica, el marco de referencia propuesto facilita la automatización de la segmentación de tumores cerebrales, reduciendo la carga de trabajo de los especialistas y mejorando la reproducibilidad del diagnóstico. Su implementación clínica podría agilizar la identificación de tumores, optimizando el seguimiento de la enfermedad y la planificación quirúrgica, aunque su adopción dependerá de su integración con sistemas médicos y validación en datos reales.

En el ámbito teórico, este estudio profundiza en el uso de clustering para la segmentación médica, resaltando sus ventajas y limitaciones frente a otros modelos del estado del arte. Los resultados obtenidos sugieren la necesidad de enfoques de aprendizaje profundo e híbridos que combinen técnicas de agrupamiento con modelos más avanzados, abriendo nuevas líneas de investigación en segmentación multimodal y modelos autorregulados.

Agradecimientos. Los autores agradecen al Centro de Investigación e Innovación en Ciencias de la Computación y Tecnología Educativa (CIICCTE) de la Facultad de Informática de la UAQ por brindar el espacio para llevar a cabo este trabajo.

References

1. Louis, D.N., Perry, A., Wesseling, P.: The WHO Classification of Tumours of the Central Nervous System: A Comprehensive Update. Acta Neuropathologica, 141(1), pp. 1–36 (2021)

2. Ostrom, Q.T., Gittleman, H., Xu, J.: Epidemiology and Prognosis of Meningiomas: A Population-Based Study. Neuro-Oncology, 19(6) (2017)
3. Molitch, M.E.: Pituitary Adenomas: Epidemiology, Diagnosis, and Management. Pituitary, 20(1), pp. 1–10 (2017)
4. Kumar, A., Jain, R., Wang, M.: Limitations of Conventional MRI in Distinguishing Brain Tumor Progression from Treatment-Related Effects. Journal of Neuro-Oncology, 148(2), pp. 231–245 (2020)
5. Smith, J.A., Patel, R., Lee, C.H.: Challenges in the Radiological Diagnosis of Brain Tumors: A Systematic Review. Neurosurgery Clinics, 33(4), pp. 455–467 (2022)
6. Chen, L., Zhang, W., Zhou, H.: Inter-Rater Variability in Manual Segmentation of Pituitary Adenomas: Implications for AI-Driven Tools. Frontiers in Endocrinology, 12 (2021)
7. Ronneberger, O., Fischer, P., Brox, T.: Deep Learning for Brain Tumor Segmentation: Pitfalls and Future Directions. Medical Image Analysis, 85 (2023)
8. World Health Organization.: World Health Organization Framework on Medical Technologies for Universal Health Coverage. (2022)
9. Patel, A., Mburu, S., Al-Husseini, M.: Global Disparities in Brain Tumor Diagnostics: Challenges and Opportunities for Low-Income Countries. The Lancet Oncology, 24(4), pp. 167–175 (2023)
10. Gupta, R., Lee, S.Y., Nkengla, L.: Open-Source AI Tools for Medical Imaging: A Pathway to Equitable Healthcare Innovation. Nature Medicine, 27(9), pp. 1512–1515 (2021)
11. Martínez-Luna, C., Rodríguez-Barrera, L., Kim, S.K.: Impact of Early Diagnosis on Survival and Quality of Life in Glioblastoma Patients: A Multicenter Cohort Study. Neuro-Oncology, 25(7), pp. 1234–1245 (2023)
12. Tanaka, H., Smith, E.R., Gupta, V.: The Role of Screening MRI in High-Risk Populations for Early Detection of Asymptomatic Meningiomas. Journal of Neuro-Oncology, 158(2), pp. 217–225 (2022)
13. Alentorn, A., Dehais, C., Houillier, C.: Survival Disparities in Primary Brain Tumors: A Population-Based Analysis of Gliomas, Meningiomas, and Pituitary Adenomas. JAMA Oncology, 7(10), pp. 1501–1509 (2021)
14. Losa, M., Mortini, P., Barzaghi, L.: Long-Term Outcomes in Pituitary Adenomas: A 20-Year Follow-Up Study. The Journal of Clinical Endocrinology & Metabolism, 105(12) (2020)
15. Johnson, E.R., Tanaka, K., Müller, H.: Optimizing MRI Protocols for Brain Tumor Detection: Focus on T2-FLAIR and Diffusion Sequences. Magnetic Resonance in Medicine, 89(4), pp. 1567–1580 (2023)
16. Wang, Q., Li, C., Smith, J.K.: The Impact of MRI Acquisition Parameters on Automated Brain Tumor Segmentation Accuracy. NeuroImage, 237 (2021)
17. Lee, S.H., Park, M., Kim, Y.G. Motion Artifacts in Brain MRI: Effects on Segmentation Algorithms and Correction Strategies. In: IEEE Transactions on Medical Imaging, 41(7), pp. 1792–1803 (2022)
18. García, M.A., López, J., Torres, A.: The Role of Signal-to-Noise Ratio in MRI for Robust Tumor Segmentation. Frontiers in Neuroinformatics, 17 (2023)
19. Nickparvar, M.: Brain Tumor MRI Dataset. Kaggle, URL: https://www.kaggle.com/datasets/mnikzad/brain-tumor-mri-dataset. (2021)
20. Gonzalez, R.C., Woods, R.E.: Digital Image Processing. Pearson (2018)
21. Lloyd, S.P.: Least Squares Quantization in PCM. In: IEEE Transactions on Information Theory, 28(2), pp. 129–137 (1982) doi: 10.1109/TIT.1982.1056489.

22. Otsu, N.: A Threshold Selection Method from Gray-Level Histograms. In: IEEE Transactions on Systems, Man, and Cybernetics, 9(1), pp. 62–66 (1979) doi: 10.1109/TSMC.1979.4310076.
23. Molina E., Escobar D.J, Silva H.H: Algoritmos de Binarización Robusta de Imágenes con Iluminación No Uniforme. Revista Iberoamericana de Automática e Informática Industrial, 15(3), pp. 252–261 (2017) doi: 10.4995/riai.2017.8847.
24. Nickparvar M.: Brain Tumor MRI Dataset. Kaglee, URL: https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset/data. (2021)
25. Soille P.: Morphological Image Analysis: Principles and Applications. Springer, pp. 63–103 (2003) doi: 10.1007/978-3-662-05088-0
26. Levine M.D., Arun K.: Shape Analysis in Medical Imaging: Applications to Brain Tumors. Medical Image Analysis, 24(1), pp. 187–202 (2015)

ISSN 1870-4069

Algorithms and Approaches Used in Medical Image Segmentation for Cell Migration Tracking: A Systematic Literature Review

Mariela Judith Domínguez-Domínguez[1], Ángel J. Sánchez-García[1], María Yesenia Zavaleta-Sánchez[1], Carlos Adrián Alarcón Rojas[2]

[1] Universidad Veracruzana,
Facultad de Estadística e Informática,
Mexico

[2] Universidad Veracruzana,
Facultad de Bioanálisis,
Mexico

dominguezmariela465@gmail.com,
{angesanchez, yzavaleta, caalarcon}@uv.mx

Abstract. Medical image segmentation plays a critical role in monitoring cellular migration, especially in tumor progression studies. This systematic literature collects the findings of 32 studies that include recent advances in computational vision methodologies applied to medical image segmentation, with a particular focus on cell migration. We can observe that traditional segmentation techniques, such as region-based methods, clustering, and edge detection, have been extensively used, but recent approaches that include deep learning architectures have been more widely used like U-Net architectures and Convolutional Neural Networks, also, findings indicate that Dice Score, Hausdorff distance, Recall and F1 score are the most used evaluation metrics. Lastly the size used for image processing was heterogeneous, ranging from 128×128 pixels, as the smallest size, to 512×512 pixels.

Keywords: Segmentation, cell migration, medical images.

1 Introduction

The term ”cell migration” refers to the process by which cells move, either individually or collectively, from one place to another in response to specific stimuli. This process is essential in various biological phenomena, whether normal or pathological, such as embryonic development and tumor cell migration [23].

This and many other mechanisms can be replicated in a laboratory setting for study through cell culture and proliferation, with the wound healing assay and the use of Boyden chambers being the most commonly used techniques to

monitor the cell migration process, as well as the morphological changes that occur over time [1, 19].

Over the years, various software systems, both specific and multi-analysis, have been developed to assist in the processing and analysis of images obtained from the aforementioned techniques [30]. However, these programs present limitations and inaccuracies that must be manually corrected by the user. This entails a greater time investment, which increases as the image databases grow larger.

In recent years, the application of computer vision methodologies in different fields has become increasingly common, especially in medicine, where the interpretation and identification of pathologies based on imaging results can sometimes be inaccurate. Previous studies have shown that the use of computer vision techniques can improve the accuracy of these analyses and could potentially be used for the detection of pathologies that may be undetectable to the human eye [14, 34].

Therefore, the aim of this work is to summarize the findings of previous research related to the implementation of computer vision methodologies for the segmentation of medical images, with a focus on tracking cell migration.

2 Related Work

Recently, the implementation of segmentation techniques in medical images as a tool for disease diagnosis and monitoring has become an area of great interest. In the study conducted by Ramesh et al. in 2021 [26], a compilation of both traditional and recent methods available for medical image segmentation was carried out. They highlight region-based methods, clustering, and edge detection as the most commonly used approaches.

On the other hand, in the work presented by Gupta and Mishra in 2024 [6], a systematic review focused on deep learning-based medical image segmentation methods for polyp detection was conducted. They provided a detailed classification of the main neural network architectures applied to segmentation, including convolutional neural networks (CNNs), encoder-decoder models (such as U-Net and its variants), recurrent neural networks (RNNs), attention-based models, and generative adversarial networks (GANs).

Since there are no systematic reviews that specifically address the methods used to more accurately track tumor cell migration, the aim of this systematic literature review is to organize and present the results obtained, highlighting both the benefits and the challenges identified regarding segmentation methods in medical images applied to the study of cell migration. By gathering the available information, this review aims to provide various options that could contribute to improving current techniques and addressing key issues identified in this field of research.

Table 1. Research questions and their Motivations.

ID	Research question	Motivation
RQ1	What computer vision approaches have been used for segmentation in wound closure images at the cellular level in medical imaging?	Identify the most effective computer vision techniques applied to cellular segmentation in wound closure assays to adapt them for segmenting liver tumor cells.
RQ2	What evaluation metrics have been used to assess wound closure segmentation algorithms at the cellular level in medical imaging?	Identify the most relevant metrics for measuring the effectiveness of segmentation algorithms to evaluate the proposed approach in this work.
RQ3	What approaches or algorithms have been used to remove noise in wound closure segmentation at the cellular level in medical imaging?	Determine the most effective noise removal techniques in the context of cellular segmentation in medical imaging.
RQ4	What are the characteristics of the images used in the segmentation process of wound closure at the cellular level?	Understand the experimental conditions under which the proposed methods were evaluated to identify the most suitable approaches for our case study.

3 Research Method

In this study, the guideline proposed by [15] was followed, which is a method to identify, evaluate, and synthesize evidence in systematic literature. This approach involves three main stages such as planning the review, conducting the method, and documenting the results.

3.1 Planning Stage

In this stage, the research questions and the search terms are defined, the data sources are selected and the search string is built.

Research questions: Table 1 presents the four research questions formulated to guide this work, along with the motivation behind their formulation.

Selected search terms: The search terms used to guide this search are shown in Table 2; these terms represent key aspects of the topic of interest. The appropriate selection of these key terms allowed us to obtain the necessary information for the development of this work.

Data sources: The data sources ware selected based on their relevance in the area of artificial intelligence and image processing. Furthermore, some are multidisciplinary to cover the health sector. These data sources are shown in Table 3.

Table 2. Search terms.

Search term	Related terms
Algorithm	Approach, method, technique
Image	Medical image, Medical imaging, Image analysis, Image segmentation.
Cell	Cellular, tumor cells.
Segmentation	Segmentation evaluation, Edge detection.
Counting	Count, Cell counting, Automated cell counting, Image-based cell counting.
Wound	Wound healing, wound closure, wound healing images, wound healing assay.
Cell segmentation	Tumor segmentation.
Accuracy	Metrics, precision.

Table 3. Selected sources.

Data source	Web site
ACM Digital Library	`https://dl.acm.org/`
IEEE Xplore	`https://ieeexplore.ieee.org/`
Science Direct	`https://www.sciencedirect.com/`
Springer Link	`https://link.springer.com/`

Search string: Based on the terms presented in Table 2, several search strings were proposed and evaluated using the Quasi-Gold-Standard method proposed by [40]. This method considers metrics such as recall and precision, which are described in equations (1) and (2) respectively:

$$Recall = \frac{\text{Retrieved relevant studies}}{\text{Relevant studies}}, \tag{1}$$

$$Precision = \frac{\text{Retrieved relevant studies}}{\text{Retrieved studies}}. \tag{2}$$

The search string presented below demonstrated the best performance in retrieving relevant studies, achieving 82% in the recall metric and 19.9% in the precision metric.

```
(‘‘Medical image’’ OR ‘‘image segmentation’’ OR ‘‘wound healing’’ OR ‘‘Tumor
segmentation’’ OR ‘‘Segmentation evaluation metrics’’)) AND (‘‘Accuracy’’ OR
‘‘Precision’’)
```

Selection criteria: The selection criteria used to choose the primary studies are shown in Tables 4 and 5, where the proposed inclusion and exclusion criteria can be found.

Table 4. Inclusion criteria.

ID	Inclusion criteria
IC1	The study must have been published between 2019 and 2024
IC2	The study must be written in English
IC3	Full-text access to the article must be available
IC4	The study must answer at least one research question

Table 5. Exclusion criteria.

ID	Exclusion criteria
EC1	The study must have been published between 2019 and 2024
EC2	Duplicate studies across databases
EC3	Studies available only as slides, book chapters, posters, or technical reports

De acuerdo, vamos a aplicar el mismo estilo de la tabla Selected sources a tu tabla Inclusion criteria.

Recordemos las características de ese estilo, que es el que me indicaste que querías seguir para la tabla Selected sources y que, por ende, es el que aplicaríamos aquí:

Utiliza tabularx para el control del ancho total de la tabla.

Define las columnas usando pancho para la primera columna y X para la segunda columna. Esto permite un ancho fijo para la primera y una distribución justificada para la segunda.

Incluye líneas verticales y horizontales.

Usa ara ajustar el espaciado entre filas.

Aquí tienes el código para tu tabla Inclusion criteria con ese estilo:

Fragmento de código

3.2 Conduction Stage

The selection process was divided into three stages, during which the inclusion and exclusion criteria were applied to identify the primary studies. In the first stage, IC1 and IC2 were applied, while in stage 2, EC1 and EC2 were applied. In the third and final stage, IC3 and IC4 were applied. This process can be observed in Figure 1.

The results of the primary study selection process described in Figure 1 are outlined in Table 6, where it can be observed that after applying stage 3, a total of 32 primary studies were obtained.

3.3 Reporting Stage

A narrative synthesis was carried out following the steps presented by [24]. In this synthesis, an analysis of the findings was performed, identifying patterns in plots and tabulations.

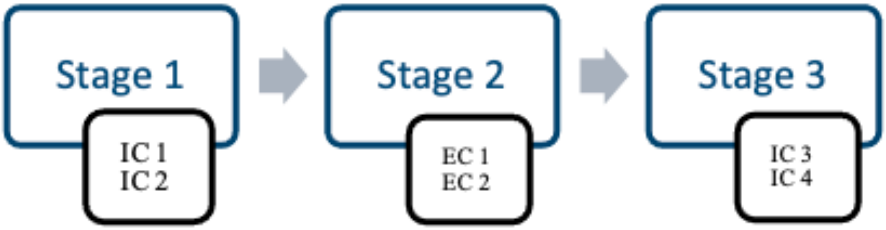


Fig. 1. Primary studies selection process.

Table 6. Results of selection process by stage.

Source	Initial stage	Stage 1	Stage 2	Stage 3
Springer Link	307	139	122	7
ScienceDirect	152	47	20	2
IEEE Xplore	18	18	18	10
ACM DL	377	352	122	13
Total	854	556	282	32

4 Results

The automated search was conducted in the data sources shown in Table 3, with ACM DL accounting for 41% of the relevant studies as shown in Fig. 2, followed by IEEE Xplore with 31%. Table 7 lists the selected studies and their corresponding data sources.

Additionally, the distribution of articles by publication type was identified, as shown in Fig. 3. Of the total articles, 69% were published in journals and 31% in conferences.

Finally, regarding the distribution of publications by year, as can be seen in Fig. 4, 21 of the 32 selected studies were published in 2024, while 8 were published in 2023 and 2022, showing research topic is current and relevant.

4.1 RQ1. What computer vision approaches have been used for segmentation in wound closure images at the cellular level in medical images?

During this work, two main computer vision approaches for segmentation were identified: texture and color. The studies that addressed this research question and used color as an approach were mainly based on genetic algorithms [19, 30], while those that used texture primarily implemented deep learning techniques [11, 16]. In both cases, their study objects were different types of tumors, and they used images obtained through imaging techniques such as magnetic resonance imaging and computed tomography, as well as images obtained from cell migration assays, including the wound closure assay.

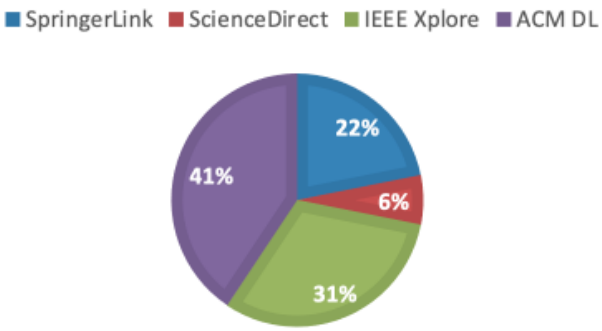


Fig. 2. Distribution of primary studies by data source.

Table 7. Selected Primary Studies

Data source	Primary studies
Springer Link	[10] [16] [11] [18] [17] [37] [8]
ScienceDirect	[14] [4]
IEEE Xplore	[25] [29] [22] [13] [20] [28] [38] [12] [27] [32]
ACM DL	[21] [36] [9] [3] [39] [31] [2] [5] [35] [7] [41] [42] [33]

4.2 RQ2. What evaluation metrics have been used to assess wound closure segmentation algorithms at the cellular level in medical images?

The most used metrics by the authors to evaluate the performance of segmentation algorithms were average precision, F1-score, and Dice score, reported in 10 of the 32 selected articles, with values ranging between 80% and 90% for each metric across all studies. This can be seen in more detail in Table 8.

4.3 RQ3. What approach or algorithm is used to remove noise in wound closure segmentation at the cellular level in medical images?

A wide variety of algorithms and techniques were found for noise removal and analysis of medical images, but two of them stood out: U-Net, which is used in 14 of the 32 analyzed articles, and Convolutional Neural Networks (CNNs), used in five of the 32 analyzed articles. This can be seen in more detail in Table 9.

4.4 RQ4. What are the characteristics of the images used in the segmentation pro-cess of wound closure at the cellular level?

We observed that the image sizes used in the analyzed studies were heterogeneous, ranging from 128x128 pixels, as the smallest size, to 512x512 pixels, as the largest size. Both grayscale and RGB images were used.

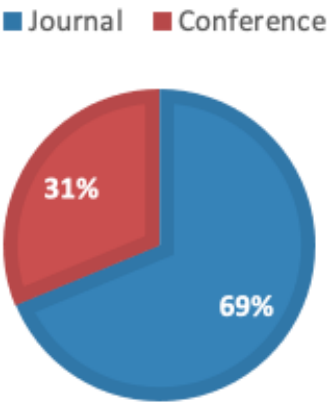


Fig. 3. Distribution of primary studies by publication type.

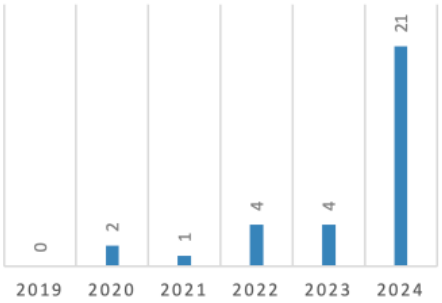


Fig. 4. Distribution of primary studies by publication year.

Table 8. Distribution of Primary Studies by evaluation metrics used.

Evaluation metrics	Primary studies
Intersection over Union (IoU)	[11]
Dice Score	[18] [29] [22] [38]
Hausdorff distance	[18] [17] [22]
Structural Similarity Index (SSIM)	[4]
Mean Squared Error (MSE)	[4]
Percentage of Misclassification (PM)	[29]
Peak signal-to-noise ratio (PSNR)	[17] [29]
Recall	[11] [17] [38]
Precision	[11] [27]
F1 score	[17] [38] [21]
Accuracy	[27] [21]

Table 9. Distribution of Primary Studies by Computer vision approach.

Approach	Primary studies
Spatial-channel Convolution Optimization (ASCO)	[11]
U-Net	[18] [37] [29] [22] [13] [28] [38] [12] [21] [36] [31] [35] [7] [41]
Deep Attention Integrated Networks (DAINets)	[33]
Encoder-decoder network for segmentation and a sub network for classification	[41] [42]
Arithmetic Optimization Algorithm (AOA)	[17]
Ultrasound Network (US-Net)	[3]
Contextual Attention Network (CAN)	[32]
Convolutional Neural Networks (CNN)	[14] [8] [27] [39] [5]
CA-Unet network	[25]
Prediction Wound Progression Framework (PWPF)	[4]

5 Conclusions and Future Work

This systematic review highlights the growing reliance on U-Net and CNN-based architectures for medical image segmentation, particularly in tracking cellular migration. These models have shown superior segmentation accuracy compared to traditional methods. However, several challenges remain, including the need for extensive labeled datasets, and the difficulty of generalizing models to different imaging modalities.

Future research should prioritize focus on improving datasets availability and synthetic data generation. Another crucial approach would be to refining existing models to achieve better generalization across diverse imaging conditions and reduce the need for manual corrections. Addressing these challenges will contribute to the development of more reliable and efficient segmentation tools for biomedical applications.

References

1. Brown, K. C., Sugrue, A. M., Modi, K. J.: An Experimental Protocol for the Boyden Chamber Invasion Assay With Absorbance Readout. Bio-protocol, vol. 14, no. 15, pp. 1–21 (2024) doi: 10.21769/BioProtoc.5040.
2. Chen, Q.-Q., Sun, Z.-H., Wei, C.-F.: Semi-Supervised 3D Medical Image Segmentation Based on Dual-Task Consistent Joint Learning and Task-Level Regularization. In: IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 20, no. 4, pp. 2457–2467 (2022) doi: 10.1109/TCBB.2022.3144428.
3. Erragzi, N., Zrira, N., Jimi, A.: US-Net: A Breast Ultrasound Image Segmentation using Deep Learning. In: Proceedings of the International Conference on Advances in Social Networks Analysis and Mining. pp. 596–602 (2023) doi: 10.1145/3625007.36273.
4. Garcia-Moreno, F. M., Ruiz-Espigares, J., Gutiérrez-Naranjo, M. A.: Using Deep Learning for Predicting the Dynamic Evolution of Breast Cancer Migration. Computers in Biology and Medicine, vol. 180, pp. 1–18 (2024) doi: 10.1016/j.compbiomed.2024.108890.
5. Guo, K., Wu, J., Wan, W.: Biomedical Image Segmentation Based on Classification Supervision. In: Proceedings of the 13th International Conference on Bioinformatics and Biomedical Technology. pp. 22–27 (2021) doi: 10.1145/3473258.34732.
6. Gupta, M., Mishra, A.: A Systematic Review of Deep Learning Based Image Segmentation to Detect Polyp. Artificial Intelligence Review, vol. 57, no. 1, pp. 1–53 (2024) doi: 10.1007/s10462-023-10621-1.
7. Haja, A., Radu, S., Schomaker, L.: A Comparison of Different U-Net Models for Segmentation of Overlapping Organoids. In: Proceedings of the 9th International Conference on Biomedical and Bioinformatics Engineering. pp. 1–10 (2022) doi: 10.1145/3574198.357419.
8. He, L., Li, M., Wang, X.: Morphology-Based Deep Learning Enables Accurate Detection of Senescence in Mesenchymal Stem Cell Cultures. BMC Biology, vol. 22, no. 1, pp. 1–17 (2024) doi: 10.1186/s12915-023-01780-2.
9. He, L., Zhang, Z., Zhang, J.: Context-Based Deep Residual Learning for Medical Image Segmentation. In: Proceedings of the 9th International Conference on Communication and Information Processing. pp. 206–212 (2023) doi: 10.1145/3638884.363891.
10. Huang, T., Yin, H., Huang, X.: Improved Genetic Algorithm for Multi-Threshold Optimization in Digital Pathology Image Segmentation. Scientific Reports, vol. 14, no. 1, pp. 1–21 (2024) doi: 10.1038/s41598-024-73335-6.
11. Ji, Z., Mu, J., Liu, J.: ASD-Net: A Novel U-Net Based Asymmetric Spatial-Channel Convolution Network for Precise Kidney and Kidney Tumor Image Segmentation.

Medical & Biological Engineering & Computing, vol. 62, no. 6, pp. 1673–1687 (2024) doi: 10.1007/s11517-024-03025-y.
12. Ji, Z., Zhao, Z., Zeng, X.: ResDSda_U-Net: A Novel U-Net-Based Residual Network for Segmentation of Pulmonary Nodules in Lung CT Images. In: IEEE Access, vol. 11, pp. 87775–87789 (2023) doi: 10.1109/ACCESS.2023.3305270.
13. Katiyar, P. S., Sarmah, R.: VU-NET: An Explainable AI Approach For Liver Segmentation. In: 15th International Conference on Computing Communication and Networking Technologies (ICCCNT). pp. 1–7. IEEE (2024) doi: 10.1109/ICCCNT61001.2024.10725563.
14. Kavitha, C., Priyanka, S., Kumar, M. P.: Skin Cancer Detection and Classification using Deep Learning Techniques. Procedia Computer Science, vol. 235, pp. 2793–2802 (2024) doi: 10.1016/j.procs.2024.04.264.
15. Kitchenham, B. A., Budgen, D., Brereton, P.: Evidence-Based Software Engineering and Systematic Reviews. Chapman and Hall/CRC (2015)
16. Kutlu, F., Ayaz, İ., Garg, H.: Integrating Fuzzy Metrics and Negation Operator in FCM Algorithm Via Genetic Algorithm for MRI Image Segmentation. Neural Computing and Applications, vol. 36, no. 27, pp. 17057–17077 (2024) doi: 10.1007/s00521-024-09994-3.
17. Li, H., Zhu, X., Li, M.: Multi-Threshold Image Segmentation Research Based on Improved Enhanced Arithmetic Optimization Algorithm. Signal, Image and Video Processing, vol. 18, no. 5, pp. 4045–4058 (2024) doi: 10.1007/s11760-024-03026-2.
18. Li, H., Nan, Y., Del Ser, J.: Large-Kernel Attention for 3D Medical Image Segmentation. Cognitive Computation, vol. 16, no. 4, pp. 2063–2077 (2024) doi: 10.48550/arXiv.2207.11225.
19. Martinotti, S., Ranzato, E.: Scratch Wound Healing Assay. Epidermal Cells: Methods and Protocols, pp. 225–229 (2020) doi: 10.1007/7651_2019_259
20. Mourad, A., Afifi, A., Keshk, A. E.: Automated Brain Tumor Segmentation in MRI using Superpixel Over-segmentation and Classification. In: 21st International Arab Conference on Information Technology (ACIT). pp. 1–8. IEEE (2020) doi: 10.1109/ACIT50332.2020.9300122.
21. Musthafa, N., Masud, M. M., Memon, Q.: Advancing Early-Stage Brain Tumor Detection with Segmentation by Modified_Unet. In: Proceedings of the 8th International Conference on Medical and Health Informatics. pp. 52–57 (2024) doi: 10.1145/3673971.3674001.
22. Osei, I., Appiah-Kubi, B., Frimpong, B. K.: Multimodal Brain Tumor Segmentation Using Transformer and UNET. In: 20th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP). pp. 1–6. IEEE (2023) doi: 10.1109/ICCWAMTIP60502.2023.10387113.
23. Pijuan, J., Barceló, C., Moreno, D. F.: In Vitro Cell Migration, Invasion, and Adhesion Assays: From Cell Imaging to Data Analysis. Frontiers in Cell and Developmental Biology, vol. 7, pp. 1–16 (2019) doi: 10.3389/fcell.2019.00107.
24. Popay, J., Roberts, H., Sowden, A.: Guidance on the Conduct of Narrative Synthesis in Systematic Reviews. A Product from the ESRC Methods Programme Version, vol. 1, no. 1 (2006)
25. Pu, L., Wan, L., Wang, X.: A Collaborative Attention Mechanism Unet for Liver Tumor CT Image Segmentation Algorithm. In: International Conference on Algorithms, Data Mining, and Information Technology (ADMIT). pp. 7–13. IEEE (2022) doi: 10.1109/ADMIT57209.2022.00010.

26. Ramesh, K., Kumar, G. K., Swapna, K.: A Review of Medical Image Segmentation Algorithms. EAI Endorsed Transactions on Pervasive Health & Technology, vol. 7, no. 27, pp. 1–9 (2021) doi: 10.4108/eai.12-4-2021.169184.
27. Rastogi, D., Sharma, A., Yadav, R.: Anomaly Detection in Medical Images Using Deep Reinforcement Learning. In: 2nd International Conference on Disruptive Technologies (ICDT). pp. 506–512. IEEE (2024)
28. Rathore, S., Sahare, P.: Design of an Efficient Model for Enhanced Liver and Tumor Segmentation Using Advanced Deep Learning Techniques. In: 4th International Conference on Intelligent Technologies (CONIT). pp. 1–6. IEEE (2024) doi: 10.1109/CONIT61985.2024.10626313.
29. Samantaray, R., Wagh, M. P., Prasad, R.: Enhanced Brain Tumor Segmentation Using Improved U-Net Architecture. In: 15th International Conference on Computing Communication and Networking Technologies (ICCCNT). pp. 1–6. IEEE (2024) doi: 10.1109/ICCCNT61001.2024.10724811.
30. Smith, K., Piccinini, F., Balassa, T.: Phenotypic Image Analysis Software Tools for Exploring and Understanding Big Image Data from Cell-Based Assays. Cell Systems, vol. 6, no. 6, pp. 636–653 (2018) doi: 10.1016/j.cels.2018.06.001.
31. Sonia, M., Kalita, I., Devi, D.: A Breast Cancer Prognosis Model using PyRadiomics and Image Segmentation from MRI Data. In: Proceedings of the 7th International Conference on Machine Vision and Applications. pp. 27–34 (2024) doi: 10.1145/3653946.365395.
32. Srinivasan, P. S., Regan, M.: Enhancing Brain Tumor Diagnosis with Substructure Aware Graph Neural Networks and Fuzzy Linguistic Segmentation. In: Second International Conference on Intelligent Cyber Physical Systems and Internet of Things (ICoICI). pp. 1613–1618. IEEE (2024) doi: 10.1109/ICoICI62503.2024.10696691.
33. Sun, M., Zou, W., Wang, Z.: An Automated Framework for Histopathological Nucleus Segmentation With Deep Attention Integrated Networks. IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 21, no. 4, pp. 995–1006 (2023) doi: 10.1109/TCBB.2022.3233400.
34. Teixeira, P. A., Sousa, P. A., Coimbra, M.: Computer Vision Challenges for Chronic Wounds Assessment. In: 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). pp. 1840–1843. IEEE (2020) doi: 10.1109/EMBC44109.2020.9175713.
35. Wang, Z., Chen, Y., Yin, J.: Application of Improved U-Net with Feature Fusion and Expectation-Maximization Attention in Kidney Tumor Segmentation of CT Images. In: Proceedings of the 4th International Conference on Bioinformatics and Intelligent Computing. pp. 113–118 (2024) doi: 10.1145/3665689.366570.
36. Wu, C., Song, C., Cheng, D.: IDMUNet: An Effective Network for Liver Tumor Segmentation. In: Proceedings of the 7th International Conference on Advances in Image Processing. pp. 49–56 (2023) doi: 10.1145/3635118.363512.
37. Xu, R., Xu, C., Li, Z.: Boundary Guidance Network for Medical Image Segmentation. Scientific Reports, vol. 14, no. 1, pp. 1–14 (2024) doi: 10.1038/s41598-024-67554-0.
38. Yadav, A. C., Alam, Z., Mufeed, M.: U-Net-Driven Advancements in Breast Cancer Detection and Segmentation. In: International Conference on Electrical Electronics and Computing Technologies (ICEECT). vol. 1, pp. 1–6. IEEE (2024) doi: 10.1109/ICEECT61758.2024.10738914.
39. Yang, B., Cao, X., Wang, H.: DCTNet: A Fusion of Transformer and CNN for Advanced Multimodal Medical Image Segmentation. In: Proceedings of the 5th

International Conference on Computer Information and Big Data Applications. pp. 762–767 (2024) doi: 10.1145/3671151.36712.

40. Zhang, H., Babar, M. A., Tell, P.: Identifying Relevant Studies in Software Engineering. Information and Software Technology, vol. 53, no. 6, pp. 625–637 (2011) doi: 10.1016/j.infsof.2010.12.010
41. Zhang, R., Zhang, R., Ma, J.: Analysis of Different Encoder-Decoder-Based Approaches for Biomedical Imaging Segmentation. In: Proceedings of the 6th International Conference on Robotics and Artificial Intelligence. pp. 105–113 (2020) doi: 10.1145/3449301.34493.
42. Zhang, X., Han, J., Li, Z.: A Multi-task Learning framework for Segmentation and Classification of Patellofemoral Osteoarthritis in Multi-Parametric Magnetic Resonance Imaging. In: Proceedings of the 5th International Conference on Artificial Intelligence and Pattern Recognition. pp. 449–456 (2022) doi: 10.1145/3573942.357404.

ISSN 1870-4069

Enhancing Embryo Image Interpretability through Language Models

Alberto León[1], Isaac Aguilar[2], Omar Paredes[1,2]

[1] Biodigital Innovation Lab,
Translational Bioengineering Department, CUCEI,
Mexico

[2] R&D Department, IVF 2.0 LTD,
United States of America

jose.aguirre5645@alumnos.udg.mx,isaac@ivf20.com,
omar.paredes@academicos.udg.mx

Abstract. Large Language Models (LLMs) have rapidly transformed numerous disciplines following the release of ChatGPT. The emergence of open-source models like LlaMA, Mistral, Phi and R1 has accelerated the Language Models adoption across diverse computational domains. Building on this foundation, Vision-Language Models (VLMs) and Multimodal Models (MMs) have extended these capabilities to process and interpret visual data alongside text. This study explores the application of Small Language Models (SMLs) in embryo evaluation, a critical area in assisted reproductive procedures. We demonstrate that SMLs can effectively interpret outputs from specialized deep learning systems, translating complex embryo features into accessible natural language descriptions for clinicians. Through comparative analysis of four models (DeepSeek, Llama 3.2 Vision, Llava:7b and Qwen2), we identify DeepSeek as most effective in generating descriptions that balance detail and conciseness. Our approach addresses two key challenges in medical AI: bridging the "black box" gap between complex deep learning and human-readable explanations and providing computationally efficient solutions suitable for clinical settings where data privacy is paramount. By enabling interpretability of specialized AI systems through lightweight language models, our methodology offers a promising direction for enhancing embryologist decision-making in *In Vitro Fertilization* (IVF) procedures while maintaining practical deployment capabilities in resource-constrained environments.

Keywords: Language models, small language models, interpretability, embryo.

1 Introduction

Language Models (LMs) have revolutionized numerous disciplines in recent years, particularly following the widespread adoption of ChatGPT as a breakthrough technology. Similar applications using language models to generate feedback through chat, text, and even images have quickly followed. Subsequently, open-source models including Llama, Mistral, Phi, and R1 have emerged, leading to the development of

language models as solutions for virtually any computational task which offer similar capabilities with lower computational requirements. These models have practical advantages in clinical settings, where efficiency and privacy are critical being the last one something important in clinical field to keep privacy of patient record and not sending this sensitive information to any other vendor, service provider or company.

The natural evolution of these models has been the emergence of Vision and Language Models (VLMs), which are LMs capable of "seeing" images, extracting their content, and generating responses to queries related to visual content. Recent work has explored their application in ophthalmology, radiology, microscopy, and embryology [1,2,5,6]. However, literature lacks focused studies on their use in reproductive medicine. Our study builds on these advances, presenting SLMs as interpretable layers for deep learning systems that process embryo images. Unlike LLMs, SLMs are deployable in constrained environments, enabling privacy-preserving workflows in sensitive medical domains.

Our study builds on this trend by proposing the use of these models to enhance interpretability of AI systems used in specialized medical image evaluation, enabling multimodality capability previously limited to private models such as ChatGPT [OpenAI] and Claude [Anthropic]. This limitation has presented obstacles for medical applications where patient data privacy is paramount.

Deep learning systems have long been characterized as "black boxes" [3], making it difficult or impossible to understand how they process information and reach conclusions. The interpretability challenge varies depending on the neural network architecture and outputs generated. As Afnan et al. assert [4], "the sacrifice of interpretability is something that cannot be justified". Language models, with their ability to abstract information and analyze it within broader contexts (dependent on their training data), may clarify complex values and provide better interpretability for sophisticated deep learning models, particularly for non-specialists. Recent work demonstrates this potential, using vision models and ChatGPT as tools for generating medical image diagnoses [4,5].

In this paper, we present a proof-of-concept protocol using SLMs as interpreters of results from other deep learning models whose outputs are abstract to human understanding. SLMs are particularly suitable for this application as they are computationally efficient and have narrower context windows, potentially reducing the risk of output deviation. We apply this approach to assisted reproduction, specifically focusing on the critical task of embryo evaluation for transfer and feature description. This application demonstrates how SLMs can bridge the gap between complex AI outputs and human-understandable information in sensitive medical contexts.

2 Methodology

2.1 Small Language Model Selection

In this study, we evaluated several state-of-the-art Small Language Models (SMLs) with multimodal capabilities, selected for their balance of performance and efficiency. These models represent different approaches to combining language and vision capabilities while maintaining reasonable computational requirements as show in Table

Table 1. SMLs used, their parameters length and publisher.

Model	Parameters	Publisher
Llama 3.2 Vision	11B	Meta
Llava	7B	Haotuan Liu
MiniCPM-V	8B	OpenBMB
DeepSeek	14B	DeepSeek

1. Each model offers distinct advantages in processing medical imagery and interpreting complex features.

Llama 3.2 Is a foundational language model developed by Meta, distinguished by its dense Transformer architecture. It demonstrates multilingual capabilities and strong logical reasoning, showing performance comparable to GPT-4 in various evaluation tasks. The model was trained on 15.6 billion multilingual tokens, with particular emphasis on data quality and diversity [6]. For this work, we utilized the Vision variant with 11 billion parameters.

Llava Large Language and Vision Assistant (Llava) is a multimodal model developed using the "Visual Instruction Tuning" strategy. It combines a visual encoder based on CLIP with a Llama language model and was trained on data automatically generated by GPT-4. Llava excels at following detailed visual instructions and executing complex reasoning tasks at a level comparable to GPT-4 [7]. For this research, we employed the 7 billion parameter variant.

MiniCPM-V belongs to a family of lightweight multimodal models developed by the OpenBMB team, specifically designed for efficient operation in constrained systems such as mobile or embedded devices. The most recent model (version 2.6) integrates adaptive vision techniques with the Qwen2 language model. It distinguishes itself through its ability to process high-resolution images (up to 1.8 million pixels) and, when combined with quantization processes, offers exceptional performance and efficiency [8]. For our work, we utilized the 8 billion parameter version.

R1 is an advanced language model developed by DeepSeek-AI, specifically designed to enhance reasoning abilities through reinforcement learning. The final version combines initial cold-start data, supervised fine-tuning, and reasoning-oriented reinforcement learning, enabling it to achieve results comparable to leading models such as OpenAI-o1-1217 in cognitive tasks and complex questions. The model offers distilled variants based on architecture like Qwen and Llama, making it more lightweight [9]. For this research, we employed the 14 billion parameter version.

2.2 Experimental Design

Research Collaboration and Data Acquisition This research was conducted in collaboration with IVF 2.0 LTD, which provided anonymized embryo images and access to their proprietary ERICA API [10] for image processing and feature extraction. A total of 30 anonymized embryo images were supplied for testing across the different SMLs.

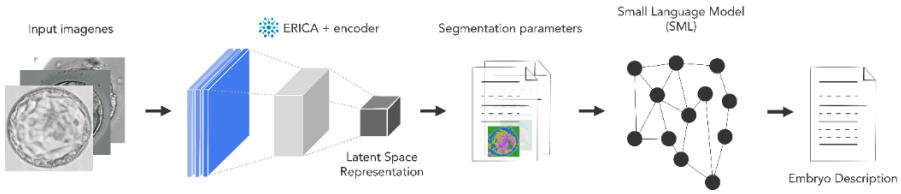


Fig. 1. Pipeline to extract features and parameters for the SML to generate a description of an embryo image.

Image Preprocessing and Feature Extraction The raw images provided by IVF 2.0 were processed by our team to ensure consistency. We cropped each embryo image to standardize dimensions across the dataset, preparing them for subsequent feature extraction.

The standardized images were processed through the ERICA API [10], which extracted morphological zone features in tensor format and generated segmentation masks. These features represented key embryological characteristics used by specialists in embryo evaluation, such as Pellucid Zone, Trophectoderm and Inner Cell Mass segmentation briefly described on Figure 1.

We designed two primary testing protocols:

Feature Interpretation Test: SMLs were prompted with comprehensive embryo features and segmentation masks to interpret the image.

Quality Assessment Test: SMLs evaluated embryo quality based on the parameters extracted from images and characteristics provided by ERICA.

2.3 Implementation Framework

Technical Integration and Experimental Process We accessed the language models through the Ollama platform [11], which provides open-source language models via a straightforward Python API. Model versions from the Ollama repository.

Using the Ollama API, we generated a fresh chat instance for each model-image combination to prevent cross-contamination between descriptions. For each image, we created a prompt containing both the image and its extracted characteristics as shown in Figure 2. The resulting descriptions were saved in a data frame for subsequent analysis, as outlined in the next section.

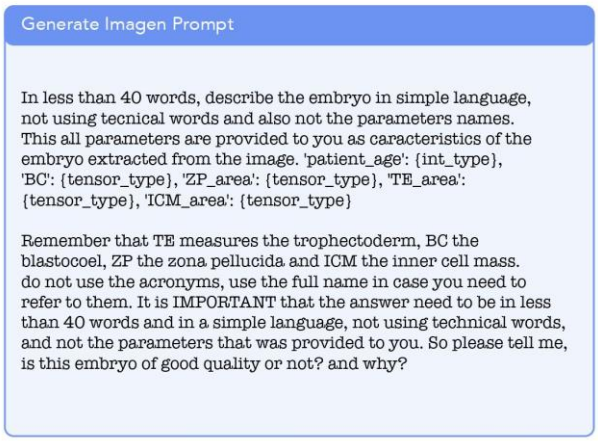


Fig. 2. Prompt used to generate the descriptions of each image for all SMLs.

Technical Requirements and Reproducibility Guidelines Our testing environment consisted of a computer equipped with a Nvidia RTX 4080 (16GB VRAM) graphics card, Intel 14700K processor, and 64GB DDR5 RAM. The most computationally intensive component was the Ollama Chat API interacting with the language models, which primarily utilized the GPU.

To replicate similar experiments, the following are necessary: the Ollama package, Python for scripting, and appropriate models selected from ollama.com. Researchers may also employ vision model encoders or implement alternative feature encoders such as OpenAI's CLIP or convolutional network models. Hardware specifications may vary as the primary computational demand is on the GPU for model inference.

3 Results

We analyzed participants' preferences regarding descriptions generated by four different language models: DeepSeek, Llama 3.2 Vision, Llava:7b, and Qwen2. Each model provided a description for 27 embryo images, resulting in a total of 108 descriptions. Participants selected the description they considered most appropriate or appealing for each image.

Our analysis revealed that DeepSeek was the most preferred model, receiving 106 total selections across all participants. Qwen2 followed with 83 selections, while Llava:7b and Llama 3.2 Vision received 59 and 49 selections respectively see in Figure 3. This distribution suggests a clear preference hierarchy among the models when interpreting morphological features.

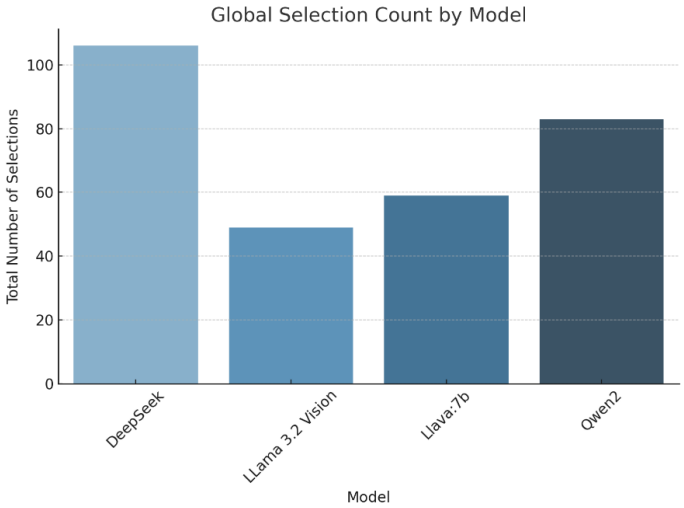


Fig. 3. Global selection counts of descriptions generated by each model.

Table 2. Selection for each model and average number of words on the model descriptions.

Model	Selection count	Avg # words
DeepSeek R1	106	45.35
MiniCPM-V	83	56.25
Llava	59	75.32
Llama 3.2 Vision	49	34.46

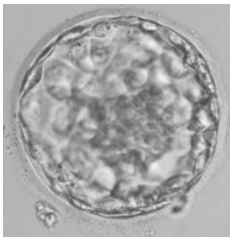
As clarification for the number of images from 30 to 27, these 3 remaining images were corrupted and unable to process so they were discarded from the experiment.

To investigate whether description length influenced participants' preferences, we analyzed the average word count of descriptions generated by each model see in Table 2. Llava:7b produced the longest descriptions with an average of approximately 75 words per description, followed by Qwen2 with about 56 words, DeepSeek with about 45 words, and Llama 3.2 Vision with the shortest descriptions averaging only 34 words.

This analysis reveals that longer descriptions did not necessarily correlate with higher preference rates. Despite generating substantially longer descriptions, Llava:7b ranked third in overall participant preferences. In contrast, DeepSeek, with its more concise descriptions of moderate length, was significantly favored by participants. This suggests that moderate description length may optimize clarity and general acceptance among users, potentially striking an ideal balance between providing sufficient detail while remaining accessible.

Interestingly, Llama 3.2 Vision, which produced the shortest descriptions, received the fewest selections. This indicates that excessively brief descriptions may fail to provide enough detail to satisfy participants' information needs when interpreting complex embryological features. The relationship between description length and preference appears to follow an inverted U-shaped curve, with optimal length falling in the moderate range represented by DeepSeek's descriptions.

Table 3. Descriptions generated by each model of embryo images 1, 5 & 11.

Images	Llama3.2 vision	Llava 7B	MiniCPM-v (Qwen2)	DeepSeek R1
	The embryo appears to have **poor quality**. While it is euploid (normal chromosome count), the measurements show unusual proportions between key structures like the inner cell mass and trophectoderm, and some areas are too small or misshapen, which can affect its developmental potential.	Based on the image's characteristics, it is difficult to determine if the embryo is of good quality as there are several factors affecting its quality. The size of the embryo, which includes the trophectoderm (TE), inner cell mass (ICM), and zona pellucida (ZP), may not be enough to assess its quality. Furthermore, other factors such as the number of cells, the presence of fragmentation or abnormalities in the embryo's shape, and the overall appearance cannot be assessed without further information.	The embryo appears to have various irregularities such as decreased circularity, increased eccentricity, and low area-to-perimeter ratio. These factors suggest it may be of lower quality compared to a well-formed embryo with high circularity and proper proportions between size metrics. The presence of an atypical ICM also indicates potential issues in the embryo's development.	The embryo appears to have **poor quality**. While it is euploid (normal chromosome count), the measurements show unusual proportions between key structures like the inner cell mass and trophectoderm, and some areas are too small or misshapen, which can affect its developmental potential.
	This embryo appears to be of good quality. The zona pellucida is thick enough, the inner cell mass is well-formed, and the trophoblast layer is even. However, it's always best to consult a doctor for an accurate assessment.	This embryo has a large and well-defined inner cell mass (ICM) and trophectoderm (TE), suggesting it may be of high quality. The large size of the blastocoel (BC) is typical for an early-stage embryo, while the zona pellucida (ZP) appears to be thinning out. Overall, these characteristics suggest that this embryo is of good quality and likely has a good chance of implanting successfully.	This embryo appears euploid with healthy cell division. The blastocoel has no significant defects, zona pellucida looks intact but slightly elongated in shape, the inner cell mass seems well developed without any signs of abnormality, and the trophectoderm is present too. Based on these features, it can be considered as a good quality embryo.	This embryo appears to be of good quality because it is euploid (the correct number of chromosomes), has normal-sized areas for its structures, and shows proper development of the inner cell mass, which will form the future baby.
	The embryo appears to be of good quality. It has a normal number of chromosomes (euploid) and its cells are well-organized. The measurements suggest a healthy balance between the outer layers and the inner cell mass.	Based on the image, the embryo appears to be of poor quality. The reason for this assessment is the low values for BC_circularity (0.257) and ICM_circularity (0.634), which indicate that the embryo has an irregular shape. Additionally, the TE_ICM_area_ratio (0.844) suggests a lack of proper development in the inner cell mass. It is important to note that these assessments are based solely on the image provided and do not consider other important factors such as chromosomal status or the overall health of the embryo.	The embryo appears euploid with a well-defined structure. However, it has low circularity values for BC, ZP, ICM, and TE, indicating potential issues in cell shape and organization. Overall, the quality seems to be average but could improve.	This embryo appears to have irregular shapes for its blastocoel, trophectoderm, and inner cell mass, with very low circularity in all areas. The zona pellucida also has a highly irregular shape. These features suggest the embryo may not be of optimal quality due to fragmentation or poor organization.

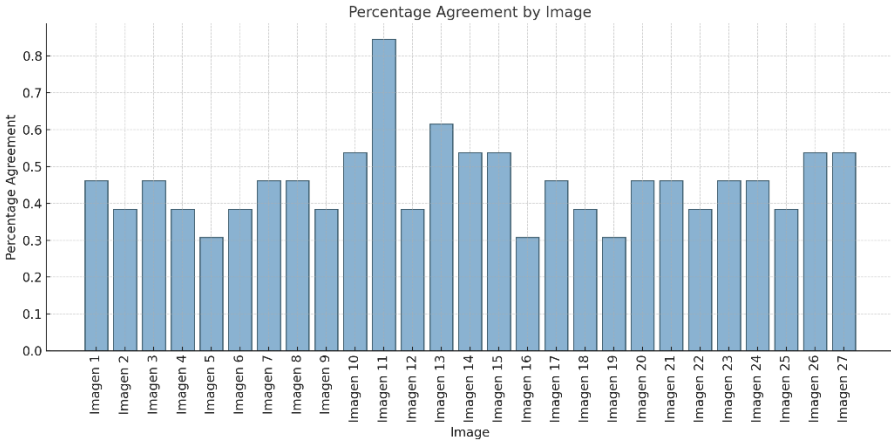


Fig. 4. Percentage of Agreement by image.

The preference patterns observed suggest that when interpreting specialized medical imagery, particularly in the context of embryo evaluation, users value descriptions that balance comprehensiveness with conciseness. This finding has important implications for the design of AI-assisted medical interpretation systems, suggesting that optimizing for moderate description length may enhance user satisfaction and potentially improve the clinical utility of such systems.

Participants evaluators were from a diverse but short number of professionals, such as embryologists, biologists and technicians in the field of medicine (medical engineers) in total 11 participants the method to get the selections and feedback was through a form (feedback was no considered in this paper because tend to be subjective and limited to the words of the description y future experiments we will try get a description to compare model vs professional), so they can see image + description by each model as see in Table 3.

4 Discussion

4.1 Variability in Model Agreement

The analysis of percentage agreement for descriptions generated by language models for embryo images revealed significant variability across the dataset Figure 4. Image 11 demonstrated the highest consensus among language models, exceeding an 80% agreement level. This indicates a strong convergence towards a particular standardized description, which was selected as the gold standard due to its precision and clarity.

The gold-standard description for image 11 Table 3, provides a clear, detailed characterization of structural irregularities in the blastocoel, trophectoderm, inner cell mass, and zona pellucida, emphasizing critical quality indicators such as fragmentation and low circularity. This precision likely facilitated high agreement among language models, as clear pathological indicators were easily identifiable, leading to a unified interpretation.

Conversely, image 5 Table 3 exhibited the lowest percentage agreement and greatest dispersion among generated descriptions. The descriptions generated for this image varied significantly, highlighting different morphological features and resulting in contrasting interpretations of embryo quality. For example, one description emphasized a thinning zona pellucida and a large blastocoel as indicators of good quality and high implantation potential, while another pointed out an elongated zona pellucida with less definitive implications for embryo viability. Such discrepancies reflect inherent ambiguity in the embryo's morphological features, resulting in varied diagnostic interpretations by the language models.

4.2 Interpretability Challenges in Embryo Assessment

This variability underscores the inherent complexity of accurately evaluating embryos with intermediate or subtle morphological features. Embryos with clearly identifiable abnormal characteristics such as those seen in third image from Table. 3, facilitate higher levels of agreement due to the unambiguous pathological features. Conversely, embryos like second imagen Table. 3, with subtle or ambiguous morphological indicators, pose challenges to automated diagnostic tools, resulting in decreased consensus and diagnostic confidence.

These findings highlight important limitations and opportunities for improvement in the use of language models for embryo assessment. Specifically, training language models with standardized, detailed, and precise descriptions as benchmarks appears essential to enhance their diagnostic accuracy, particularly for embryos presenting intermediate or unclear morphological characteristics. Future research could focus on incorporating highly standardized references during training phases to improve model convergence and accuracy, thus facilitating more consistent and reliable embryo assessments.

4.3 Model Performance and User Feedback

The performance of the models in terms of efficiency and hallucination met our initial expectations. The survey conducted among biology professionals yielded predominantly positive feedback regarding the models' ability to accurately describe embryo characteristics. Participants specifically noted the models' effectiveness in evaluating cell quality, suggesting that the descriptions generated by SMLs were clinically relevant and potentially useful in practice.

DeepSeek emerged as the preferred model among professionals, likely due to its optimal balance of description length and detail. This preference pattern suggests that concise, yet comprehensive descriptions may be most effective for clinical interpretation of embryological features. The moderate description length (averaging 45 words) provided by DeepSeek appears to strike an ideal balance between providing sufficient detail while remaining accessible to users.

4.4 Implications for AI Interpretability

Our findings demonstrate that Small Language Models (SMLs) can effectively generate descriptions and evaluations based on analysis from another AI model or Deep

Learning system with relatively little context. This capability represents a promising approach to AI interpretability, particularly in specialized medical domains like embryology. By translating complex feature representations from deep learning models into natural language descriptions, SMLs can potentially bridge the gap between sophisticated AI systems and human practitioners.

This interpretability layer addresses a critical limitation of deep learning systems in healthcare—their "black box" nature. By providing human-understandable explanations of AI-derived assessments, SMLs may enhance trust, facilitate error detection, and improve the clinical utility of AI systems in assisted reproduction technologies.

In summary, this analysis emphasizes the importance of clearly defined, standardized descriptors in training language models for embryological applications, particularly in cases where the morphological quality of embryos is not clearly delineated. The study demonstrates the potential of SMLs as interpretability tools for more complex AI systems, suggesting a promising direction for developing more robust assistants for interpretability tasks in healthcare and beyond.

5 Conclusion and Future Work

5.1 Summary of Findings

This research demonstrates that SMLs can effectively generate accurate descriptions of embryonic images using a Zero-Shot approach. Their key advantage is providing interpretability for complex Deep Learning models through accessible natural language descriptions. Among the tested models, DeepSeek performed best by balancing detail and conciseness, suggesting moderate description length optimizes clarity for specialist users.

5.2 Limitations

Our study has several limitations: the accuracy of SML-generated descriptions requires systematic validation against expert assessments; our evaluation relies on preference judgments rather than objective measures; and the sample size (27 images) limits generalizability across the full spectrum of embryo morphologies encountered clinically.

5.3 Future Work

Future research directions include:

- Collecting descriptions from senior embryologists to measure semantic distance with SML outputs.
- Evaluating textual quality using standard metrics like BLEU scores.
- Training models with specialized datasets containing detailed embryo descriptions and clinical outcomes.
- Investigating integration into clinical embryology workflows.

– Exploring additional data modalities like time-lapse imaging or genetic testing results.

6 Conclusion

SMLs show promise as interpretability tools for complex AI systems in reproductive medicine. By translating abstract features into accessible descriptions, they address the "black box" problem in healthcare AI. With domain-specific training, these models could develop into robust tools for embryo evaluation that enhance accuracy and clinical utility in assisted reproduction.

Acknowledgments. This study was possible thanks to IVF 2.0 LTD for providing data and computational resources. Thanks to Gustavo Guzmán for the figures.

References

1. Hong Lee, K., Lee, R., Eun Kwon Y.: Validation of a Deep Learning Chat X-ray Interpretation Model: Integration Large-Scale AI and Large Language Models for Comparative Analysis with ChatGPT. Diagnostics (2023).
2. Lingxuan, Z., Weiming M., Ancheng L.: Step into The Era Of Large Multimodal Models: A Pilot Study On Chatgpt-4(Visions) Ability To Interpret Radiological Images. International Journal of Surgery (2024).
3. Lee, T., Natalwala J., Chapple, V.: A brief history of artificial intelligence embryo selection: from black-box to glass-box (2024). doi: https://doi.org/10.1093/humrep/dead254.
4. Afnan, M.A.M., Liu, Y., Conitzer V.: Interpretable, not black-box, artificial intelligence should be used for embryo selection. Human Reproduction (2021): hoab040.
5. Lim, G., Elangovan, K., Jin, L.: Vision Language Models in Ophthalmology (2024).
6. Bosbach., W., Jan F., Senge, B.: Ability Of Chatgpt To Generate Competent Radiology Reports For Distal Radius Fracture By Use Of RSNA Template Items And Integrated AO Classifier. Current problems in diagnostic radiology (2023).
7. Meta AI.: Llama 3: A Family of Large Language Models. (2024) url: https://arxiv.org/abs/2402.06687.
8. Liu, H., Li, C., Wu, Q.: Visual Instruction Tuning. (2023) url: https://arxiv.org/abs/2304.08485.
9. OpenBMB. (2024). MiniCPM-V: Lightweight Multimodal Models for Efficient Deployment. https://arxiv.org/abs/2402.15733
10. DeepSeek-AI. (2025). DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. https://arxiv.org/abs/2501.12948
11. Chavez, A., Flores-Saiffe, A., Mendizabal, G., Drakeley, A., Cohen, J., Embryo Ranking Intelligent Classification (ERICA): Artificial Intelligence clinical assistant prediction embryo ploidy and implantation (2020). https://doi.org/10.1016/j.rbmo.2020.07.003.
12. Ollama Homepage, https://ollama.com/ last accessed 2025/03/07.
13. Zhang, J., Huang, J., Jin, S., & Lu, S. (2023). Vision-Language Models for Vision Tasks: A Survey. IEEE Transactions on Patter Analysis and Machine Intelligence, 46, 5625-5644. https://www.semanticscholar.org/reader/f38bf22a5ceef785d6a15263fad3f22f623a3e6b

ISSN 1870-4069

Revolutionizing the Fight against Child Cyberbullying: Using Holograms and Voice Recognition as Allies

María del Carmen Hidalgo Baeza[1,2], Alberto Ochoa Zezzatti[3,4], Víctor Manuel Casas Gómez[1], Irma Yazmín Hernández Baez[5]

[1] Universidad Tecnológica Fidel Velázquez, Mexico

[2] Universidad Internacional de Aguascalientes Manuel de Velasco Martínez, Mexico

[3] Universidad Autónoma de Ciudad Juárez, Mexico

[4] CADIT, Universidad Anáhuac, Mexico

[5] Universidad Politécnica de Morelos, Mexico

mcarmen.hidalgo@utfv.edu.mx, alberto.ochoa@uacj.edu.mx, victor.casas@utfv.edu.mx, Ihernandez@upemor.edu.mx,

Abstract. This research examines the feasibility and effectiveness of integrating holograms and voice recognition technologies as an innovative solution to combat child cyberbullying. By creating an empathetic virtual environment and enabling the early detection of emotional signals, the goal is to provide emotional support and prevent cyberbullying among children aged 9 to 14. The study addresses ethical and privacy considerations and underscores the importance of the responsible implementation of these emerging technologies. The primary focus is to thoroughly investigate the viability and effectiveness of this novel strategy, which combines holograms and voice recognition to establish a safe and compassionate virtual space. Early detection of emotional cues is crucial for identifying potential cyberbullying situations, enabling proactive intervention and timely emotional support. Beyond examining the technical feasibility and practical effectiveness of this solution, the research also carefully explores the ethical and privacy challenges associated with implementing emerging technologies in the context of child protection. Special attention is given to ensuring the responsible deployment of these tools; considering the sensitivity of the target population and establishing protocols that protect the integrity and privacy of participating children. The anticipated outcomes of this study aim not only to contribute insights into the technical viability and efficacy of the proposed solution but also to provide valuable perspectives on how to address child cyberbullying ethically and responsibly within an ever-evolving digital environment.

Keywords. Child cyberbullying, holograms, voice recognition, emotional well-being, innovative technologies, children and technology, digital ethics.

1 Introduction

El Child cyberbullying is a form of harassment among children and adolescents that occurs through digital technologies such as social media, instant messaging, or online video games.

This type of violence involves intentional and repeated harassment, humiliation, threats, and exclusion. Its psychological impact is profound, especially on children aged 9 to 14 a critical stage in the development of their identity and emotional stability.

According to a UNICEF report [1], 1 in 3 young people in the world has been a victim of online bullying. Nationally, the National Survey on the Availability and Use of Information Technologies in Households (ENDUTIH) [2] revealed that in Mexico, 23.5% of children and adolescents aged 12 to 17 have experienced cyberbullying, with a higher prevalence among girls.

In light of this alarming reality, it is urgent to adopt a comprehensive approach that recognizes the emotional vulnerability of minors and proposes proactive solutions. In this context, the innovative combination of holograms and voice recognition technologies emerges as a disruptive and empathetic tool.

These technologies aim not only to counteract the effects of cyberbullying but also to create safe digital spaces that provide emotional support for children.

The use of interactive holograms allows for a constant, compassionate virtual presence, functioning as a digital friend who listens and responds empathetically to the child's experiences, offering comfort and reducing feelings of isolation.

Meanwhile, emotional voice recognition acts as a digital sentinel capable of detecting emotional nuances in a child's speech, identifying patterns associated with bullying, anxiety, or distress. This enables early and personalized intervention that could prevent more severe consequences.

This technological approach goes beyond merely reacting to incidents; it empowers children with tools to manage their emotions, express their experiences, and navigate the digital environment with greater confidence.

Thus, technology is no longer seen as part of the problem but as a proactive ally in supporting children's emotional well-being.

However, the use of these technologies also raises ethical questions about privacy, consent, and the responsible use of emotional data. As these tools continue to evolve, it is essential that their implementation is grounded in a solid ethical framework that places the child and their right to protection and development at the center.

2 Theoretical Framework

2.1 Holography

Holography is a photographic technique that creates three-dimensional images by projecting a beam of light onto a refractive material. When the material receives light from an appropriate angle, it projects an image in three dimensions.

2.2 Hologram

A hologram is a three-dimensional image of an object that can be viewed from any angle. This means that the light reaching the viewer's eyes from the hologram is physically the same as that emitted by the original object. Holograma.

2.3 History of Holography

In 1947, Hungarian physicist Dennis Gabor was searching for a method to improve the resolution and definition of the electron microscope when he accidentally discovered a new technique for forming images. The object he used to create his first hologram was a transparent circular slide containing the names of three physicists he considered significant: Huygens, Young, and Fresnel. He called this process holography, from the Greek holos (whole) and graphos (to write). Although his discovery was initially unsuccessful, the invention of the laser later enabled the development of numerous scientific and technological applications based on holography. In recognition of his contributions, Gabor was awarded the Nobel Prize in Physics in 1971.

With technological advancement, techniques for generating images or videos using modeling software such as Autodesk and projecting holograms using projectors or monitors have become common. The process typically begins by capturing a physical model of any object or person with cameras positioned from multiple angles. This allows the creation of a reference image for the final hologram design. The captured images, known as perspectives, are sent to the modeling software where they are edited and shaped into the desired 3D figure.

Another technique, known as the particle model method, involves generating images through heated air particles. The image, produced by a projector or laser, is reflected by these particles, resulting in a three-dimensional image. Because the hologram is a reflection of the projected image, it appears in full color. The air can be heated using a radiator. As the particles rise in wave patterns, they form a sort of mirror through the hot air currents. Dust particles in the air also contribute to better light reflection and dimension at specific angles. One major advantage of this type of hologram is its ability to be interactive and penetrable it exists in a nonphysical medium, so the image does not distort when touched. These characteristics make it particularly useful for presentations and exhibitions.

2.4 Characteristics of Light Waves

Light is a form of electromagnetic radiation visible to the human eye. It has a dual nature, behaving both as a wave and as a particle (photon). Light waves can propagate even in a vacuum, without the need for a material medium.

When light strikes a surface, reflection occurs, allowing objects to be seen by returning part of the light to the environment [3]. Depending on the surface, reflection can be classified as:

- Specular reflection: Occurs on smooth surfaces such as mirrors, where the reflected rays remain parallel [4].
- Diffuse reflection: Occurs on rough surfaces, causing the rays to scatter in different directions.
- Extended reflection: A combination of specular and diffuse reflection, typical of semi-polished surfaces.
- Scattered reflection: Irregular dispersion that does not follow the classical laws of reflection.

The law of reflection applies: the angle of incidence is equal to the angle of reflection, and both rays (incident and reflected) lie in the same plane with respect to the normal (a line perpendicular to the surface) [5].

3 Method

3.1 Methodology

This study was conducted using a prospective approach for data collection. According to information provided by the Secretariat of Education, Science, Technology, and Innovation (CCT-SECTI), the municipality of Nicolás Romero has a total of 452 educational institutions, of which 78 are at the primary level and 42 at the secondary level [6]. The target population consisted of girls, boys, and adolescents between the ages of 9 and 14, enrolled in the fourth, fifth, and 6th grades of primary school, as well as the 1st and 2nd grades of secondary school, in both public and private institutions.

To calculate the sample size, the formula for a finite population (N) was used. This is an adaptation of the formula for infinite populations, adjusted to account for the effect of a limited population:

$$n = (N * Z^2 * p * (1 - p)) / (e^2 * (N - 1) + Z^2 * p * (1 - p)),$$

where:
n = sample size,
N = total population (total number of students),
Z = Z critical value of 1.96 (for 95% confidence level),
p = estimated proportion of the population with the characteristic of interest (0.5, for maximum variability),

```python
@author: Carmen
"""

def calculate_sample_size(N, Z=1.96, p=0.5, e=0.05):
    q = 1 - p
    numerator = N * (Z**2) * p * q
    denominator = ((N - 1) * (e**2)) + ((Z**2) * p * q)
    n = numerator / denominator
    return round(n)

# Application with N = 68232
sample_size = calculate_sample_size(68232)
print(f"Required sample size: {sample_size}")
```

Fig. 1. Sample Size Calculation in Python.

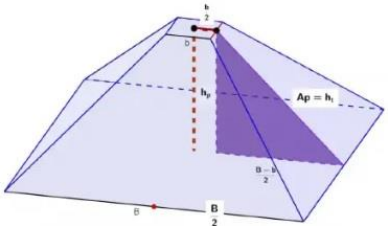


Fig. 2. Holographic pyramid.

e = margin of error (typically 0.05 for 5%).

This formula is an adaptation of the sample size calculation for infinite populations, adjusted to account for the limited size of the population.

As a data collection technique, a survey was conducted using a structured questionnaire consisting of 17 closed-ended questions with multiple-choice options, divided into five sections:

1. Demographic Information.
2. Experiences with Cyberbullying (children and adolescents aged 9 to 14).
3. Use of Emerging Technologies to Detect Cyberbullying.
4. Perceptions of Parents and Educators on Cyberbullying and Emerging Technologies.
5. General Opinions.

3.2 Holographic Implementation

3.2.1 Holographic Pyramid

The holographic pyramid is a system that is used to project all kinds of elements in three dimensions on a metal structure, cardboard on which the light-emitting equipment rests, which will be the pyramid as shown in figure 2.

When the film is illuminated with the same light source, a three-dimensional image of the original object is produced:

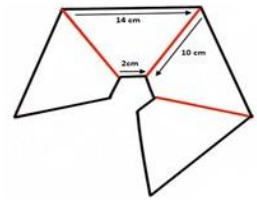


Fig. 3. Paper mold for trapezoidsde.

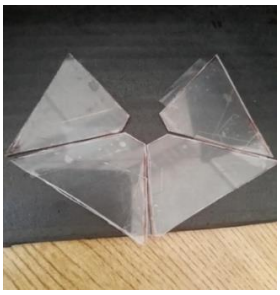


Fig. 4. Four-piece trapezoid-shaped cuts.

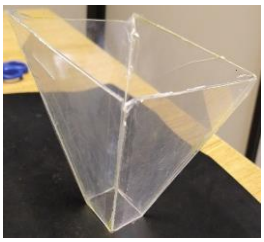


Fig. 5. Holographic Pyramid.

To build the pyramid, several tests were carried out on the following materials.

- In plastic the image looks relatively good but somewhat opaque.
- In acrylic, depending on the thickness, the image was cut or duplicated.
- On acetate the image is seen with less definition.

For the construction of the pyramid, it is necessary to draw a trapezoid as shown in figure 2 with the following measurements: minor base 2 centimeters, major base 14 centimeters and height 10 centimeters.

The trapezoids are drawn and then the trapezium-shaped molds are cut, four cuts are made on CD covers, in the shape of a trapezoid as shown in figure 4.

The four pieces are glued with tape and secured with silicone, the final result is shown in Figure 4.

3.2.2 Hologram Projector Box

The following materials were used for the construction of the hologram projector box:

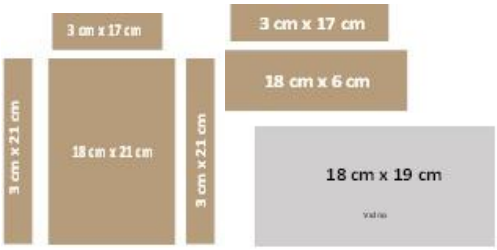


Fig. 6. Projector Box Parts.

Fig. 7. Projector Box for the Hologram.

- Cardboard: for the structural prototype, due to its low cost and easy handling.
- Silicone: as an adhesive, adding stability and sealing.
- Matte black paint: applied internally to reduce reflections.
- Glass: as a projection surface, allowing the adequate transmission and reflection of light.

The measurements of each of the pieces are shown in Figure 6.
This modular design allows holographic images to be reproduced with good visual quality in low-light environments, offering an effective tool for educational demonstrations, optical experiments or multimedia applications. The result of the construction of the project box is shown in Figure 7.

3.3 Artificial Intelligence (AI) System Components

3.3.1 Cyberbullying Detection System

- Trained with NLP (Natural Language Processing) models.
- Able to detect:
    - Offensive language, threats, insults,
    - Sarcasm or microaggressions through semantic analysis,

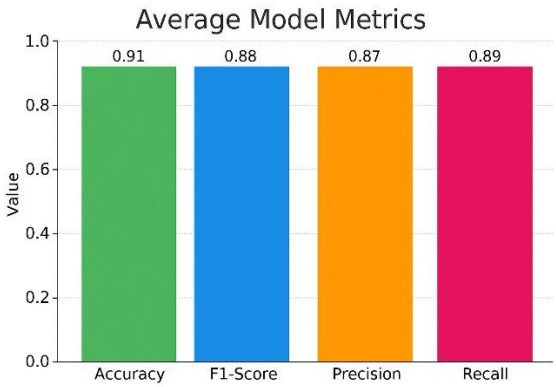


Fig. 8. Average model metrics.

- Offensive images and memes with computer vision,
- Real-time integration into chats, social networks, messaging apps, etc.

- Real-time integration into chats, social networks, messaging apps, etc.

3.3.2 Model and Architecture

Implement the artificial intelligence model capable of denouncing offensive or cyberbullying comments in free text.
Base model: BERT (Bidirectional Encoder Representations from Transformers)
Architecture: 12 layers of transformers, 110 million parameters
Dataset used: HateXplain, Cyberbullying Detection Dataset, OLID (Offensive Language Identification Dataset)
Techniques applied supervised fine.tuning, data augumetation, tokenization with WordPiece.

- Evaluation metrics: Accuracy, F1-score, Recall y Precision.
- Results achieved:
  - Accuracy: 91.3%,
  - F1-score: 0.88 (bullying class) indicating high detection performance,
  - Training Time: 6 hours on GPU (NVIDIA RTX 3090).

Figure 8 shows the average metrics obtained by the model during validation.

3.3.1 Interactive Hologram

- Projection using hologram technology (Pepper's Ghost).
- Design with an empathetic, friendly and reassuring personality.
- Interact with the child:

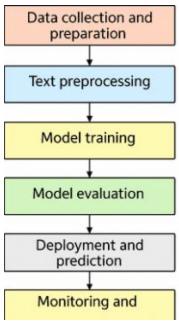


Fig. 9. Machine Learning Model.

- Offers comfort and tools to respond to bullying.
- Gives advice on how to act.
- Automatic help protocols can be activated.

3.3.4 Emergency Response Module

Connected with:

- Contactosrusted contacts (Parent, guardians).
- Educational institutions.
- In severe cases, local emergency services.

Must be customized according to country/laws/locatione

3.3.5 Ethics and Privacy Module

- Protection of minors' data.
- End-to-end encryption.
- Informed consent from parents and schools.
- Periodic ethical auditsn.

3.4 Development process

3.4.1 Phase 1: Research and Prototyping

- Case studies of child bullying.
- Compilation of real datasets.
- Hologram Prototype Design (Software & Hardware).

3.4.2 Phase 2: Training the AI Model

- Using the BERT Model for Language Analysis.
- Train on specific bullying datasets.

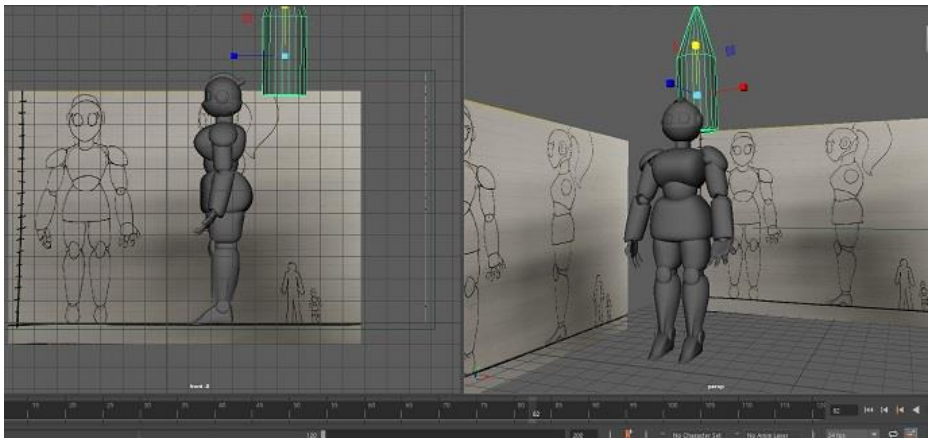


Fig. 10. Front and side character modeling with Autodesk Maya.

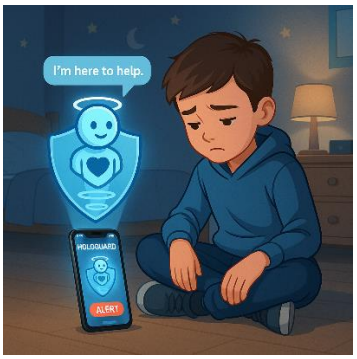


Fig. 11. Integration of Artificial Intelligence, hologram and response protocol.

– Accuracy testing, minimization of false.

3.4.3 Development of the Hologram

- Diseño3D design of the wizard (friendly character).
- Integration with display devices.

3.4.4 Phase 3: Pilot tests

Deploying the system in a controlled school environment.

3.4.5 Total Integration

Connection between Artificial Intelligence + Hologram + Response Protocol.
Simulation of real cases.

3.5 Exploring Prevention, Follow-up and Argumentation Variables in Child Cyberbullying

The preliminary experimental design to explore variables that could influence the effectiveness of strategies for the prevention and management of childhood cyberbullying. The objective is to identify effective combinations between technology, educational intervention and social participation. The variables are:

1. Prevention Strategies:
   - Levels: Digital education, awareness and promotion of healthy online behaviors.
   - Method: Comparison between groups with and without intervention.
2. Tracking Technologies:
   - Levels: Holograms, voice recognition and combination.
   - Method: Measurement of effectiveness in early detection and intervention.
3. Caregiver Involvement:
   - Levels: Low, moderate and high.
   - Method: To evaluate the impact according to the level of early supervision.
4. Frequency of Interventions:
   - Levels: Scheduled versus interventions based on system alerts.
   - Method: To compare the effect on incidence and emotional impact.
5. Emotional Feedback from Technology:
   - Levels: Technology with or without emotional response.
   - Method: Analysis of its influence on the child's response and the overall effectiveness of the system.
6. Integration into the School Curriculum:
   - Levels: School curriculum incorporating cybersecurity education versus standard curriculum.
   - Method: To evaluate the incidence of cyberbullying in different school contexts.
7. Post-Incident Psychological Support:
   - Levels: Different access to support services.
   - Method: Evaluation of well-being according to the level of care received.

8. Impact Assessment in the School Community:

- Levels: Degree of Participation of the school community.
- Method: Analysis of the effect on sustainability and effectiveness of interventions against cyberbullying.

This design provides a comprehensive view of the factors that affect the effectiveness of interventions against child cyberbullying.

4 Discussion of Results

This study explores the effectiveness of strategies for the prevention and management of childhood cyberbullying through a preliminary experimental design. It is proposed to evaluate various variables using Likert screens to identify effective combinations between technology, education and social participation.

1. Cyberbullying Prevention:

Participants will assess the clarity, usefulness and effectiveness of the educational strategies implemented and speech recognition.

2. Technological Monitoring:

Analyze the perceived ease of use, efficacy, and security of technologies such as holograms and speech recognition.

3. Caregiver Involvement:

Assess the level of involvement, quality of emotional support, and collaboration with technology.

4. Frequency of Interventions and Emotional Feedback:

To compare scheduled and alert-triggered interventions and assess the impact of emotional feedback on child well-being.

5. Integration Curricular

Inclusion of cybersecurity content in the school curriculum, highlighting its relevance and usefulness.

6. Subsequent psychological support:

Evaluate the availability, accessibility and empathetic quality of post-incident care services.

7. Community Impact:

Value school participation and the perception of a safer and more supportive community.

The results will allow us to identify patterns of acceptance, correlation between variables and key elements to optimize future interventions.

The implementation of innovative technologies requires guaranteeing the emotional protection and privacy of children. To do this, it is essential to follow clear guidelines on privacy and consent in data collection, ensuring that it is informed and voluntary, and establishing transparent policies on its use and access. Information security must be supported by effective cybersecurity measures, as well as confidential and anonymous handling of sensitive data. Technological transparency is also essential, which implies clear communication with children and their caregivers about the operation and purpose of the technologies used.

Likewise, a continuous ethical evaluation must be carried out that considers all the actors involved and respect for children's rights must be integrated in all phases of the project. These elements are key to strengthening the trust, legitimacy and effectiveness of the technologies used in the prevention of child cyberbullying.

5 Conclusion

The integration of emerging technologies, such as interactive holograms and voice recognition systems, is an innovative proposal in the prevention of child cyberbullying. This technological synergy not only addresses current problems, but also projects a sustainable and adaptive emotional support model, aligned with the well-being and protection of children in digital environments.

From a psychoeducational perspective, the constant presence of the hologram acts as an emotional support agent that validates the child's experiences and reinforces their resilience and self-esteem. At the same time, the ability to recognize speech allows the early identification of linguistic patterns associated with risk situations, enabling preventive intervention and not merely reactive intervention.

Likewise, the system promotes a more inclusive and emotionally safe educational environment, by integrating continuous feedback and affective accompaniment. Its evolutionary design allows it to adapt to the transformations of child development, guaranteeing a personalized and contextualized response.

The consolidation of this proposal requires a research and development agenda focused on the following axes:

- Optimization of artificial intelligence algorithms sensitive to the emotional context, for greater precision in the interpretation of affective signals.
- Incorporation of emerging technologies (virtual reality, artificial emotional intelligence) that enhance the accompaniment experience.
- Cultural and contextual adaptability, to ensure effective implementation in diverse contexts.
- Interdisciplinary collaboration with experts in children's mental health, ensuring the therapeutic rigor of the technological design.

- Development of ethical frameworks and robust privacy protocols, fundamental for the trust and acceptance of the system.
- Long-term impact monitoring, to assess the sustained effect on users' emotional health.
- Active participation of the educational community, as a key agent in the implementation and continuous improvement of the system.
- Constant updating in the face of technological and social changes, guaranteeing the validity of the proposed model.

These lines of action seek to strengthen the multidisciplinary and inclusive approach to intervention, consolidating a digital environment that prioritizes children's rights, emotional health, and safety.

Acknowledgments. This article has been funded by the Fidel Velázquez Technological University, to whom we express our gratitude for its valuable support.

References

1. Instituto Nacional de Estadística y Geografía (INEGI): Encuesta Nacional sobre Disponibilidad y Uso de Tecnologías de la Información en los Hogares (ENDUTIH) 2020 (2021) url: https://www.inegi.org.mx/programas/dutih/2020/
2. UNICEF: UNICEF Poll: More Than a Third Of Young People in 30 Countries Report Being A Victim Of Online Bullying. (2019) url: https://www.unicef.org/press-releases/unicef-poll-more-third-young-people-30-countries-report-being-victim-online-bullying.
3. Giambattista, A., Richardson, B. M., Richardson, R. C.: College physics (fourth Ed.). McGraw-Hill Education (2012)
4. Young, H. D., Freedman, R. A. University physics with modern physics (12th Ed.). Pearson/Addison-Wesley (2010)
5. Hecht, E.: Optics (4th Ed.). Addison Wesley (2002)
6. Secretaría de Educación, Ciencia, Tecnología e Innovación del Estado de México.: Catálogo de Centros de Trabajo (CCT-SECTI). (2024) url: https://seduc.edomex.gob.mx/sis/catalogoct/.
7. Bostrom, N., Yudkowsky, E.: The Ethics of Artificial Intelligence. In: W. M. Ramsey & K. Frankish (Eds.), The Cambridge Handbook of Artificial Intelligence. Cambridge University Press, pp. 316–334 (2014) doi: 10.1017/CBO9781139046855.020.
8. Bauman, S., Cross, D., Walker, J.: Principles of Cyberbullying Research: Definitions, Measures, and Methodology. Routledge. (2013) doi: 10.4324/9780203107723.
9. Cheng, L., Li, X., Zhai, Q.: Cyberbullying and Psychological Distress: The Moderating Roles of Emotional Intelligence and Gender. Computers in Human Behavior, 121 (106776) (2021) doi: 10.1016/j.chb.2021.106776.
10. Creswick, H., Hill, P., Lodge, J. M.: The Role of Artificial Intelligence in Cyberbullying Detection: A Systematic Review. Computers in Human Behavior, 103, pp. 259–277 (2019) doi: 10.1016/j.chb.2019.09.019.
11. Dede, C.: Comparing Frameworks for 21st Century Skills. In: J. A. Bellanca & R. S. Brandt (Eds.), 21st Century Skills: Rethinking How Students Learn, pp. 51–76 (2010)

12. Epstein, J. L.: School, Family, and Community Partnerships: Preparing Educators and Improving Schools. Westview Press. (2010)
13. Floridi, L.: The Ontological Interpretation of Informational Privacy. Ethics and Information Technology, 7(4), pp. 185–200 (2005) doi: 10.1007/s10676-006-0001-7.
14. Huang, Y., Chou, C.: An Analysis of Multiple Factors of Cyberbullying Among Junior High School Students in Taiwan. Computers in Human Behavior, 26(6), pp. 1581–1590 (2010) doi: 10.1016/j.chb.2010.06.005.
15. Kish, F., Azuma, R. T.: A Study of the Limitations of Head-Worn Displays. Proceedings of SPIE. The International Society for Optical Engineering, 2177, pp. 158–169 (1994) doi: 10.1117/12.175274.
16. Kazdin, A. E., Rabbitt, S. M.: Novel Models for Delivering Mental Health Services and Reducing the Burdens of Mental Illness. Perspectives on Psychological Science, 8(1), pp. 22–29 (2013) doi: 10.1177/1745691612466676.
17. Livingstone, S., Haddon, L., Görzig, A.: Risks and Safety on the Internet: The Perspective of European Children. EU Kids Online. (2011) doi: 10.1007/978-94-007-1967-4_4.
18. Picard, R. W., Vyzas, E., Healey, J.: Toward Machine Emotional Intelligence: Analysis of Affective Physiological State. In: IEEE Transactions on Pattern Analysis and Machine Intelligence, 23(10), pp. 1175–1191 (2001) doi: 10.1109/34.954607.
19. Pérez-Fuentes, M. C., Molero Jurado, M. M., Gázquez Linares, J. J.: Cybervictimization, Self-Esteem and Social Support in Adolescents: A Structural Equation Model. International Journal of Environmental Research and Public Health, 18(2), pp. 702 (2021) doi: 10.3390/ijerph18020702.
20. Rabiner, L. R., Juang, B. H.: Fundamentals of Speech Recognition. Prentice Hall (1993)
21. Russell, S., Norvig, P.: Artificial Intelligence: A Modern Approach (3ª ed.). Prentice Hall (2010)
22. Slonje, R., Smith, P. K., Frisén, A.: The Nature of Cyberbullying, and Strategies for Prevention. Computers in Human Behavior, 29(1), pp. 26–32 (2013) doi: 10.1016/j.chb.2012.05.024.
23. Smith, P. K., Mahdavi, J., Carvalho, M.: Cyberbullying: Its Nature and Impact in Secondary School Pupils. Journal of Child Psychology and Psychiatry, 49(4), pp. 376–385 (2008) doi: 10.1111/j.1469-7610.2007.01846.x
24. Soni, D., Roberts, S.: Predicting Cyberbullying Incidents on Social Media Using Machine Learning Approaches. Journal of Artificial Intelligence Research, 64, pp. 355389 (2019) doi: 10.1613/jair.1.11344.
25. Turkle, S.: Alone Together: Why We Expect More from Technology and Less from Each Other. Basic Books (2011)
26. Tokunaga, R. S.: Following You Home from School: A Critical Review and Synthesis of Research on Cyberbullying Victimization. Computers in Human Behavior, 26(3), pp. 277–287 (2010) doi: 10.1016/j.chb.2009.11.014.
27. Tangen, D., Campbell, M.: Cyberbullying Prevention: One Primary School's Approach. Australian Journal of Guidance and Counselling, 20(2), pp. 225–234 (2010) doi: 10.1375/ajgc.20.2.225.
28. Warschauer, M., Matuchniak, T.: New Technology and Digital Worlds: Analyzing Evidence of Equity in Access, Use, and Outcomes. Review of Research in Education, 34(1), pp. 179–225 (2010) doi: 10.3102/0091732X09349791.
29. Xu, J. M., Jun, K. S., Zhu, X.: Learning from Bullying Traces in Social Media. Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 656–666 (2012) doi: 10.5555/2382029.2382133.

ISSN 1870-4069

Clasificación de datos de cotización de la bolsa mexicana de valores usando aprendizaje automático

José Gonzalo Ramírez Rosas, Jorge de la Calleja,
Araceli Ortiz Carranco, Martin Neri-Suárez,
Salvador Antonio Arroyo Diaz

Universidad Politécnica de Puebla,
México

{jose.ramirez, jorge.delacalleja, araceli.ortiz,
martin.neri, salvador.arroyo}@uppuebla.edu.mx

Resumen. Las predicciones en el sector bursátil han sido un factor clave para que los inversionistas puedan tomar mejores decisiones y se minimice el riesgo que afrontan. Tradicionalmente, esta actividad se ha realizado manualmente o bien usando algunas herramientas de cómputo, sin embargo, en los últimos años la inteligencia artificial se ha empezado a aplicar en el área financiera para detectar patrones de comportamiento y así obtener mejores resultados en la predicción de datos. En el presente artículo se muestran los primeros resultados de un trabajo en desarrollo para la clasificación de los precios de cotización de algunas de las empresas que cotizan en el sector de productos de consumo frecuente de la Bolsa Mexicana de Valores, con el fin de apoyar en la toma de decisiones a los inversionistas. Para realizar los experimentos se obtuvieron los datos históricos (precios de cotización) de las empresas Bimbo, Chedraui, Femsa, Soriana y Walmex del periodo de 2010 al 2024. El software Weka se empleó para obtener los resultados de los algoritmos de aprendizaje automático utilizados tales como: el clasificador simple de Bayes, k-vecinos más cercanos, redes neuronales artificiales y máquinas de vectores de soporte, destacando árboles de decisión con una exactitud de clasificación con un promedio del 99%.

Keywords: Aprendizaje automático, clasificación, precios de cotización, finanzas.

Classification of Mexican Stock Exchange Quote Data Using Machine Learning

Abstract. Forecasting in the stock market has been a key factor in helping investors make better decisions and minimize the risk they face. Traditionally, this activity has been performed manually or using computing tools; however, in recent years, artificial intelligence has begun to be applied in the financial sector to detect behavioral patterns and thus obtain better results in data prediction. This article presents the initial results of an ongoing project to classify the stock prices of some of the companies listed in the consumer products sector of the Mexican Stock Exchange, with the aim of supporting investors in their decision-making.

To carry out the experiments, historical data (listed prices) were obtained from the companies Bimbo, Chedraui, Femsa, Soriana and Walmex from 2010 to 2024. The Weka software was used to obtain the results of the machine learning algorithms used such as: the simple Bayes classifier, k-nearest neighbors, artificial neural networks and support vector machines, highlighting decision trees with an average classification accuracy of 99%.

Keywords: Machine learning, classification, quote prices, finance.

1. Introducción

Dentro del sector financiero, el mercado bursátil juega un papel fundamental en la economía de los países. A través de éste, las empresas pueden acceder a financiamiento para llevar a cabo distintos tipos de inversión permitiéndoles expandirse, innovar y alcanzar sus metas financieras. Al mismo tiempo, el público inversionista, ya sean personas físicas o morales, busca incrementar su patrimonio mediante los rendimientos que ofrecen los mercados financieros. Estos rendimientos dependen de la variación en los precios de cotización de los activos financieros, los cuales reflejan el valor en moneda nacional de un activo en un momento determinado.

Es importante señalar que, el mercado bursátil se caracteriza por su dinamismo y volatilidad, lo que implica que los precios de los activos pueden fluctuar drásticamente en cortos periodos de tiempo. Esta volatilidad genera incertidumbre en los inversionistas, quienes deben tomar decisiones en un entorno en constante cambio. Además, se presenta el problema de la información asimétrica, donde no todos los inversionistas tienen acceso a la misma cantidad y calidad de información, lo que puede llevar a decisiones desinformadas Baca y Marcelino (2016) [1].

Así mismo, otro factor que influye en la volatilidad del mercado son los eventos externos como crisis económicas, cambios en políticas públicas y desastres naturales, los cuales pueden causar caídas abruptas en los precios de los activos financieros. Estos factores dificultan la predicción de los movimientos del mercado y aumentan el riesgo de inversión. Ante estos desafíos, es fundamental contar con herramientas que permitan analizar grandes volúmenes de datos y detectar patrones que ayuden a reducir la incertidumbre García (2015) [2].

Por otra parte, dentro del sector bursátil el dato de cotización se representa mediante valores como el precio de apertura, el precio más bajo, el precio más alto y el precio de cierre de un activo durante un periodo determinado. Estos datos son fundamentales para calcular las ganancias y pérdidas, así como para tomar decisiones estratégicas sobre la compra o venta de activos.

Dichas decisiones en el mercado de valores a menudo se reducen a una elección binaria: comprar o vender un activo en función de su comportamiento histórico y su tendencia futura. Para facilitar este proceso de decisión, los modelos de aprendizaje automático analizan las variables y generan predicciones basadas en patrones previamente identificados.

De lo anterior, el aprendizaje automático ha demostrado ser una tecnología clave para mejorar la toma de decisiones en el ámbito bursátil. Gracias a su capacidad de analizar grandes cantidades de datos históricos y detectar correlaciones, esta herramienta permite estimar tendencias del mercado con mayor exactitud. De esta manera, los inversionistas pueden minimizar riesgos y maximizar oportunidades de inversión Raschka y Mirjalili (2019) [3].

En este contexto el aprendizaje automático se puede entender como una rama de la inteligencia artificial (IA), que ha revolucionado el análisis financiero al proporcionar modelos predictivos y de clasificación más precisos y eficientes.

En el presente trabajo se muestran los primeros resultados de la evaluación del desempeño de algoritmos de aprendizaje automático para clasificar datos de valores de cotización de la BMV. Así, en una primera aproximación se podrá conocer cuál algoritmo es el que mejores resultados proporciona para que los inversionistas puedan usarlo como apoyo para la toma de decisiones para comprar o vender un activo financiero.

2. Trabajos relacionados

Para el desarrollo de la presente sección, la búsqueda de los trabajos se realizó en las bases de datos de *Scopus, Google Scholar y EBSCO,* utilizando como palabras clave en idioma inglés: *stock market and artificial intelligence, deep learning and finance;* seleccionando los artículos más cercanos con uso de algoritmos de aprendizaje automático en el sector bursátil.

De esta búsqueda se encontró que varios algoritmos de aprendizaje automático se han estado aplicando en la predicción y clasificación de datos financieros, destacando los algoritmos de: máquinas de vectores de soporte, redes neuronales artificiales, árboles de decisión, modelos de series de tiempo y modelos de memoria a corto plazo (LSTM). Enseguida se describen entonces aquellos trabajos más afines con el que se está desarrollando.

Para comenzar, Henrique, Sabreiro y Kimura en 2019 [4] expusieron que el algoritmo de máquina de soporte vectorial es una técnica efectiva para realizar predicciones a corto plazo, especialmente cuando se utilizan ventanas de tiempo reducidas. Su capacidad para capturar patrones en datos recientes la hace una opción confiable en escenarios de alta frecuencia, no obstante, su precisión tiende a disminuir en pronósticos a largo plazo debido a la creciente volatilidad del mercado, por lo tanto, factores como cambios inesperados en las tendencias y la acumulación de errores afectan su desempeño, lo que limita su aplicación en horizontes temporales más amplios.

Así también, Strader, Rozycki, Root y Huang en 2020 [5] sugieren que el aprendizaje automático posee un potencial significativo en la predicción del mercado, sin embargo, su eficacia depende de diversos factores, como la calidad de los datos y la robustez de los modelos empleados; por lo anterior para mejorar la precisión de estas predicciones, es fundamental realizar más estudios sobre la integración de múltiples fuentes de datos y la optimización de los algoritmos, así como modificar los hiperparametros, esto

permitirá desarrollar modelos más sofisticados y adaptativos, capaces de captar mejor la dinámica del mercado.

De la misma forma, los modelos de aprendizaje automático, como las redes neuronales y los árboles de decisión, han demostrado un desempeño sólido en la predicción del mercado, su capacidad para identificar patrones complejos y adaptarse a grandes volúmenes de datos los convierte en herramientas valiosas para el análisis financiero; además, la precisión de las predicciones mejora significativamente cuando se combinan varios métodos de inteligencia artificial. Tal como la integración de enfoques complementarios que permite reducir errores y optimizar el procesamiento de la información, así como obtener resultados más confiables en diferentes condiciones del mercado Sahrab, Kang y Jin, 2021[6].

Coy, Granados y García en 2021[7] expusieron que los modelos de series temporales, como ARIMA y GARCH, continúan siendo herramientas fundamentales en la predicción financiera, su capacidad para modelar la dependencia temporal y la volatilidad los hace ampliamente utilizados en el análisis de tendencias y riesgos de los mercados, no obstante, la creciente complejidad de los datos financieros ha evidenciado la necesidad de modelos más avanzados. En este contexto, las redes neuronales recurrentes han surgido como una alternativa prometedora, ya que pueden capturar patrones no lineales y adaptarse mejor a las dinámicas cambiantes del mercado.

Por su parte, Cuevas, Alvares, Azcona y Rodríguez en 2019[8] en su estudio desarrollaron un modelo para pronosticar la proyección de los ingresos de una empresa a través del algoritmo de máquinas de soporte vectorial, si bien el estudio no fue desarrollado para el sector bursátil, fue empleado en el área financiera, teniendo como resultado que este algoritmo es adecuado para la clasificación y predicción de series de tiempo en el ámbito financiero.

Chhajer, Shah y Kshirsagar en el 2022 [9] demostraron que los *Long short-term memory* (LSTM) mostraron un rendimiento superior en la predicción a largo plazo gracias a su capacidad para manejar secuencias temporales de manera efectiva, su arquitectura permite retener información relevante a lo largo del tiempo, lo que mejora la precisión en la identificación de tendencias y patrones complejos en los datos de cotización. Por otro lado, los modelos de máquina de soporte vectorial y redes neuronales, también obtuvieron buenos resultados, aunque presentaron limitaciones en la captura de relaciones temporales más complejas. Si bien pueden ser útiles en ciertos contextos, su desempeño es menos eficiente en comparación con LSTM cuando se trata de datos secuenciales de largo plazo.

Cabe señalar que, García, Garzón y López en 2023 [10] concluyeron que las técnicas de *soft computing*, como las redes neuronales, la lógica difusa y los algoritmos genéticos, ofrecen ventajas significativas sobre los métodos tradicionales de predicción de mercados, su capacidad para procesar grandes volúmenes de datos y reconocer patrones no lineales les permite abordar la naturaleza dinámica y cambiante de los mercados financieros. Además, estas técnicas destacan por su mayor flexibilidad y adaptabilidad, lo que las hace especialmente eficaces para modelar incertidumbre y variabilidad, a diferencia de los enfoques convencionales, pueden ajustarse de manera más precisa a condiciones de mercado complejas, mejorando así la toma de decisiones en entornos financieros.

De la misma manera, Lin y Lobo en 2024 [11] destacan el crecimiento en el uso de técnicas de inteligencia artificial, como redes neuronales y algoritmos de aprendizaje

profundo, en la predicción del mercado, estas herramientas han demostrado ser altamente efectivas para identificar patrones complejos y mejorar la capacidad de anticipación en entornos financieros dinámicos, además la combinación de inteligencia artificial con análisis técnico y fundamental ha permitido mejorar significativamente la precisión en las predicciones, esto al integrar múltiples fuentes de información con enfoques que ofrecen una visión más completa del comportamiento del mercado, optimizando así la toma de decisiones y reduciendo la incertidumbre.

Igualmente, se destaca el uso de algoritmos como las redes neuronales y Long-Short-Term Memory (LSTM), para la predicción de los índices bursátiles en especial el IPC de México y el Standard & Poor´s (S&P), los cuales fueron eficaces para la contribución de la gestión económica y optimización de recursos en la toma de decisión, arrojando como resultados positivos para la predicción al algoritmo LSTM presentando errores por debajo del 4% Andrade et al 2024 [12]

También, otro estudio de Ricchiuti y Sperlí en 2025 [13] aplicaron los LSTM utilizado en el mercado de criptomonedas que ha demostrado la eficiencia en el retorno de los rendimientos, pues la predicción con respecto a los precios de cotización de estos activos demostró un retorno del 39%

Sin embargo, existen estudios que están utilizando la inteligencia artificial basados en datos de procesamiento de lenguaje natural para poder clasificar las acciones del mercado bursátil y con ello poder hacer una predicción con respecto a las tendencias que presentan los activos Aritonang, Wiryono y Faturohman 2025 [14]

Así mismo, la inteligencia artificial aplicada en las finanzas ha tomado gran relevancia, pues se ha tenido evidencia científica de que las personas están utilizando los algoritmos de aprendizaje automático para gestionar sus carteras de inversión, esto con el fin de maximizar sus rendimientos y minimizar los riesgos que presentan sus inversiones Banerjee, 2025 [15]

En un estudio de Ning, Zhaang y Wang en 2025 [16], experimentaron con los algoritmos de aprendizaje automático para minimizar los riesgos en las inversiones, los hallazgos fueron que a través del uso de los datos de las empresas que cotizan en la bolsa de Shanghái y Shenzhen por el periodo de 2007 al 2020 estabilizó el riesgo de la caída en las acciones y se pudo mejorar el desempeño de las inversiones en el mercado de capitales de aquella bolsa bursátil.

De lo presentado anteriormente, se puede notar que los algoritmos de árboles de decisión y redes neuronales destacaron como los de mejor desempeño en la clasificación de datos de cotización, como lo señalan Sahrab, Kang y Jin en su estudio de 2021, sin embargo, como se ha visualizado el algoritmo *Long-Short-Term Memory* ha demostrado resultados favorables en la parte bursátil como lo señalan los estudios de Andrade et al en 2024 y de Chhajer, Shah y Kshirsagar en 2022, pero, en su esencia éste algoritmo no presenta una funcionalidad para los datos de serie de tiempo como lo exponen en su mismo trabajo.

3. Metodología y métodos

La presente investigación tiene un corte de enfoque cuantitativo dado que se está midiendo la exactitud de la clasificación de los datos de cotización de las empresas seleccionadas. Desde el punto de vista de la finalidad, se trata de una investigación

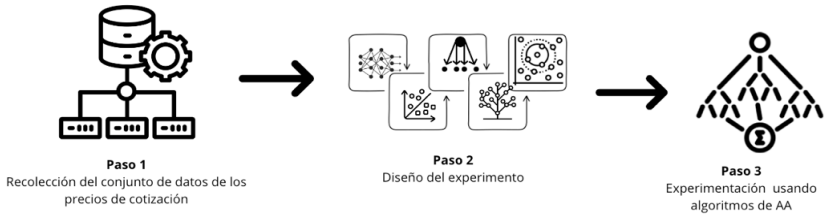


Fig. 1. Etapas principales para realizar la clasificación de los datos de cotización.

aplicada porque se podrá conocer el algoritmo que mejor desempeño de clasificación tendrá en los datos de cotización.

3.1. Metodología

En la figura 1 se muestra un diagrama de las principales etapas para el desarrollo del presente trabajo, que son: recolección de datos, el diseño del experimento y la experimentación usando los algoritmos de aprendizaje automático del software WEKA (*Waikato Environment for Knowledge Analysis*); a continuación, se describe cada una de éstas.

Etapa 1. En este paso, se diseñó un sistema de descarga de datos en lenguaje Python utilizando la librería *yfinance*, lo que permitió acceder a información relevante de estas compañías. Dado que las bases de datos no son de acceso público, el acceso a los datos se realiza a través de la API de Yahoo Finanzas, la cual recopila información de las bolsas de valores de distintos países, sin embargo, es importante destacar la responsabilidad en el manejo de estos datos, asegurando su correcta interpretación y uso.

Así también, se llevó a cabo la selección de las empresas BIMBO, CHEDRAUI, FEMSA, SORIANA y WALMEX que pertenecen al sector de bienes de consumo frecuente en México, estas entidades representan una parte significativa del mercado, ya que con base en el Financiero [17] aportaron una inversión en el año 2024 en promedio de 94 mil millones de pesos mexicanos, y son consideradas como centros de consumo listados en el mercado de capitales local con base en el método que se describe en la figura 1.

Es preciso señalar que, la información al ser pública las empresas son auditadas por el máximo organismo financiero en México que es la Secretaría de Hacienda y Crédito Público (SHCP), en este sentido la información recopilada se entiende verídica. Este sistema facilita la recopilación de datos históricos de cotización y otros indicadores financieros esenciales para su clasificación y estudio, tales como precio de cotización de apertura, bajo, alto y de cierre, también se determina el volumen de operaciones, así como la rentabilidad y volatilidad, se tiene un atributo de decisión binaria de sí o no de invertir; por último, la recolección de datos en promedio fue de 3700 por cada una de las cinco empresas analizadas.

Tabla 1. Resultados de clasificación correcta de los algoritmos.

	BIMBO (%)	CHEDRAUI (%)	FEMSA (%)	SORIANA (%)	WALMEX (%)
Árboles de decisión	99.9	99.9	99.9	100	100
Clasificador Simple de Bayes	94.5	95.7	92.7	91.2	96.5
K-vecinos más cercanos	92.8	95.7	91.1	90.4	95.9
MVS	98.3	97.9	97.7	91.4	98.8
Redes Neuronales	97.7	98.4	97.9	96.9	98.3

Fuente: Elaboración propia con base en los resultados por *weka*. En esta tabla se puede observar los resultados en porcentaje de las clasificaciones correctas de los algoritmos.

Etapa 2. En este paso se diseñó el experimento en el que cada algoritmo seleccionado se ejecutó cinco veces de forma independiente, modificando el *random seed* en cada una de éstas, el rendimiento final se determinó promediando los resultados de estas cinco ejecuciones; lo anterior utilizando un *cross validation fold* de 10, por lo que se realizaron 50 experimentaciones de cada algoritmo.

Etapa 3. Para este paso, se experimentó con los algoritmos seleccionados y los datos descargados en el paso 1 los cuales fueron procesados y analizados utilizando la herramienta WEKA, una plataforma especializada en minería de datos y aprendizaje automático. WEKA permite utilizar distintos algoritmos de clasificación y evaluación de modelos para identificar relaciones dentro de los datos financieros.

En este análisis, se emplearon los hiper parámetros por defecto de cada algoritmo, para obtener una primera aproximación a la clasificación de los datos de forma correcta e incorrecta, lo que permitió evaluar el comportamiento de distintos algoritmos sin la necesidad de realizar ajustes manuales previos; por último, este paso sirvió como punto de partida para identificar cuáles modelos podrían ser más efectivos en la predicción y segmentación de los datos de las empresas seleccionadas a través de los algoritmos de árboles de decisión, clasificador simple de Bayes, k-vecinos más cercanos, máquinas de soporte vectorial y red neuronal.

4. Resultados experimentales

Para realizar los experimentos se usaron datos de cinco empresas de las más representativas que cotizan en la Bolsa de Valores en México del mercado de capitales. La recolección de datos se hizo considerando el periodo del año 2010 al 2024, tomando en cuenta los atributos de los precios de cotización que son: precio de apertura, precio máximo, precio alto, precio bajo, precio de cierre, precio de cierre ajustado, rendimiento, volatilidad; y teniendo como valor de decisión para la clasificación si se invierte o no.

Tabla 2. Resultados de la mejor matriz de confusión para el algoritmo de árboles de decisiones.

Conjunto de datos	a	b
Walmex	5253	0
	0	1885

Fuente: Elaboración propia con base en los resultados por *weka.* En esta tabla se puede observar la clasificación del conjunto de datos de Walmex.

Tabla 3. Resultados de la mejor matriz de confusión para el algoritmo clasificador simple de Bayes.

Conjunto de datos	a	b
Walmex	1790	63
	66	1819

Fuente: Elaboración propia con base en los resultados por *weka.* En esta tabla se puede observar la clasificación del conjunto de datos de Walmex.

Tabla 4. Resultados de la mejor matriz de confusión para el algoritmo k vecinos más cercanos.

Conjunto de datos	a	b
Chedraui	1437	64
	62	1362

Fuente: Elaboración propia con base en los resultados por *weka.* En esta tabla se puede observar la clasificación del conjunto de datos de Chedraui.

Tabla 5. Resultados de la mejor matriz de confusión para el algoritmo red neuronal simple.

Conjunto de datos	a	b
Chedraui	1861	16
	41	1739

Fuente: Elaboración propia con base en los resultados por *weka.* En esta tabla se puede observar la clasificación del conjunto de datos de Chedraui.

Tabla 6. Resultados de la mejor matriz de confusión para el algoritmo de maquinas de soporte vectorial.

Conjunto de datos	a	b
Walmex	1840	13
	28	1857

Fuente: Elaboración propia con base en los resultados por *weka.* En esta tabla se puede observar la clasificación del conjunto de datos de Walmex.

La experimentación fue realizada usando el software *Weka* como ya se comentó en párrafos anteriores, utilizando las implementaciones de los algoritmos de árboles de decisión, clasificador simple de Bayes, máquinas de vectores de soporte, redes neuronales artificiales y k-vecinos más cercanos, y usando los valores de los hiper parámetros que están por defecto. Así, algunos de los valores usados fueron: para AD se tomó como factor de confianza de 0.25, para el CSB se consideró un tamaño de lote de 100, para k-nn se usaron 3 vecinos y la distancia Euclidiana, para MVS se utilizó un

Tabla 7. Resultados de las métricas precisión y media F.

Nombre del algoritmo	Precisión	Medida F
Árboles de decisión	1	1
Clasificador simple de Bayes	0.94332	0.9416
K-vecinos más cercanos	0.93248	0.93244
Máquinas de soporte vectorial	0.9702	0.96856
Red neuronal simple	0.97896	0.97868

Fuente: Elaboración propia con base en los resultados por *weka*. En esta tabla se puede observar el promedio de las métricas de precisión y medida F por cada algoritmo.

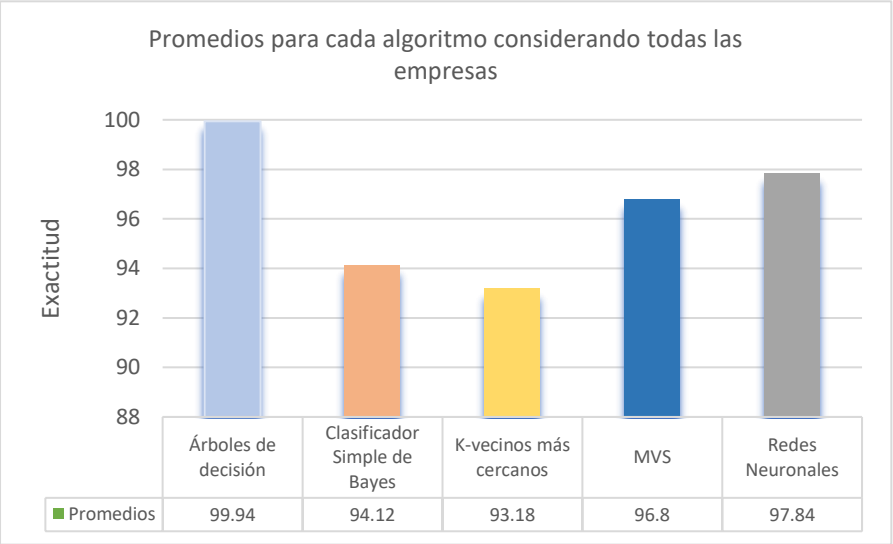


Fig. 2. La presente figura expone los promedios por empresa respecto al algoritmo con mejor desempeño que fue el de árbol de decisión.

kernel de función lineal, mientras que para las RNA se tuvo una tasa de aprendizaje de 0.3 y 500 épocas para el entrenamiento. Nuevamente es importante mencionar que al tratarse de los primeros experimentos, se ha decidido no modificar estos valores que están previamente establecidos.

En la Tabla 1 se muestra el promedio de 5 ejecuciones de los algoritmos, usando *10-fold cross-validation*. Se puede observar que el algoritmo de árboles de decisión obtuvo los mejores resultados para todas las empresas, logrando inclusive una exactitud del 100% para las empresas de Soriana y Walmex, y del 99.9% para Bimbo, Chedraui y Femsa. Los segundos mejores resultados fueron obtenidos con los algoritmos de máquinas de vectores de soporte y redes neuronales, con valores del 97% y 98% en algunos casos. Así también, se puede observar que el algoritmo que obtuvo los resultados más bajos fue k-vecinos más cercanos, teniendo un 90.4% para la empresa de Soriana.

En las tablas de la 2 a la 6 se muestran las mejores matrices de confusión de cada conjunto de datos, de los diferentes algoritmos implementados en *Weka*, siendo para los cinco casos la nomenclatura marcada de la siguiente manera: a una clasificación tipo no, y b una clasificación tipo sí.

En lo que respecta a las métricas de precisión y medida F, se tomaron en cuenta los promedios de los cinco conjuntos de datos de cada algoritmo que se aplicó en el software *weka*, obteniendo los resultados que se describen en la tabla 7.

Por último, como se puede observar en la figura 2, el algoritmo de árboles de decisión fue el que tuvo mejor desempeño con respecto a la clasificación de los datos de las cinco empresas; en el caso de las empresas Bimbo, Chedraui y Femsa, aunque no llegaron a tener una clasificación del 100% como los datos Soriana y Walmex, fue el algoritmo mejor evaluado con respecto a los demás.

5. Conclusiones

Los resultados obtenidos permiten concluir que el algoritmo de árboles de decisión demostró el mejor desempeño en la clasificación de los datos. Su efectividad radica en su capacidad inherente para modelar decisiones binarias de manera eficiente, lo que lo hace particularmente adecuado para el análisis bursátil. Esto concuerda con lo reportado en la literatura, específicamente en los trabajos de Sahrab, Kang y Jin 2021 [6].

Cabe señalar que, el conjunto de datos utilizado para cada empresa contiene 3,700 registros por lo que, se puede inferir que la cantidad de datos no limitó significativamente la capacidad de clasificación. La literatura sugiere que estos algoritmos pueden mejorar su rendimiento a medida que el volumen de datos aumenta, lo que abre la posibilidad de mejorar mediante el uso de conjuntos de datos más extensos.

Finalmente, el alto rendimiento del algoritmo de árboles de decisión, particularmente cuando se aproxima al 100%, podría ser indicativo de sobreajuste a los datos de entrenamiento. Por lo tanto, se recomienda realizar un ajuste de hiper parámetros, como la profundidad máxima del árbol para evaluar su desempeño.

Como trabajo a futuro, en primer lugar se considera ampliar el número de registros de otras empresas pertenecientes al sector V -bienes de consumo frecuente en México, en segundo lugar obtener los datos de otros sectores tal como el sector III- Industrial y el sector VIII que corresponde a Tecnología de la Información, para realizar los experimentos correspondientes como lo expuesto en el presente artículo; por último, se modificarán algunos valores de los hiperparametros de los algoritmos con el objetivo de mejorar los resultados obtenidos.

Referencias

1. Baca, G., Marcelino, M..: Ingeniería Financiera. México: Editorial Patria (2016)
2. García, V. M.: Análisis Financiero: Un enfoque integral. México: Editorial Patria (2015)
3. Raschka, S., Mirjalili, V.: Python Machine Learning. México: Marcombo (2019)
4. Henrique, B., Sobreiro, V., Kimura, H.: Uso de regresión vectorial de soporte (SVR) para la predicción de precios de acciones. Expert Systems with Applications, pp. 226–251 (2019)

5. Strader, T., Rozycki, J., Root, T., Huang, Y.: Revisión de estudios sobre la predicción del mercado de valores mediante aprendizaje automático. Journal of International Technology and Information Management Journal of Internatio, pp. 63–83 (2020) doi: 10.58729/1941-6679.1435.
6. Sahrab, M., Kang, Y., Jin, L.: Eficiencia de la IA en la predicción del mercado de valores mediante el aprendizaje automático. Electric Power Systems Research, pp. 1–11 (2021)
7. Coy, G., Granados, O., Garcia, O.: Predicción de la serie temporal del indicador bancario de referencia (ibr) con redes neuronales. Revista Mutis, pp. 65–76 (2021) doi: 10.21789/22561498.1748.
8. Cuevas, M., Alvares, S., Azcona, M.: Capacidad predictiva de las Máquinas de Soporte Vectorial. Una aplicación en la planeación financiera. Revista Cubana de Ciencias Informáticas, pp. 59–75 (2019)
9. Chhajer, P., Shah, M., Kshirsagar, A.: Theapplications of Artificial Neural Networks, Support Vector Machines, and Long–Short Termmemoryforstockmarketprediction. Decision Analytics Journal, pp. 1–12 (2022) doi: 10.1016/j.dajour.2021.100015.
10. García, M., Jalal, A., Garzón, L.: Métodos para predecir índices bursátiles. Ecos de Economía, pp. 51–82 (2023)
11. Lin, C., Lobo, J.: Stock Market Prediction Using Artificial Intelligence: A Systematic Review of Systematic Reviews. Social Sciences & Humanities Open, pp. 1–11 (2024) doi: 10.1016/j.ssaho.2024.100864.
12. Andrade-Gorjoux, L.E., González-Contreras, J.F., Montiel-Pérez, J.Y.: Comparación entre Codificación Predictiva Lineal (LPC) y Red Neuronal Artificial Long Short-Term Memory (LSTM) para la predicción financiera del Índice de Precios y Cotizaciones (S&P/BMV IPC). POLIBITS, 65(1), pp. 19–26 (2024)
13. Ricchiuti, F., Sperlí, G.: An Advisor Neural Network Framework Using LSTM-Based Informative Stock Analysis, Expert Systems with Applications (2024) doi: 10.1016/j.eswa.2024.125299.
14. Aritonang, P.K., Wiryono, S.K., Faturohman, T.: Hidden-Layer Configurations in Reinforcement Learning Models for Stock Portfolio Optimization, Intelligent Systems with Applications (2024) doi: 10.1016/j.iswa.2024.200467.
15. Banerjee, S.: Portfolio Management With the Help of AI: What Drives Retail Indian Investors to Robo-Advisor Electronic, Journal of Information Systems in Developing Countries, 91(1) (2025) doi: 10.1002/isd2.12346.
16. Ning, F., Jin, J., Zhang, L.: Risk Analysis of China's Financial Market Collapse Based on Cloud Computing and Machine Learning Algorithms, Expert Systems, 42(1) (2025) doi: 10.1111/exsy.13426.
17. El Financiero (2024) https://www.elfinanciero.com.mx/empresas/2024/03/26/empresas-como-bimbo-walmart-y-femsa-invertiran-184-mil-877-mdp/

ISSN 1870-4069

Responsible Use of AI as a Transversal Theme in an Interaction Design Course: A Report on Participation in the Instructional Co-Design Process

Scarlett Itzel Xochicale Flores[1], Angelica Rodríguez Vallejo[1], Soraia Silva Prietch[2], Josefina Guerrero García[1], Juan Manuel González Calleros[1]

[1] Benemérita Universidad Autónoma de Puebla, Facultad de Ciencias de la Computación, Mexico

[2] Benemérita Universidad Autónoma de Puebla, Facultad de Ciencias de la Electrónica, Mexico

{scarlett.xochicalef, angelica.rodriguezval}@alumno.buap.mx, sp223570639@alm.buap.mx, {josefina.guerrero, juanmanuel.gonzalez}@correo.buap.mx

Abstract. The goal with this paper is to present the experience of one of the participating teams in a study carried out with students of the "Interaction Design" course to identify the advantages and disadvantages of using an adapted version of the card sorting technique in the instructional co-design process of lesson plans that aim to combine course content with the discussion on the responsible use of AI tools in higher education. In many situations, the voices of undergraduate students are not heard during the curriculum or lesson plan design process. The intention of conducting the study with them was to give them the opportunity to design and co-create classes tailored to their needs and learning style. In order to achieve this goal, three workshop sessions were held to discuss the principles of responsible use of AI, collaborative lesson planning, and implementation and evaluation of a co-designed lesson plan. As a result, the codesign activities are instantiated with the perceptions of the team that agreed to share their participation experience. A list of takeaway recommendations that can be useful for this research group (or others) is presented as the main contribution of this paper.

Keywords: Codesigned lesson plans, card sorting, artificial intelligence in higher education, AIED.

1 Introduction

With the widespread use of generative Artificial Intelligence (AI) tools by students in higher education reported in different publications (Digital Education Council, 2024; Freeman, 2025), there is a need to promote discussions on their responsible use and to co-design lesson plans for courses in any discipline or major that include the "responsible use of AI tools in higher education" as a transversal subject, and for this

process to be carried out collaboratively with higher education (HE) actors (e.g., professors, students, industry professionals). The rationale for including students in the process is that they are considered "primary users" of the system referred to as education. As such, integrating the students into the instructional co-design process contributes to an equitable process for a group of HE actors whose voices, in many cases, remain unheard (Hidalgo and Perines, 2018). This work is part of a broader doctoral research project, in the context of which the study reported in this paper was conducted.

This larger project is supported by an epistemological framework that blends the following paradigms: Social Constructivism (Kim, 2001), Critical-Reflective (Freire, 1967, 2028) and Transformative (Creswell and Creswell, 2018), a theoretical framework that includes: Instructional Design (Branch, 2009), Active Learning Methods (Bonwell and Eison, 1991; Tharwat and Schenck, 2023), Co-design (Schuler and Namioka, 1993; Muller et al., 1997; Baranauskas et al., 2013), Artificial Intelligence in Education (AIED) and Principles of the Responsible Use of AI (Floridi and Cowls, 2019; UNESCO, 2021; Alam, 2023; Kosslyn, 2023; Nguyen et al., 2023; Aler-Tubella et al., 2024), a methodological framework guided by qualitative research methods (Flick, 2015; Creswell and Creswell, 2018), Critical Participatory Action Research (Kemmis, McTaggart and Nixon, 2014), and Instructional Co-design (Drajati et al., 2023; Barton and Fanshawe, 2024).

In our previous work, a systematic literature review was carried out and a related study has been conducted, both of which support the present study decisions (Prietch *et al*., 2024; Prietch, Guerrero and González, 2024). In this previous study a more diverse group participated, including faculty members, graduate students, an undergraduate student, and an industry professional. Some takeaway lessons included: (a) the possibility of power dynamics evidenced by the interaction between a faculty member and an undergraduate student, as well as follow-up interviews with participants, which motivated us to conduct separate workshop sessions involving groups with similar educational levels; (b) a more focused way to co-design a single class instead of an entire curriculum to include de use of generative AI tools and the discussion of the responsible use of AI in education; and, finally, (c) the possibility to co-design a transversal theme class contextualized into a real course, which could play an important role in the collaborative design of lessons in a more concrete way for participants.

The study reported in this paper refers to an intervention within the larger research project carried out with undergraduate students of the "Interaction Design" course in an information technology program at a Mexican public university. As stated by (Preece, Rogers and Sharp (2015, p. 23), "Interaction is the process by which two or more entities, whether people or systems, exchange information, influencing each other through various communication and response channels".

Both concepts are integrated to create the term interaction design, which focuses on facilitating intuitive, efficient, and satisfactory interaction between people and a product, a system, whether digital or physical, taking on singular importance. In this context, to Muñoz Arteaga et al., (2014, p. 32), "the goal of interaction design [is] to create systems that satisfy the needs of the people who use them, in a way that is spontaneous and satisfying.".

The overall objective of this particular study was to identify the advantages and disadvantages of using an adapted version of the card sorting technique[1] in the instructional co-design of lesson plans that combined course's content with the transversal theme (the responsible use of AI tools in higher education). Specifically, this paper presents the experience of one of the student teams participating in the study, called "Pancitos" (and in a later session renamed as "Umizumi").

This paper is organized into the following sections: Sections 2 and 3 summarize salient references related to our study, Section 4 describes the methodology used, Section 5 reports the lesson plan co-designed by the "Pancitos" team, Section 6 presents the discussion of the results, and finally, Section 7 presents the conclusions.

2 Principles for the Responsible Use of AI

With the intention of discussing real-world cases from the perspective of humans (users), this study took into consideration nine principles of responsible use of AI systems (for easier reference, "AI" is used in the paper). This section presents definitions of the principles identified in specialized references (Floridi and Cowls, 2019; UNESCO, 2021; Alam, 2023; Nguyen et al., 2023; Aler-Tubella et al., 2024): (1) Human-centered AI, (2) Social and environmental well-being, (3) Bioethics and ethics, (4) Justice, (5) Explainability, (6) Privacy, (7) Accountability, (8) Security and (9) Transparency.

Principle 1 (Human-centered AI) concerns placing humans in the loop, being in control of decisions, responsible for the design, use, evaluation, tracking and impact assessment of AI. Principle 2 (Social and environmental well-being) states that critical thinking should be applied in the way to avoid negative impacts on society, the environment, or the global economy. Principle 3 (Bioethics and ethics) refers to four basic aspects: non-maleficence (not causing harm to others); beneficence (promoting and doing good); autonomy (respecting people's self-determination); and justice (seeking and promoting equitable access). The definition of Principle 4 (Justice) indicates that existing biases may be reinforced and even magnified in society if concerns about justice and equity in the way we interact with AI are not taken into consideration. Principle 5 (Explainability) seeks to improve the interpretability and transparency of AI; in this sense humans should be aware of and understand their responses to be in control of their own actions. Principle 6 (Privacy) indicates the awareness of informed users consent to make decisions on how to protect their personal or professional data. Principle 7 (Accountability) is about clearly stating each stakeholder's acknowledgement and responsibility for their actions. Principle 8 (Security) states that generative AI must be used with safety in mind, considering oneself and others; they should not be utilized to harm or endanger people. Finally, Principle 9 (Transparency) states that humans should be transparent in their actions, for instance, to disclose what and how generative AI responses were used in academic

[1] The *card sorting* technique allows for gathering perspectives on the mental models of diverse participants and can be applied in three different ways: open, close and hybrid card sorting (Preece, Rogers and Sharp, 2015). In this study, the closed method was used in the lesson plan template (printed in a poster format), since its fields were previously defined, and the three stack of cards provided to the participants included concepts related to each of them.

work. These nine principles take us to a broader discussion on academic integrity (Yusuf et al., 2024) that go beyond plagiarism issues; however, this discussion is not addressed in this paper.

3 GenAI in Education

The use of generative artificial intelligence (GenAI) has escalated in recent years, particularly in higher education, generating a broad debate within the academic community. Thus, for instance, Sánchez and Carbajal (2023) emphasize the need to design appropriate pedagogical strategies for their integration into the classroom and warns about the potential risks associated with the use of these tools. Mollick and Mollick (2023) analyzed how AI can facilitate the implementation of evidence-based teaching strategies using tools that implement large-scale language models (LLMs). In addition, their work demonstrates how ChatGPT can support the rapid and personalized generation of teaching materials, as well as of evaluations and explanations adapted to diverse students.

Moreover, the integration of participatory methodologies into teaching has proven effective in improving classroom learning. This is the case of card sorting, which is used to organize information and design interactive activities. In this area, López and Albar (2024) analyzed and concluded that card sorting in the classroom increases the motivation and participation of students, allowing them to be active during their learning process. Likewise, this technique facilitates information structuring, as students organize concepts in a logical way. Bivens and Welhausen (2021) highlighted that the use of open card sorting serves as a cognitive support strategy that helps students develop analytical skills. Also, the authors structured learning activities into different levels of complexity using Bloom's Taxonomy.

The studies described throughout this section provide the theoretical and methodological framework that supports the approach used in the work presented in this paper. As highlighted by Sánchez and Carbajal (2023), and Mollick and Mollick (2023), the use of AI tools in education has been emphasized, while López and Albar (2024), and Bivens and Welhausen (2021) demonstrate that card sorting can improve the active structuring of learning content. These references reinforce the approaches used in the study reported in this paper, since instructional co-design allowed students to develop progressive skills, motivating them to organize content in a reflective way.

4 Methodology

The study was conducted in the “Interaction Design” course with the collaboration of a professor and a doctoral researcher. It took place in February 2025 in three 90-minute workshop sessions, totaling 4.5 hours. The group consists of 30 enrolled students; however, a few did not participate in all sessions (26 students in sessions 1 and 2; 25 in Session 3).

The sessions were organized as follows. In Session 1, topics related to the responsible use of AI were presented and discussed. The materials used included slides, a brainstorming poster, sticky notes, and colored markers to encourage oral and written

participation. In Session 2, the hands-on activity session of instructional co-design of lesson plans was conducted using an adapted card sorting technique. For this session, posters were distributed with a large template for co-design[2] with areas to be occupied by cards, an envelope with three stacks of cards[3], as well as sticky notes and markers. The co-design template included five areas (fields) to be filled out: (1) Student profile (considering team data), (2) Principles of the responsible use of AI, (3) Course Content, (4) Objective and Methodology, and (5) 7-Step Class Activity Details. The envelopes included sets of cards for Field 2 (with 9 cards), Field 3 (with 5 cards), and Field 4 (with 13 cards) of the template. Fields 1 and 5 were designed to be filled out using sticky notes. Field 5 organized the class moments according to Bloom's Taxonomy to define activities and its learning goals. Each stack of cards had different colors matching the fields in the poster template. The cards were designed with the following information: section title and logo, topic title (thematic unit, principle of the responsible use of AI, or objective and methodology number), descriptive content according to its title, and references.

At the end of Session 2, a vote was held to choose which team would implement the co-designed lesson at the next session. A team named "Justice for AI" received the most votes. In Session 3, the "Justice for AI" team implemented and evaluated their co-designed lesson, and members of the "Pancitos" (now "Umizumi") team and other teams participated as peer students. The lesson co-designed by the team with the most votes required some adjustments before implementation, which was done in collaboration with the session moderator during the period between Session 2 and Session 3.

As previously mentioned, this paper reports on the experience of one of the participating teams during the study. The following section provides details of this experience.

5 A Team's Experience Report

The results of the three sessions discussed in Section 4 are described in the following subsections.

5.1. Session 1: Participation in the Initial Discussion

As mentioned in the methodology section, this first session focused on the presentation and discussion of the responsible use of artificial intelligence (AI). From the perspective of some students, the session was informative and enriching, as it allowed them to reflect on a topic that, while used in their daily lives to optimize time, they had not yet analyzed in depth.

[2] Poster - Template for the instructional co-design of lesson plans, https://bit.ly/4iBLA4O

[3] Three stacks of cards, https://bit.ly/3DHUij5

(a) (b) (c)

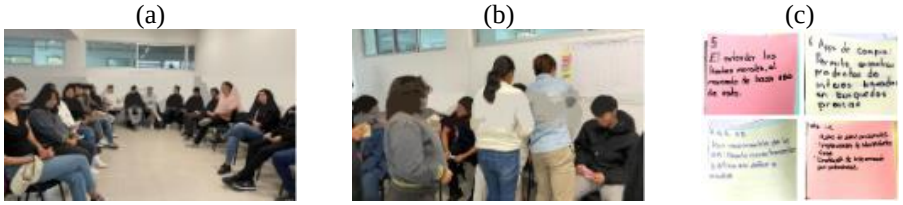


Fig. 1. Pictures of (a) group discussion, (b) opinions posting and (c) sample reflections.

One of the most valuable aspects was the distinction between responsible use of AI and the concept of responsible AI, which led them to consider its implications in various fields, especially in education and professional settings. During this session, student participation focused on active listening and discussion, paying attention to the ideas shared by their classmates and the moderator. Subsequently, the students contributed to the brainstorming session, recording reflections on sticky notes and posting them on the collective poster. Figure 1 illustrates the group discussion (a), sticky notes posting (b), and some sample reflections (c).

As a result of the question "What does 'responsible use of AI' mean to you?", 24 responses were identified on the sticky notes, among which the following reflections were presented: "Understanding the moral limits when using it [AI]"; "Using it [AI] correctly and ethically without harming anyone"; "Knowing how to use AI with a responsible objective, i.e., ethical, academic, etc. With a good purpose and not excessively because it wastes/occupies a lot of resources." In the responses, a frequent association between responsible use and ethical use is noted, with at least thirteen sticky notes including the prefix "ethic*" or related words.

For the question "What are the positive or negative aspects of using predictive AI in social media or other web-based systems?", 28 responses were obtained on sticky notes. Here are some of the reflections: "Shopping apps: Allow you to find products of interest based on previous searches"; "Polarization of ideas, which prevents dialogue and discussion of opinions"; "They increase screen time, using user preference algorithms"; "Theft of personal data, identity theft, restriction of information based on preferences". Different types of examples were cited, evidencing that students are aware of a variety of web-based systems that may embed predictive AI algorithms and some of their implications.

This space for dialogue and analysis allowed for the expression of different perspectives, with the discussion of the ethical implications and challenges of integrating AI into academic training and professional futures being particularly relevant. In this sense, it is essential to establish clear principles that regulate its appropriate use, ensuring a positive impact on education and professional practice.

5.2. Session 2: Co-designed Lesson Plan

In this second session, a team of four women worked together to develop an instructional co-design lesson plan using an adapted version of the card sorting technique. The teamwork strategy was carried out collaboratively, beginning with a

(a) (b) (c)

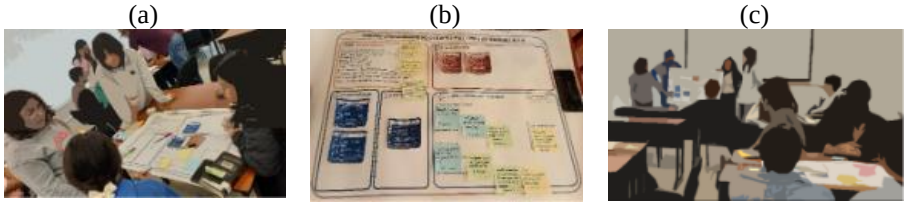


Fig. 2. Pictures of the: (a) team in the co-design process, (b) co-designed lesson plan, and (c) "Pancitos" team presentation

general exploration of the materials provided, particularly in reading and selecting from the card sets.

For the first field of the template, called "Student Profile", general team information was shared, including an average age of 22. Additionally, a brief internal survey was conducted to answer questions about the use of generative AI, identifying the most used tools within the team: ChatGPT, DeepSeek, Gemini, Copilot, and Meta, with a usage frequency of five business days per week. A vote was then held to describe prior knowledge about the responsible use of AI from the individual team perspective, highlighting its misuse for legal purposes, environmental impact, transparency, and ethics.

In the next field of the template, "Principles of the Responsible Use of AI", a new vote was held to select the two most important principles for our proposal; in this case, the following were chosen: (1) Human-Centered AI and (2) Responsibility. The cards were moderately supportive in the choice of principles, since, although they served as a reminder of Session 1, in the end the group discussion influenced decision making.

For the "Course Content" field, each member received a card corresponding to a thematic unit, read it in depth, and shared their opinions with the team. Based on these discussions, a vote was held again to discard the topics that interested us the least, leaving "Unit 3: Interaction Design Process". Within this unit, the topic "3.5 Usability and User Experience" was selected, as everyone agreed that it would be new and interesting content to teach in class if the team were selected to lead Session 3. It is worth mentioning that the group had access to the course syllabus at any time. This syllabus is on the official website of the academic secretariat. Complementarily, lecturers must present their course syllabus at the beginning of each semester, as part of university protocol. From our point of view, we considered important to remember the topics of the course through the cards. In this sense, the cards "Course Content" were of great support, since they facilitated the identification and comparison of their contents, allowing for a more structured and informed decision making.

In the "Objective and Methodology" field, a similar process was followed: each member received a card with different pedagogical strategies, and, after a vote, it was decided to combine two approaches based on the theory of behaviorism: "*Positive reinforcement*: to shape behaviors" and "*Programmed teaching*: structured learning sequence with immediate feedback". The main reason for choosing behaviorism was the description on the cards, given that when all participants read that the pedagogical strategy was based on positive reinforcement, and its operationalization in gamification, the team understood that people are motivated by winning or being rewarded for an action.

Figure 2 shows photos of the team during the co-design process, the co-designed lesson plan, and the presentation of the co-designed lesson to the group members.

For the last field, "7-Step Class Activity Details", each team member conducted a ChatGPT consultation using descriptive prompts. The team members then wrote their responses individually on sticky notes and held a vote to select the overall seven best steps to be used in the co-designed class. The selected prompt was: *"Hello, please help me create an hour and a half class on Usability and User Experience, while also covering two other topics: Human-Centered AI and Accountability in the Use of AI with Behaviorism Methodologies: Positive Reinforcement and Programmed Teaching, complying with the 7 levels of Bloom's Taxonomy, distributed over specific times"*.

The full result from ChatGPT for the prompt can be found at https://bit.ly/3DLgRUe.

However, due to the limited space for notes, the result was summarized. In Step 1, "Quick knowledge questions" (5 minutes) activate initial interest and prepare students for the content that will be seen during the class. In Step 2, with the "Explanation of concepts (examples)" (10 minutes), it facilitates the understanding of the information at the time of exemplification, thus ensuring that students understand the basics before applying them. As observed in Step 3, "Analysis of a given interface and proposal for improvement" (15 minutes), which includes an analysis of an interface and its improvements, this allows us to put into practice the concepts previously learned, thus reinforcing the connection between theory and practice. In Step 4, "Analyze AI cases and their impact on UX" (20 minutes), critical thinking is promoted to evaluate how AI influences the user experience, and to improve the discussion among students, positive reinforcement [*giving away candies*] is implemented to reward interventions or moments where students contribute well-founded ideas. In Step 5, "Compare cases of good and bad UX design" (15 minutes), it is proposed that students perform a structured comparison of real examples (experiences) that help to consolidate the topic. In this way it is observed that the programmed teaching allows students to identify design patterns through the logical sequence of analysis with the examples of the previous stages. Subsequently, for Step 6 –"Conceptual design of an interface with UX principles and AI ethics" (15 minutes) and "Create/Design a UX proposal with responsible AI" (15 minutes)-, as in Step 4, we proposed to implement positive reinforcement to encourage creativity and initiative for students to design innovative solutions, when assigning them a task of developing a proposal that is focused on accessibility and ethics. As for positive reinforcement, we relate it more to stage 7, "Reflect on responsibility in UX design with AI + Questions (Kahoot)" (10 minutes), where a Kahoot quiz is provided to evaluate learning in a dynamic way, in addition to creating an environment of participation, motivation and interest by rewarding interaction as well as student performance. Additionally, the programmed teaching provides immediate feedback on the answers, reinforcing the knowledge acquired in previous stages. Finally, the lesson co-design was presented to the rest of the group.

5.3. Session 3: Participation in the classmates' co-designed class

In the third session, the "Justice for AI" team implemented and evaluated the lesson they co-designed, while the "Pancitos" (now "Umizumi") team members and other members of the group took on the role of students. The group's participation was characterized by an atmosphere of respect and collaboration, as the presentation team

(a) (b)

Fig. 3. Pictures of the: (a) teams presenting their interface proposals and (b) interface proposal of the "Umizumi" team.

consistently emphasized the use of illustrative examples to facilitate understanding of the content on the topic of "Types of Interaction".

As the next step, the team proposed the development of an inclusive ATM powered by AI, designed to serve people with various disabilities. In this way, the ATM included a soundproof booth that provides privacy and security, incorporating an AI voice assistant that personalizes the interface according to the user's needs. Among its main features of AI voice assistance, it is based on asking the user about specific needs and thus being able to adjust its interface in response to the user's needs. Furthermore, the team proposed implementing Braille buttons to facilitate navigation for people with visual impairments. The proposal also included automatic screen adjustment, which adjusts the device's height based on the user's height or wheelchair users. Finally, an on-screen sign language interpreter was added, allowing interaction for people who are Deaf. Thus, the proposal integrated the principles of accessibility and inclusion using AI, aligned with the objectives of the practice and highlighting the importance of designing technologies that correspond to the interaction needs of diverse users.

At the end of the class, three teams had the opportunity to present their proposals and receive feedback from the lecturers. As for the perspective as a team, we think we did not necessarily consider presenting our work because we noticed that, among the first teams, the ideas explained in their proposals did not differ much in terms of our own proposal. At the time we noticed that the team responsible for the class first provided an example that covered all the points that were requested in the activity.

Although this session was perceived as well structured following the instructional co-design steps, "Umizumi" team considered that, due to the time required for presentations, thorough feedback was missing to all teams. Feedback was provided by the professor and the moderator, exploring ideas that could be improved or further investigated such as privacy, security, human factors and user experience.

5.4. Sessions' Evaluation

After we ran the three workshop sessions, an online questionnaire was made available to students (participants). Fourteen out of 26 participants responded to the closed and open questions regarding self-evaluation and workshop sessions evaluation. All respondents participated in the three sessions and indicated their responses as follows:

— *My greatest strengths in the sessions were:* attentiveness, willingness to participate, active listening and participating in small team projects, prior knowledge of AI, team communication, contributing with ideas during team activities, active and inclusive participation, patience and detailed explanations, critical thinking and understanding and analyzing the themes.

— *I am aware that I can improve in the following personal aspects in future similar activities:* punctuality, previous knowledge of AI, actively participating in group activities, clarifying ideas and sharing them, improving the ability to write texts, managing time, creativity, empathy and learning about ethics applied to AI use.

— *In the activities developed in the three sessions I learned:* Types of interaction (1), Learning techniques (1), Proper use of AI to benefit the local community and systems' design (2), Methods of working and analyzing information (2), Teamwork (2), Application to various areas (3), Methodologies for class design (3), IA and its importance and risks (3), and The responsible use of IA and its contextualized principles, and awareness of these aspects (9).

— *Session 1 - Content and Duration:* the content was very good (12), ideal duration (7), duration was not enough (5), and the content was good, and the duration was sufficient (2). Participants added that it was entertaining, gained a lot of experience and learning, easy to understand, interactive, "There were several points about AI that were missing; I think it only covered half of it" (P8), and "It was the session I enjoyed the most because we discussed the use of AI and its ethical implications" (P9). Overall, many respondents expressed that more time should be spent on this activity to calmly discuss the topics and to share everyone's ideas.

— *Session 2 - Content and Duration:* the content was very good (10) with ideal duration (8) and duration not sufficient (2), the content was good and the duration sufficient (3), and neither one was well planned (1). Some participants added: "I liked the dynamics" (P1), "[…] we didn't have time to produce our ideal results" (P7), "[…] we were able to see how we can apply AI to develop a lesson" (P9) and "Some content was too complex for our educational level" (P13).

— *Session 2 - Please provide your critical feedback on the printed materials used and their instructions for use:* some of the comments were: "the possibility of choosing the methodology gave greater freedom" (P6), "I really liked […] they can be shared with team members" (P7), "[…] I found it very convenient to work with cards and not have to think and search what to include in each part of the preparation; we just chose the various options" (P9), "A little confusing in some concepts, however with the support of the moderator it was easy to do it" (P11), very didactic (P12, P13), quality materials, not too expensive and reusable (P14).

— *Session 3 - Content and Duration:* overall the content was considered good or very good (13), and the duration was considered ideal or sufficient (11). Two comments that express some of the respondents' perceptions were: "I think they spent too much time on the introduction and not enough time on the activity, which didn't allow us to really come up with a good proposal" (P6), and "It was good because we were able to observe the implementation of all the steps outlined in our colleagues' previous activity" (P9).

6 Discussion

The instructional co-design process of the lesson plan, using the adapted card sorting technique, allowed for collaborative structuring and organization of class content. This process encouraged active participation among team members who participated in the sessions, in addition to developing the students' critical thinking. Specifically, Session 2 focused on the classification and selection of elements for the lesson plan, proved to be a practical and interactive way to facilitate decision-making regarding content, methodologies and objectives among the wide range of student opinions. Some of the students evaluated these contents are too complex for their educational level, they needed more time to discuss them, or further explanation could clarify.

Concerning the instructional co-design template given to students, in a positive way, the template composes a valuable tool for educational planning, as it allows structuring in an organized way the specific contents of interest, in the context of integrating the "Principles of the Responsible Use of AI" as a transversal theme into the "Interaction Design" course. Its design facilitates the creation of a detailed outline that guides the development of the class session, promoting a clear and efficient methodological approach. The use of this template has multiple advantages, firstly, it provides a clear, delimited, organized and systematic structure, in specific fields allowing a better structuring. It also encourages collaboration between teams by providing several options, which at the same time leave small room for participants to move away from the central themes, promoting attention maintenance, reflection and critical analysis.

Moreover, after participating in the class taught by their classmates, the "Umizumi" team considered improvements to the co-designed lesson plan, specifically in the "7-Step Class Activity Details" section, including better-structured activities to measure meaningful student learning. Furthermore, they identified the importance of conducting pilot tests to anticipate potential difficulties and thus improve their learning experience.

6.1. Takeaway recommendations

From this study, we have structured a list of recommendations from which we can move forward in our broader research project and provide some takeaways to other researchers who work on related topics:

— *Responsible use of AI as a Transversal theme in Higher Education (HE) Courses.* In this study, the instructional co-design of lesson plans was meant to integrate the transversal theme into the Interaction Design course; however, as GenAI is being used by many, with no distinction about their occupation, any other course in any knowledge field can benefit from this discussion. Also, this instructional co-design process can be an opportunity for HE actors to introduce social themes in computing or engineering career courses (Garrett *et al*., 2020).

— *Real-life examples considering a set of the principles of responsible use of AI*. These are useful for showing what is currently happening around the world regarding the topic, for uncovering critical thinking on positive and negative aspects considering presented examples, and for providing a previously (but recently) reflected transversal topic for participants to be able to integrate it into a main course content unit. It is recommended that the examples presented for

discussion are of interest to participants, of their life contexts to generate a sense of relevance and critical thinking among the group (Freire, 1967, 2018). Furthermore, according to students' evaluation, the size of the group should be considered so everyone can disclose their ideas and discuss them in depth with no rush.

— *Written and verbal opinions*. Allowing participants to provide written and verbal opinions contribute to maintaining a safe environment for those who are shy or do not want to express their opinions publicly. Also, the use of illustrative examples on a slideshow, a brainstorming poster, sticky notes and colored markers can function as a visual, auditory and motor stimulation activity.

— *Template improvements (printed poster for the instructional co-design).* A wider design area for the "Student Profile" field is needed, as multiple sticky notes can be visually overloading and can obstruct information, hindering quick reading comprehension. Some fields may require descriptions so as to facilitate the understanding of the students who participate, to avoid delays due to clarifications, or so students can optimize work by clarifying their main questions. A guide would be helpful in completing the fields where small descriptions of the expected answer format are placed in each field of the template.

— *Adapted card sorting integrated into the instructional codesign process*. The closed card sorting technique was chosen to provide pre-defined cards to participants of specific fields on the template. This choice was made considering the potential need to retrieve information or to introduce new concepts. The principles of the responsible use of AI were topics of a recent discussion and course content was disclosed to students at the beginning of the semester; however, a quick reminder of the topics could help in recalling them. Learning theories, pedagogical strategies and their objectives are not information technology course contents, so those cards can be new information to support choices on these matters. From some students' perspective choosing options from cards can be easier than searching for new information, and cards were considered of quality, very didactic, and reusable.

— *Learning taxonomy as a new stack of cards.* During Session 2 we realized that we could also have provided a set of learning taxonomy cards, not only Bloom's Taxonomy, but providing others such as Marzano's, Dave's, Fink's (BUAP, 2020), and letting participants decide freely on activities according to their chosen learning goals, instead of deciding for them (a 7-step class). This can be applied to other theoretical frameworks researchers wish to include in the instructional co-design process (e.g., topics related to informal training content).

— *Human-Human Collaboration and Human-AI Collaboration*. Human-AI collaboration was planned to be used in all fields if teams so wished. The intention was that students could have the experience of responsibly using GenAI in education and could gather new information for wild cards and generate ideas to organize learning activities according Bloom's Taxonomy. Some students' positive perception was that towards the use of IA to create lessons, meaning that a guided learning activity using AI in the classroom can support students to be

aware of responsible ways to do it. We observed that both types of collaboration can enrich learning experiences if appropriately planned and moderated.

— *Time and criteria for evaluating co-designed lessons plans*. The only co-designed lesson plan fully and deeply evaluated between sessions 2 and 3 was the most voted one ("Justice for AI" team) since they had the opportunity to implement it in a real classroom setting. The other teams did not receive specific feedback because of time constraints. This is an important activity; one alternative could be a shorter class co-design to give all the same opportunity to implement their proposals, especially if participants rely too much on GenAI's suggestions and they do not have much experience of evaluating whether the activities' durations are appropriate. Also, this alternative might lead to biased voting for teams with popular students. The same recommendation is valid for presenting activity results during the implemented co-designed class, which in Session 3 fell short in time and criteria for a small number of teams who did not have the opportunity to present their proposals.

— *Additional suggestions for equitable strategies to gather feedback*. For an equitable strategy where all teams would have the opportunity to present their proposals, we present two ideas: the first would be an art gallery style, where each team places the proposed interface in a shared space, resembling posters on a wall, slides or interactive whiteboards (e.g., Miro), so that later students can walk through the gallery and peruse the other teams' ideas, and leave comments on each of the proposals. The second alternative is to brainstorm ideas, with each team presenting one or more key points of their proposal, i.e. a distinctive or innovative aspect of their design. Similarly, sticky notes can be used to paste them on the blackboard or a collaborative digital tool to visualize and organize ideas. Afterwards, an open discussion could be held to compile the ideas, perhaps grouping them by themes and discussing the best (innovative or interesting) elements of each solution.

7 Conclusions

We posit that the objective of the study - *to test the use of an adapted version of the card sorting technique in the instructional co-design process of lesson plans that combined subject content with the discussion on the responsible use of AI tools in higher education* - was achieved, since participants were able to carry out, conclude and present their co-designed lesson plans, and at least one team had the opportunity to implement and evaluate their proposal.

As a positive point, the set of delivered artifacts (poster, cards, and sticky notes) was easy to understand and encouraged collaboration among participants. On the downside, the co-design time was considered short, and the results (co-designed lesson plans) consisted only of posters, cards, and sticky notes. For a reusable and implementable lesson plan, these would need to be converted into text in a more descriptive and detailed manner. This became apparent when the "Justice for AI" team needed to design the slides and activities to implement their co-designed lesson plan. The moderator raised many questions during their online meetings that remained unanswered at the conclusion of the co-design. To answer these questions and increase the chances of

successful implementation of the co-designed lesson, extra adjustment sessions were held.

The experience of the "Pancitos" team presented in this paper provides a snapshot of the activities carried out during the three workshop sessions, including photos and details about their perceptions. We understand that including the point of view of a single team may be considered a limitation of our work. It should be noted, however, that other student teams produced interesting results. The paper focuses on the impressions of the one team that was eager to collaborate in sharing and writing about their experience when the invitation was made to all participants. Fourteen participants collaborated also with their responses to a final evaluation questionnaire.

Acknowledgments. We thank SECIHTI (Secretaria de Ciencia, Humanidades, Tecnología e Innovación) for the doctoral scholarship of Soraia Prietch. We appreciate the collaboration with the professor (responsible for the Interaction Design course) and the participation of all students involved in this study. We also would like to thank the anonymous reviewers whose insightful comments greatly helped to improve an earlier version of this paper.

References

1. Alam, A.: Developing a Curriculum for Ethical and Responsible AI: A University Course on Safety, Fairness, Privacy, and Ethics to Prepare Next Generation of AI Professionals. In: Rajakumar, G., Du, KL., Rocha, Á.: Intelligent Communication Technologies and Virtual Mobile Networks. Lecture Notes, 171 (2023)
2. Aler Tubella, A., Mora-Cantallops, M. Nieves, J.C.: How to Teach Responsible AI in Higher Education: Challenges and Opportunities. Ethics Inf Technol, 26(3) (2024) doi: 10.1007/s10676-023-09733-7.
3. Baranauskas, M.C.C., Martins, M.C., Valente, J.A: Codesign de Redes Digitais: Tecnologia e Educação a Serviço da Inclusão Social. Penso, Porto Alegre (2013)
4. Barton, G., Fanshawe, M.: The LAB School Project: A Socio-Ecological Investigation Into the Intersection Between Literacy, The Arts and Wellbeing in a Rural Early Years Classroom Setting. The Australian Journal of Language and Literacy, 47(3), pp. 403–426 (2024) doi: 10.1007/s44020-024-00070-w.
5. Bivens, K.M., Welhausen, C.A.: Using a Hybrid Card Sorting-Affinity Diagramming Method to Teach Content Analysis. Communication Design Quarterly, 9(3), pp. 4–13 (2021) doi: 10.1145/3468859.3468860.
6. Bonwell, C., Eison, J.: Active Learning: Creating Excitement in the Classroom. AEHE-ERIC Higher Education Report. No.1. ED340272. Washington, DC: Jossey-Bass (1991)
7. Branch, R.M.: Instructional Design: The ADDIE Approach. Textbook, Springer, eBook (2009) ISBN: 978-0-387-09506-6.
8. BUAP - Benemérita Universidad Autónoma de Puebla.: Módulo II. Taxonomías del aprendizaje. Diplomado: Evaluación de los Aprendizajes (2020)
9. Creswell, J.W., Creswell, J.D.: Research Design: Qualitative, Quantitative, and Mixed Methods Approaches. Sage publications, 5th ed (2018)
10. Digital Education Council.: Digital Education Council, 2024 https://www.digitaleducationcouncil.com/post/digital-education-council-global-ai-student-survey-2024.
11. Drajati, N.A., So, H.J., Rakerda, H.: Exploring the Impact of TPACK-Based Teacher Professional Development (TPD) Program on EFL Teachers' TPACK Confidence and

Beliefs. Journal of Asia TEFL, 20(2), pp. 300–315 (2023) doi: 10.18823/ASIATEFL.2023.20.2.5.300.
12. Flick, U.: El diseño de investigación cualitativa. Ediciones Morata SL (2015) doi: 10.23935/2016/01018.
13. Floridi, L., Cowls, J.: A Unified Framework of Five Principles for AI in Society. Harvard Data Science Review, 1(1) (2019) doi: https://doi.org/10.1162/99608f92.8cd550d1.
14. Freeman, J.: Student Generative AI Survey 2025. HEPI. Homepage. http://bit.ly/3Ynh72V (2025)
15. Freire, P.: Educação como prática da liberdade. Rio de Janeiro: Paz e Terra (1967)
16. Freire, P.: Pedagogia do oprimido. Rio de Janeiro: Paz e Terra (2018)
17. Garrett, N., Beard, N., Fiesler, C.: More Than "If Time Allows" the Role of Ethics in AI Education. In Proceedings of the AAAI/ACM. In: Conference on AI, Ethics, and Society, pp. 272–278 (2020) doi: 10.1145/3375627.3375868.
18. Hidalgo, N., Perines, H.: Dar voz a los protagonistas: La participación estudiantil en el proceso de enseñanza-aprendizaje. Revista Educación, pp. 438–464 (2018) doi: 10.15517/revedu.v42i2.27567.
19. Kim, B.: Social Constructivism. In: Orey, M.: Emerging Perspectives on Learning, Teaching, and Technology, pp. 1–8 (2001)
20. Kosslyn, S.M.: Active Learning with AI: A Practical Guide. Alinea Learning (2023)
21. López, L.M., Albar, J.M.: El Card Sorting como metodología de diseño experiencial en el aprendizaje de la Programación didáctica con alumnos de Bellas Artes. Dialnet (2024)
22. Mollick, E.R., Mollick, L.: Using AI to Implement Effective Teaching Strategies in Classrooms: Five Strategies, Including Prompts. SSRN Electronic Journal (2023) doi: 10.2139/ssrn.4391243.
23. Muller, M. J., Haslwanter, J. H., Dayton, T.: Participatory Practices in the Software Lifecycle. In: Helander, M.G., Landauer, T.K., Prabhu, P.V.: Handbook of Human-Computer Interaction, 2nd ed., pp. 255–297 (1997) doi: 10.1016/B978-044481862-1/50077-7.
24. Muñoz Arteaga, J., Céspedes Hernández, D.: Temas de diseño en Interacción Humano-Computadora. Iniciativa Latinoamericana de Libros de Texto Abiertos (LATIn) (2014)
25. Nguyen, A., Ngo, H.N., Hong, Y.: Ethical Principles for Artificial Intelligence in Education. Educ Inf Technol, 28, pp. 4221–4241 (2023) doi: 10.1007/s10639-022-11316-w.
26. Preece, J., Rogers, Y., Sharp, H.: Interaction design: Beyond Human-Computer Interaction, 4th ed., Wiley (2015)
27. Prietch, S.S., Aguilar-González, G., Cordero-Cid, L.A.: Interdisciplinary Co-Design Process of Instructional Lesson Plans for Promoting the Responsible Use of AI. Avances En Interacción Humano-Computadora, 9(1), pp. 223–228 (2024) doi: 10.47756/aihc.y9i1.172.
28. Prietch, S.S., Guerrero, J.G., González, J.M.C.: Una investigación acción participativa crítica para promover el uso responsable de herramientas de IA con actores de la educación superior. In: Cervantes, E.E.V., García, J.G., Rangel, Y.N.: Metodologías para el aprendizaje y la inteligencia artificial . United Academic Journals, pp. 136–152 (2024)
29. Sánchez Mendiola, M.S., Carbajal Degante, E.C.: La inteligencia artificial generativa y la educación universitaria. Perfiles Educativos, 45(Especial), pp. 70–86 (2023) doi: 10.22201/iisue.24486167e.2023.Especial.61692.
30. Schuler, D., Namioka, A.: Participatory Design: Principles and Practices. Lawrence Erlbaum Associates (1993) doi: 10.1201/9780203744338.
31. Tharwat, A., Schenck, W.: A Survey on Active Learning: State-of-the-Art, Practical Challenges and Research Directions. Mathematics, 11(4), 820 (2023) doi: 10.3390/math11040820.
32. UNESCO, D.: Recomendación sobre la ética de la inteligencia artificial (2021) https://unesdoc.unesco.org/ark:/48223/pf0000381137_spa.
33. Yusuf, A., Pervin, N., Román-González, M.: Generative AI and the Future of Higher Education: A Threat to Academic Integrity or Reformation Evidence from Multicultural

Perspectives. International Journal of Educational Technology in Higher Education, 21(21) (2024) doi: 10.1186/s41239-024-00453-6.

ISSN 1870-4069

Optimización de la Arquitectura Pix2Pix: Un Estudio de Reducción de Capas y Calidad de Imagen

Alexander Tapia Cortez, José Alejandro Tejeda Sánchez, Yolanda Moyao Martínez, David Eduardo Pinto Avendaño, José de Jesús Lavalle Martínez

Benemérita Universidad Autónoma de Puebla, México

{alexander.tapiaco, jose.tejedasanc}@alumno.buap.mx, {yolanda.moyao, david.pinto, jose.lavalle}@correo.buap.mx

Resumen. En este trabajo en proceso se explora la modificación de la arquitectura Pix2Pix a través de la reducción de capas y cambios de filtros de convolución para encontrar un balance entre la complejidad de la arquitectura y la calidad de las imágenes sintetizadas. Se utilizó un problema clásico en la traducción de imagen a imagen, bocetos a imagen realista aplicado a un conjunto de entrenamiento de flores. El error cuadrático medio (MSE) de la comparación entre las imágenes generadas por la arquitectura original Pix2Pix y la arquitectura propuesta es 0.6718, mientras que el índice de similitud estructural (SSIM) obtenido fue de 0.6347, lo que indica un nivel moderado de similitud perceptual. Estos resultados motivan a seguir explorando con la modificación de capas para mejorar la eficiencia sin comprometer significativamente la calidad visual de las imágenes generadas.

Palabras clave: GAN condicional, arquitectura, filtro.

Pix2Pix Architecture Optimization: A Study of Layer Reduction and Image Quality

Abstract. In this work in progress, we explore the modification of the Pix2Pix architecture through layer reduction and convolution filter modifications to find a balance between architectural complexity and the quality of the synthesized images. A classic problem in image-to-image translation, from sketches to realistic images, was applied to a training set of flowers. The mean square error (MSE) of the comparison between images generated by the original Pix2Pix architecture and the proposed architecture was 0.6718, while the structural similarity index (SSIM) obtained was 0.6347, indicating a moderate level of perceptual similarity. These results motivate further exploration with layer modification to improve efficiency without significantly compromising the visual quality of the generated images.

Keywords: Conditional GAN, architecture, filter.

1. Introducción

En el campo de la inferencia visual, una rama de investigación interesante es la traducción de imagen a imagen, esto se refiere al proceso de introducir una imagen y esta a su vez genera una nueva imagen, este proceso aplica transformaciones mediante convoluciones y convoluciones transpuestas con el objetivo de generar nuevas ilustraciones que mantienen coherencia visual del contenido original. Las aplicaciones de esta técnica van desde la transferencia de estilos y clasificación de objetos hasta la recuperación y edición semántica de imágenes. [2]

Las soluciones a este problema que aplican redes generativas adversarias (GANs) son notables [2] y permiten optimizar el entrenamiento de la red al condicionarla, este condicionamiento introduce en el entrenamiento la imagen, estableciendo así la dirección del aprendizaje. [10] El modelo que se destaca es Pix2Pix. Este estudio está centrado en contrastar diversas variaciones en la arquitectura de la red y analizar su impacto en la calidad de las imágenes generadas, se toma como inspiración de diseño, ejemplos como el de trabajos presentados en el dominio de imágenes médicas, los cuales permitieron una reducción en el coste de recursos. [1]

Uno de los principales desafíos que enfrentan redes como Pix2Pix es encontrar un equilibrio entre la capacidad de representación y la eficiencia computacional. Si el modelo tiene demasiadas capas y filtros, puede volverse demasiado complejo, lo que aumenta la posibilidad de sobreajuste y dificulta la generalización de nuevas imágenes.

Por otro lado, una red con pocas capas o filtros puede tener dificultades para capturar las características necesarias para generar imágenes de alta calidad.

El modelo propuesto aborda estos problemas de manera eficiente al reducir el número de capas y filtros mientras mantiene un balance adecuado entre velocidad y precisión. Al ajustar la arquitectura de manera cuidadosa, el modelo busca preservar las características esenciales necesarias para la generación de imágenes realistas, al mismo tiempo que mejora el tiempo de entrenamiento y reduce los recursos necesarios. Esto no solo hace que el modelo sea más accesible en términos de hardware, sino que también facilita su implementación en escenarios donde los recursos son limitados. La evaluación comparativa de los resultados tras estas modificaciones permite obtener información valiosa sobre cómo simplificar la red sin comprometer de manera significativa la calidad de la salida generada.

El artículo está organizado como se indica a continuación: En la sección 2 se presenta la fundamentación y teoría de las GANs. En la sección 3 se analiza el diseño de la red generativa propuesta. Posteriormente, en la sección 4 se muestran los resultados de las pruebas comparativas entre la arquitectura original y la arquitectura propuesta. Finalmente se presentan las conclusiones del trabajo.

2. Estado del arte

La traducción de imagen a imagen es un proceso en el que se transfiere información desde un dominio fuente a un dominio objetivo, manteniendo el contenido de la representación original. [11] Este proceso tiene una amplia gama de aplicaciones, como

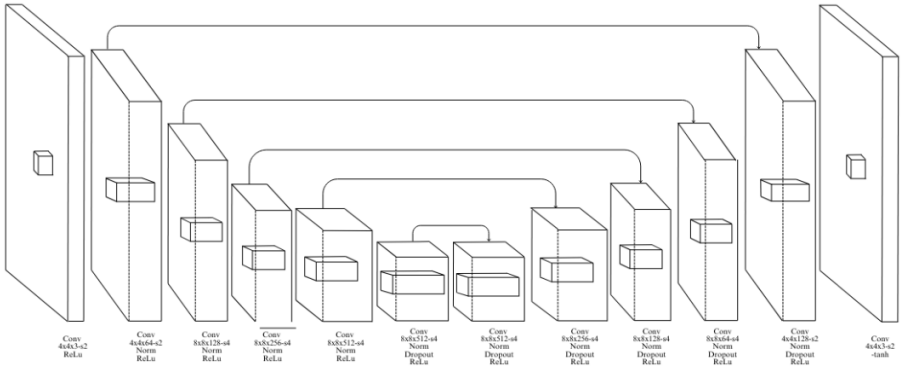


Fig. 1. En la arquitectura de la red generadora propuesta se cuenta con 12 capas convolucionales cada una seguida por un bloque de capas de normalización y ReLu, algunas capas tienen sus excepciones, el input y output de la red son imágenes de 256X256 por 3 filtros 3rgb.

en el trabajo de Kim. S, [5] quien utilizó una red residual con saltos conectados a Pix2Pix para asistir a los arquitectos en la distribución y conectividad de espacios durante el diseño de espacios, su modelo planteado alcanzó un 96% de precisión en el planteamiento de flujos de movilidad. [5] Por otro lado, la investigación de Lui, S. propone una alternativa más económica y accesible para estudios relacionados a la detección de cáncer de mama tras utilizar Pix2Pix para la generación de imágenes IHC que logró superar algoritmos tradicionales. [8] La versatilidad para la adaptación de dominios de Pix2Pix, se puede potencia como muestra el trabajo de Sun J. quien redujo de cinco pares de bloques de convolución y convolución transpuesta a tres, en la red generadora y agregó capas completamente conectadas al final de la red discriminadora, para reducir el nivel de ruido en las SPECT de baja dosis MP. [12] De hecho, investigaciones más recientes han encontrado que las mejoras en los resultados de tareas específicas están fuertemente vinculadas a modificaciones en la arquitectura de las redes, lo que demuestra, cómo la evolución de las técnicas sigue siendo un factor clave para avanzar en este campo [3].

No obstante, a pesar de estos avances, los modelos más populares de GANs continúan enfrentando desafíos importantes. Uno de los problemas más destacados es su incapacidad para generar imágenes de alta resolución [13], además, las GANs siguen siendo propensas a problemas inherentes, como la inestabilidad durante el entrenamiento y el colapso de modo, lo que puede llevar a la generación de resultados inconsistentes o de baja calidad. [6]

Varios estudios han explorado modificaciones en las arquitecturas de Pix2Pix y otros modelos GAN para mejorar su eficiencia y desempeño. Isola [4] presentaron Pix2Pix y optimizaron su arquitectura utilizando redes convolucionales de mapeo directo para tareas de traducción de imágenes, demostrando que la reducción de capas y la simplificación de la red pueden mejorar la eficiencia computacional sin sacrificar demasiado la calidad de la imagen generada [4]. Liu [8] propuso una modificación en la cantidad de filtros en las capas intermedias de una GAN, con el objetivo de reducir

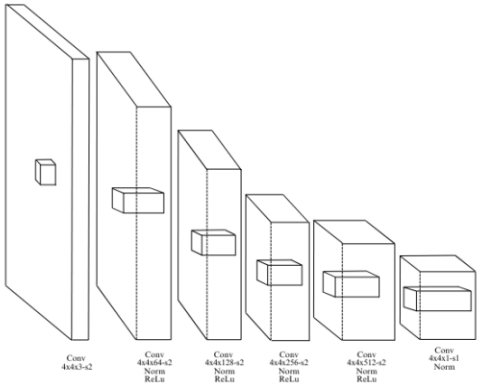


Fig. 2. La arquitectura del del discriminador de tipo *PatchGan* al igual que la del generador contiene capas de convolución, normalización y ReLu, similar a las capas donde se hace la reducción del modelo generador.

el tamaño del modelo y mejorar la velocidad de entrenamiento, sin que esto impactara negativamente en la fidelidad de las imágenes generadas [8]. Estos trabajos resaltan que la simplificación controlada de la arquitectura en las redes GAN es una estrategia efectiva para optimizar el modelo sin comprometer la calidad.

Aunque existen técnicas diseñadas para mitigar estos efectos, estas soluciones aún están en desarrollo y no se abordarán en profundidad. Sin embargo, es importante reconocer que estas estrategias ya están siendo aplicadas en algunos casos con resultados positivos como en el trabajo Liu, M, donde habilitaron saltos de capa para mitigar la pérdida de información durante las transformaciones de convolución [9] técnica que replicamos en nuestro modelo propuesto como se describe a continuación.

3. Metodología

3.1. Redes generativas adversarias

Este modelo comprende un par de redes neuronales interconectadas, denominadas Generador (G) como se muestra en la Figura 1 y Discriminador(D) representado en la Figura 2. El objetivo del generador es producir imágenes sintéticas que resulten lo más realistas posible, mientras que el discriminador recibe pares de imágenes auténticas y sintéticas con fin de clasificar correctamente a cada una según su origen. Ambas redes son entrenadas simultáneamente en un proceso competitivo. [2] Para la implementación de estas redes se utilizaron bloques de capas de normalización, Relu y convolucionales, estas últimas son un pilar del campo de visión por computadora por el alto desempeño que muestran en la comprensión de imágenes. [7]

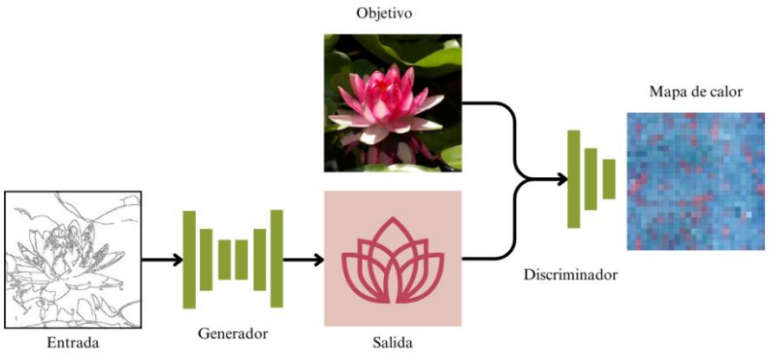


Fig. 3. Proceso de sintetización (generador) y evaluación (discriminador) de la GAN.

Fig. 1. Muestras del grupo objetivo a generar

3.2. Ajustes de la red original

La operación de convolución es fundamental en esta clase de redes por lo que modificar estas operaciones es un enfoque clave para realizar ajustes a la red original. Dentro de los ajustes que se hicieron, se logró una reducción de bloques de capas de tipo Convolución, Normalización, Relu.

Se modificaron los valores de *stride* y *kernel*, a 4 y 8 respectivamente que fueron seleccionados para lograr un equilibrio entre la velocidad de entrenamiento y la calidad de las imágenes generadas. El *stride* de 4 permite una reducción más rápida del tamaño de la imagen, lo que acelera el proceso de entrenamiento al disminuir la cantidad de operaciones requeridas.

Este valor intermedio evita la pérdida excesiva de información mientras mejora la eficiencia computacional. Por otro lado, el *kernel* de 8x8 aumenta el área de cada convolución, lo que permite al modelo capturar patrones más grandes y estructuras más amplias en la imagen. Esto mejora la coherencia y la capacidad del modelo para reconocer formas y características globales. Al combinar un *stride* alto con un *kernel*

Fig. 2. Muestras del conjunto de entrada.

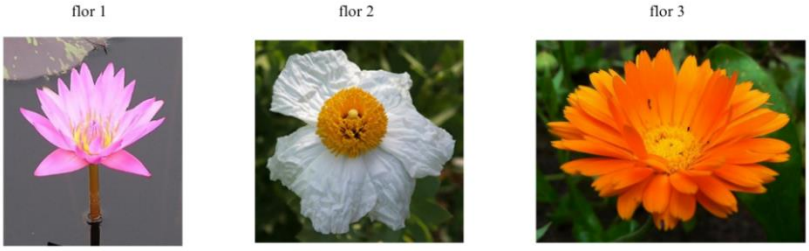


Fig. 6. Objetivo dentro del output de nuestro conjunto de datos.

más grande, se obtiene una mayor reducción de la dimensionalidad sin sacrificar los detalles importantes, lo que permite que el modelo aprenda características generales antes de enfocarse en los detalles en las capas posteriores. Este balance es crucial para mantener una alta calidad de generación sin aumentar innecesariamente el tiempo de entrenamiento o los recursos computacionales. Finalmente se redujo la cantidad de filtros en algunas capas, respetando la filosofía original del modelo, que sigue la estructura de reloj de arena como se muestra en la Figura 1. En cuanto al discriminador, se optó por utilizar el mismo de la arquitectura Pix2Pix como se ve en la Figura 2.

4. Evaluación y entrenamiento de los modelos

Para comparar Pix2Pix con el diseño propuesto, se planteó el problema de convertir una imagen estilo boceto de una flor a una imagen realista como se muestra en la Figura 3, donde los retos para el modelo consisten en generar texturas, colores y contexto lo más fieles posible a la realidad. Ambos modelos fueron entrenados con una taza de aprendizaje de $2e^{-4}$, con un lote de 50 y 80 épocas.

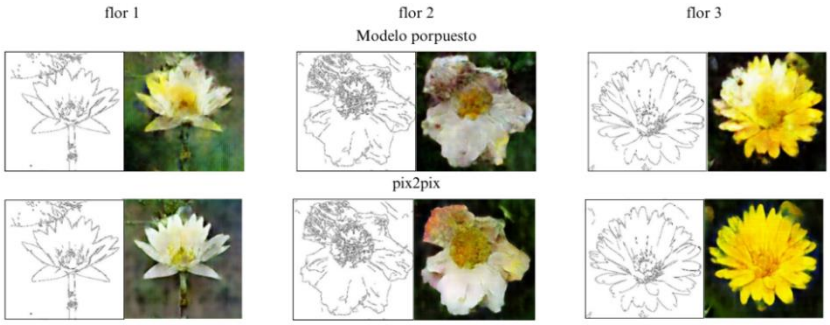


Fig. 7. Comparación de los resultados del modelo propuesto contra pix2pix.

4.1. Conjunto de entrenamiento

El conjunto de entrenamiento propuesto está conformado por 3500 imágenes variadas de flores como se ve en la Figura 4, de las cuales el modelo fue eligiendo el 90% de las imágenes (3150) y el resto (350) se consideraron para la etapa de evaluación, de los modelos. Para generar la entrada que se les dará a ambas redes, se procesaron estas imágenes para conservar únicamente los bordes más significativos como se muestra en la Figura 5. Es importante resaltar que la red generadora no tiene acceso a la imagen objetivo, sino únicamente al input, mientras que el discriminador trabaja con la salida del generador combinándola con la imagen objetivo.

4.2. Métricas de evaluación

El Error Cuadrático Medio (MSE) y la Desviación Estándar son métricas útiles para evaluar la diferencia entre imágenes generadas por GANs. El MSE mide la discrepancia pixel a pixel entre una imagen real y una generada, penalizando fuertemente las diferencias grandes, lo que permite detectar errores significativos en la generación.

Además, es una métrica computacionalmente eficiente y fácil de interpretar. Por otro lado, la desviación estándar evalúa la dispersión de los valores de píxeles en la imagen generada, lo que ayuda a identificar si la distribución de intensidades es consistente con la imagen real.

Una métrica adicional ampliamente utilizada en la evaluación de imágenes generadas es el Índice de Similitud Estructural (SSIM). A diferencia del MSE, que se enfoca en diferencias absolutas, el SSIM considera aspectos perceptuales como la luminancia, el contraste y la estructura de las imágenes, lo que lo hace más representativo de cómo el ojo humano percibe la calidad visual. El SSIM produce un valor entre -1 y 1, donde 1 indica una similitud estructural perfecta. Esta métrica es especialmente valiosa cuando se desea evaluar no solo la precisión numérica, sino también la fidelidad visual de las imágenes generadas.

En el caso de modificaciones en la arquitectura de Pix2Pix, estas métricas permiten cuantificar el impacto de los cambios en la calidad de las imágenes generadas. Por

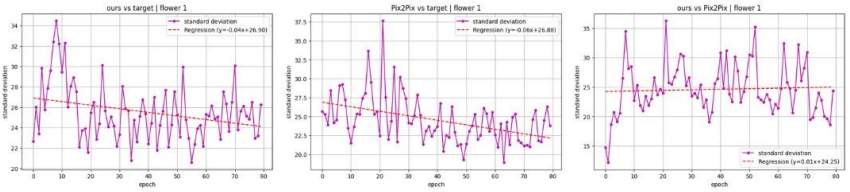


Fig. 8. Gráficas de comparativas de la desviación de la flor 1.

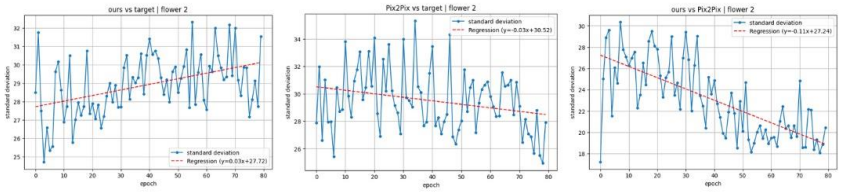


Fig. 9. Gráficas de comparativas de la desviación de la flor 2.

ejemplo, ajustar el tamaño de los filtros en la red puede afectar la precisión, la variabilidad y la percepción visual de la generación. Medir el MSE, la desviación estándar y el SSIM facilita una comparación objetiva y perceptual de los resultados, ayudando a determinar si los ajustes arquitectónicos mejoran o degradan el desempeño del modelo tanto desde una perspectiva técnica como visual.

5. Resultados

Para evaluar el desempeño de ambas arquitecturas (Pix2Pix y la propuesta), se calculó el Error Cuadrático Medio comparando 500 imágenes generadas con sus respectivas imágenes reales de referencia. Para cada par de imágenes, se obtuvo la diferencia pixel a pixel, se elevó al cuadrado y se promedió sobre el total de píxeles de la imagen. Luego, se calculó el promedio de estos errores individuales en las 500 muestras, obteniendo un valor representativo del desempeño del modelo que es de 0.6718. Este enfoque permite cuantificar de manera objetiva qué tan cercanas son las imágenes generadas por cada arquitectura a las originales, proporcionando una base sólida para comparar el impacto de la reducción de capas convolucionales y la modificación de los filtros en la calidad de la generación.

Las imágenes en ambas redes tienen una resolución de 256x256 píxeles, si bien no es perfecto, sin embargo, permite evaluar el desempeño de las arquitecturas. El manejo del color es bastante similar en ambos casos, aunque si hay una diferencia significativa en cuanto a la textura. En todas las imágenes, la silueta de las flores está correctamente

definida, lo que sugiere que el modelo ha logrado abstraer diversas formas de flores como se muestra en la Figura 7.

Para evaluar el desempeño del modelo, se hicieron comparaciones entre las imágenes en escala de grises generadas por los modelos, como se muestra en la Figura 7 con las imágenes objetivo que se muestran en la Figura 6, así como entre sí. Como métrica de comparación, se utilizó la desviación estándar de la siguiente manera:

$$\sigma = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2. \quad (1)$$

En la ecuación 1. *N* es el número total de píxeles y μ representa la media de estos. Esta fórmula se aplicó en cada época del entrenamiento para obtener gráficas que permitan visualizar el desempeño de los modelos a lo largo del proceso de entrenamiento.

Dado que por cada época se obtiene la desviación, se necesita una métrica que muestre la tendencia de la desviación, por lo que se usa la regresión lineal para representar este comportamiento.

Al comparar los resultados del generador con la imagen objetivo, se espera que la desviación siga una tendencia decreciente, lo que indica que las imágenes generadas son cada vez más parecidas. En los casos de la flor 1, mostrada en la Figura 8 y la flor 3, mostrada en la Figura 10, la pendiente de la línea de regresión es negativa, es decir que la desviación estándar disminuye progresivamente. Sin embargo, hay casos donde el modelo propuesto diverge la similitud de resultados, esto es apreciable en la Figura 9 para el caso de la flor 2, donde la desviación estándar aumenta.

Se realizó una comparación entre los resultados del modelo propuesto y el modelo Pix2Pix, donde se espera que la desviación estándar tienda a 0. Esto indicará que ambas imágenes generadas por los modelos son muy parecidas. Una desviación estándar de 0 implicaría que ambas imágenes son idénticas, lo que sugeriría que los ambos modelos se comportan de manera exactamente igual. Además, se espera que la pendiente de la línea de regresión lineal tenga una pendiente negativa, lo que reflejaría la tendencia de la desviación estándar a aproximarse a 0. En este caso, los resultados son más prometedores que en la comparación con la imagen objetivo. El comportamiento de ambos modelos se muestra más estable, con una pendiente cercana a 0 en la regresión lineal.

Además, se observa en las gráficas que el modelo propuesto y el modelo Pix2Pix generan imágenes muy parecidas al inicio del entrenamiento. Sin embargo, alrededor de la época 10, empiezan a divergir para luego volver a asemejarse.

Esto sugiere que el modelo propuesto presenta un comportamiento muy similar al del modelo Pix2Pix, siendo el caso de mayor éxito en la flor 2, como se muestra en la Figura 9.

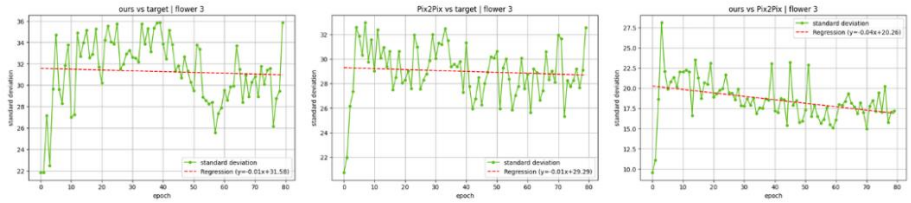


Fig. 10. Gráficas de comparativas de la desviación de la flor 3.

Para evaluar la similitud estructural entre las imágenes generadas por el modelo original Pix2Pix y el modelo propuesto, se utilizó el Índice de Similitud Estructural como métrica de referencia perceptual. El cálculo del SSIM se realizó durante 80 épocas de entrenamiento, comparando en cada una de ellas la imagen generada por la arquitectura modificada contra la correspondiente imagen generada por el modelo base Pix2Pix. Este enfoque permitió analizar cómo evolucionaba la similitud visual entre ambas salidas a lo largo del proceso de aprendizaje. Al finalizar el entrenamiento, se obtuvo un valor promedio de SSIM de 0.6347, lo cual indica un nivel de similitud estructural moderado entre ambas arquitecturas. El SSIM se define como:

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}. \quad (2)$$

En la ecuación 2 μ_x y μ_y representan las medias de las imágenes x e y, σ_x^2 y σ_y^2 son las varianzas, y σ_{xy} la covarianza entre ambas.C_1 y C_2 son constantes de estabilización. Este índice produce un valor entre -1 y 1, donde 1 representa una similitud estructural perfecta. En este contexto, el valor obtenido de 0.6347 sugiere que, a pesar de la simplificación arquitectónica, el modelo propuesto logra mantener una representación visual coherente y comparable a la del modelo original.

6. Conclusión

La reducción en la cantidad de capas convolucionales y la modificación de los filtros en Pix2Pix resultó en un Error Cuadrático Medio de 0.6718, lo que indica que las imágenes generadas por el modelo modificado tienen un nivel de similitud considerable con las generadas por la arquitectura original Pix2Pix. Adicionalmente, se obtuvo un Índice de Similitud Estructural de 0.6347, lo cual sugiere que las imágenes producidas no solo presentan una correspondencia numérica aceptable, sino que también mantienen una coherencia visual perceptible en términos de estructura, forma y contraste. Este valor de SSIM representa una similitud estructural moderada, lo que refuerza la viabilidad de aplicar simplificaciones arquitectónicas sin comprometer significativamente la calidad visual.

Este resultado sugiere que la optimización en la arquitectura no afectó drásticamente la precisión ni la percepción visual de la generación, lo que puede traducirse en un modelo más eficiente computacionalmente y más adecuado para aplicaciones en las que se prioriza la velocidad y el uso reducido de recursos.

Este trabajo es relevante para escenarios donde se busca reducir el consumo de recursos computacionales, como en dispositivos con capacidad de cómputo limitada o en entornos donde la rapidez en la generación de imágenes es prioritaria. A futuro, se podrían explorar otras modificaciones en la arquitectura, como el ajuste fino de hiperparámetros, el uso de estrategias de regularización, o la implementación de técnicas de compresión de redes neuronales, con el fin de continuar optimizando el desempeño del modelo sin comprometer su precisión ni su calidad visual.

Referencias

1. Ali, O., Ali, H., Ali, S.A.: Implementation of a Modified U-Net for Medical Image Segmentation on Edge Devices. IEEE Transactions On Circuits And Systems II: Express Briefs (2022) doi: 10.48550/arXiv.2206.02358.
2. Creswell, A., White, T.: Generative Adversarial Networks: An Overview. IEEE Signal Processing Magazine, pp. 53–65 (2018) doi: 10.48550/arXiv.2206.02358.
3. Ikhsan, G., Suciati, N.: The Comparative Study of Adding Edge Information to Pix2pix Architecture for Face Image Generation. In: 6th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE), pp. 160–165 (2022) doi: 10.1109/ICITISEE57756.2022.10057739.
4. Isola, P., Zhu, J.Y., Zhou, T.: Image-to-Image Translation with Conditional Adversarial Networks. CVPR (2017) doi: 10.1109/CVPR.2017.632.
5. Kim, S., Lee, J., Jeong, K.: Automated Door Placement in Architectural Plans Through Combined Deep-Learning Networks of ResNet-50 and Pix2Pix-GAN. Expert Systems with Applications, 244, pp. 122932 (2024) doi: 10.1016/j.eswa.2023.122932.
6. Kossale, Y., Airaj, M., Darouichi, A.: Mode Collapse in Generative Adversarial Networks: An Overview. In: 8th International Conference on Optimization and Applications (ICOA), pp. 1–6 (2022) doi: 10.1109/ICOA55659.2022.9934291.
7. Li, Z., Liu, F., Yang, W.: A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects. IEEE Transactions on Neural Networks and Learning Systems, 33(12), pp. 6999–7019 (2022) doi: 10.1109/TNNLS.2021.3084827.
8. Liu, S., Zhu, C., Xu, F.: Breast Cancer Immunohistochemical Image Generation through Pyramid Pix2pix. ArXiv (2022) doi: 10.48550/arXiv.2204.11425.
9. Liu, M.Y., Tuzel, O.: The Conditional Adversarial Loss For Image-to-Image Translation. CVPR (2019) doi : 10.48550/arXiv.1611.07004.
10. Mirza, M., Osindero, S: Conditional Generative Adversarial Nets. ArXiv (2014)
11. Pang, Y., Lin, J., Qin, T.: Image-to-Image Translation: Methods and Applications. IEEE Transactions on Multimedia, 24, pp. 3859–3881 (2022) doi: 10.1109/TMM.2021.3109419.
12. Sun, J., Du, Y., Li, C.: Pix2Pix Generative Adversarial Network for Low Dose Myocardial Perfusion SPECT Denoising. Quant Imaging Med Surg (2022) doi: 10.21037/qims-21-1042.
13. Wang, L., Sindagi, V.: High-Quality Facial Photo-Sketch Synthesis Using Multi-Adversarial Networks. In: 13th IEEE International Conference on Automatic Face & Gesture Recognition, pp. 83–90 (2018) doi: 10.1109/FG.2018.00022.

ISSN 1870-4069

Optimizing Best Response Dynamics-based Facility Location Games Using Reinforcement Learning

Andrés Burjand Torres Reyes, Rolando Menchaca-Méndez,
Francisco Hiram Calvo-Castro

Instituto Politécnico Nacional,
Centro de Investigación en Computación,
Mexico

atorresr2024@cic.ipn.mx, rmen@cic.ipn.mx, hcalvo@cic.ipn.mx

Abstract. In this article, we propose a model based on Best Response Dynamics (BRD) to examine the behavior of a group of rational agents when an external regulatory entity enforces control policies that influence the agents' dynamics. BRD is valuable for analyzing economic and social phenomena, as it captures the tendency of agents to seek to maximize their individual benefits—a common behavior in these contexts. However, these models frequently converge to a Nash equilibrium, which may not represent a socially optimal outcome. To address this limitation, we suggest introducing an external regulatory agent that employs reinforcement learning to enhance the convergence time to Nash equilibria or, ideally, to guide the system toward socially optimal solutions. We utilize an environment modeled after a Facility Location Game (FLG) to train a reinforcement learning agent and assess the impact of its policies on the FLG's behavior. This methodology presents a novel application of game theory and reinforcement learning for regulating complex systems, with potential implications in economics, social systems, robotics, and engineering. We present preliminary results to support our findings.

Keywords: Reinforcement learning, multi-agent systems, game theory, best response dynamics, facility location games.

1 Introduction

In today's interconnected world, addressing economic, social, and environmental challenges requires a deep understanding of complex systems. These systems often consist of agents (such as individuals, organizations, or entities) that interact and make decisions to optimize their individual benefits. Although such self-interested behaviors frequently result in predictable outcomes, such as Nash equilibria, they may also prevent the system from reaching a social optimum. Even when collective benefit is maximized, an agent can unilaterally adjust its strategy to increase personal gain, disrupting the system's balance.

This paper presents work in progress; it focuses on the initial steps toward addressing the challenge of dynamically regulating self-interested agents in

complex systems to reconcile individual incentives with collective welfare. Specifically, we show preliminary results on improving convergence time to a Nash Equilibrium in a Facility Location Game with players under Best Response Dynamics behavior using a Reinforcement Learning framework.

Game theory provides powerful tools for analyzing agent interactions, with Best Response Dynamics (BRD) being a prominent method for modeling rational decision-making. BRD is a process in which agents use a local search method to achieve outcomes that benefit them individually while driving the system, in general, toward Nash equilibrium [6]. It is particularly effective in modeling economic and social scenarios in which agents aim to maximize individual benefits. This model studies the interaction of rational agents who make decisions to maximize an objective function based on the system's current state, a formulation rooted in the domain of game theory of potential games [17]. However, it suffers from exponential convergence times and an inability to adapt to external interventions or collaborative behaviors.

To overcome these challenges, this work introduces an external regulatory agent equipped with reinforcement learning capabilities. This agent operates independently of the economic agents and intervenes in the system's dynamics by applying incentives or taxes. Unlike traditional models, this regulatory agent dynamically adapts its strategies to optimize regulatory policies, enabling more effective interventions in diverse scenarios. Its primary objectives are as follows:

- Accelerating convergence to Nash Equilibria: Reinforcement learning techniques reduce convergence times from exponential to polynomial, making the process more efficient.
- Guiding the system toward a social optimum: Ensuring outcomes that maximize collective welfare.

The effectiveness of this approach is evaluated using a Facility Location Game, a canonical optimization problem with applications in economics and operations research.

This research contributes an innovative framework for regulating complex systems, bridging game theory and machine learning to address real-world challenges in economic and social contexts.

2 Justification

Efficient regulation of decentralized systems is critical in domains such as energy markets, traffic routing, and public resource allocation, where static, one-size-fits-all policies struggle to address real-time dynamics' inherent volatility and complexity. In energy markets, for instance, the rise of distributed renewable energy sources (e.g., solar panels, wind farms) and fluctuating demand patterns necessitate adaptive pricing and grid-balancing mechanisms to prevent blackouts or the curtailment of renewable generation. Static tariff structures or fixed supply-demand models often fail to account for sudden

weather changes, shifts in consumer behavior, or equipment failures, leading to inefficiencies and instability [9].

Similarly, in traffic routing, rigid signal timings or preprogrammed navigation systems cannot respond to real-time congestion caused by accidents, road closures, or surges in ride-sharing demand. Adaptive traffic management systems, powered by IoT sensors and machine learning, dynamically reroute vehicles and adjust signal cycles to minimize delays and emissions [5].

Public resource allocation—such as distributing emergency aid during disasters or optimizing vaccine delivery during pandemics—also demands real-time adjustments to evolving needs, supply chain disruptions, or demographic inequities. Static policies risk misallocating resources and leaving vulnerable populations underserved, which is a significant concern in fields like public health [2].

These challenges underscore the need for decentralized regulatory frameworks that integrate real-time data, predictive analytics, and feedback loops to balance efficiency, equity, and resilience in dynamic environments.

3 Theoretical Framework

3.1 Game Theory Foundations

As stated by [14], "Game Theory aims to model situations in which multiple participants interact or affect each other's outcomes". These situations are often considered as strategic games, and, according to [7], involve:

- A set of players (the participants) $N = 1, 2, ..., n$,
- Strategy profiles $A = A_1 \times A_2 \times ... \times A_n$, which are the combination of actions chosen by all players in the game, where A_i is the set of actions available to player i
- Utility (or payoff) functions $u_i : A \to \Re$

A *Nash Equilibrium* (NE) is a strategy profile x^* satisfying:

$$x^* \in \mathrm{BR}(x^*)\,,$$

where BR denotes the best-response mapping. In other words, no player can increase their payoff by unilaterally deviating from x^*.

A social optimum is considered a situation that maximizes the total welfare of all players in a game. Mathematically, it is defined as a situation that maximizes a social welfare function, aggregating all players' utilities. The social welfare function $W : A \to \Re$ aggregates individual utilities. One of the most common aggregation methods is the utilitarian social welfare, which sums the utilities of all players:

$$W(a) = \sum_{i \in N} u_i(a), \tag{1}$$

where $a = (a_1, a_2, a_3, ..., a_n)$ is an action profile. An action profile $a^* \in A$ is socially optimal if:

$$a^* \in \arg\max_{a \in A} W(a) = \arg\max_{a \in A} \sum_{i \in N} u_i(a), \tag{2}$$

A very important consideration is the fact that a social optimum is not necessarily a Nash Equilibrium, since certain individual incentives could lead players to deviate from such situations.

According to [13], *potential game* are the ones where a function $\Phi : S \to \mathbb{R}$ exists such that for every player i, any strategy profile $s = (s_i, s_{-i})$, and any alternative strategy s'_i, the change in player i's utility satisfies:

$$u_i(s'_i, s_{-i}) - u_i(s_i, s_{-i}) = \Phi(s'_i, s_{-i}) - \Phi(s_i, s_{-i}), \tag{3}$$

where u_i denotes the utility function of player i.

3.2 Best Response Dynamics (BRD)

Best Response Dynamics (BRD) models rational decision-making in strategic games by assuming agents iteratively update their strategies to maximize individual utilities based on others' actions. This process can be represented as a directed graph where nodes correspond to action profiles, and edges denote transitions via unilateral best-response deviations. At each step, a player switches to a strategy that maximizes their payoff given the current actions of others, driving the system toward equilibrium states. [12]

While BRD converges to Nash equilibria in potential games [17], traditional models face critical limitations:

- They cannot incorporate external interventions,
- They fail to account for learning processes or collaborative behaviors among agents,
- Many existing regulatory frameworks are static and unable to adapt to dynamic systems.

3.3 Reinforcement Learning

Reinforcement Learning (RL) is a machine learning technique where an agent interacts with an environment and tries to obtain the policy that yields the maximum possible reward from said environment, using trial and error as its basis. Using the concept of delayed rewards, RL encapsulates that actions can affect both present and future rewards, improving the decision-making process. RL is also very useful in uncertain environments, because it is designed to keep the focus on proposed objectives, using the Markov Decision Processes to formalize the interaction between an agent, its actions, and its goals. One of the core challenges of RL is strategically balancing exploration (when the agent tries new actions in the hope of better results) and exploitation (when

the agent uses actions that proved helpful in the past), a dilemma absent from supervised and unsupervised learning algorithms. This holistic approach makes RL particularly suitable for real-time decision-making tasks where uncertainty and long-term planning are crucial. [16]

3.4 Facility Location Games

The Facility Location Game is an optimization problem where the goal is to determine which facilities to open and how to assign customers cost-effectively. Given a set of facilities F and a set of customers U, each facility $i \in F$ has a fixed, non-negative opening cost f_i. Additionally, serving a customer $j \in U$ from a facility $i \in F$ incurs a non-negative service cost c_{ij}, which depends on the specific facility–customer pair. The objective is to minimize the total cost, which consists of the sum of the opening costs of the selected facilities and the service costs of assigning each customer to an open facility. This requires making two key decisions: selecting the facilities to open and determining the optimal assignment of customers to these facilities while ensuring every customer is served [8, 11].

Facility Location Games, as potential games, share the property of guaranteed convergence to a Nash equilibrium, although the upper bound is considered exponential (i.e., $O(n^m)$). Another important characteristic of potential games is that they have a central function, called the potential function, which is optimized by the actions of all the players [18].

4 Related Work

A state-of-the-art search reveals several articles that analyze and apply Best Response Dynamics as a modeling technique. For example, [3] studied public goods games played on networks with possibly non-reciprocal relationships between players, where they explained how and why a Nash equilibrium is not always achieved in games on directed networks (which implies unequal relationships); this paper poses an interesting background since taking into consideration the nature of the relation between the agents in the model can improve its plausibility, and it could be interesting to study how the regulating agent could help achieve a Nash equilibrium.

Some researchers have utilized evolutionary game theory to understand spatial collective decision-making behaviors, such as [22]. In this case, they developed incentive mechanisms (reward and punishment) to investigate asynchronous BRD of anti-coordinating agents. This approach could be fruitful compared to the one proposed for this thesis. It is also interesting to point out the distinction between coordinating and anti-coordinating agent actions—former when, if one strategy prevails, agents in the system will be favored to follow it; latter when individuals take the opposite action if most game partners make the same choice [19].

A recent study by [4] explores discrete opinion dynamics in social networks with stubborn agents, where conformists adopt the most common opinions

from their neighbors, but stubborn agents remain unaffected by others. This research transforms the opinion dynamics into an n-strategy evolutionary game model with best-response updating, shedding light on how information influences strategy evolution. When agents have complete information about all their neighbors' opinions, the game becomes a potential game, guaranteeing the existence of at least one pure-strategy Nash equilibrium (PNE) and ensuring convergence to a PNE through asynchronous BRD. However, multiple PNEs often arise, complicating predictions of the evolutionary outcome. An interesting extension of this work is provided when information is limited, as agents can observe their neighbors with a probability of less than one. In this case, the game results in a unique stationary strategy distribution if stubborn agents are present, and the corresponding PNE becomes globally stable under both synchronous and asynchronous dynamics. This finding introduces the idea that a combination of stubborn agents and limited information can function as an equilibrium-selection mechanism, making the evolutionary outcome more predictable. The authors use numerical simulations on various network types, demonstrating how the distribution of stubborn agents and the available information level can significantly influence the final evolutionary outcome, showing how agents with different opinions converge to the PNE.

Reinforcement Learning is a well-respected machine learning paradigm in the literature. It is common to find uses in control theory, for instance, [10], where they used RL for heading control for unmanned sailboats using a backstepping sliding mode approach; here they propose an RL-based controller that enhances tracking performance and robustness against disturbances by integrating adaptive compensation mechanisms. The simulations show it outperforms existing methods, demonstrating RL's effectiveness in improving control strategies for uncertain and dynamic systems. There are a variety of uses as well in financial applications, such as [21], where they present a novel approach to equity portfolio optimization by integrating spectral analysis, portfolio theory, and deep reinforcement learning. In [15], the researchers show a predictive-based reinforcement learning (PRL) model to improve credit assessment for manufacturers and importers; by integrating predictive analytics and reinforcement learning, PRL enhances credit-scoring accuracy, decision-making, and financial stability.

It is also possible to find multiple articles that use both game theory and reinforcement learning to model complex problems. For instance, [1] explores the intersection of these two fields to model cyber-physical human systems by "proposing a computationally feasible approach to simultaneously model multiple humans as decision-makers, instead of determining the decision dynamics of the intelligent agent of interest and forcing the others to obey certain kinematic and dynamic constraints imposed by the environment." This multi-agent method could have certain advantages over the use of a regulatory agent, but it is also more complex; there is also an opportunity to find the intersection between both ideas, because many social and political scenarios include groups of

decision-makers capable of adaptation while having a regulatory agent with a different nature.

A promising direction for advancing this field is integrating evolutionary game theory with reinforcement learning, leveraging non-cooperative game theory to address the dynamic and complex nature of the agents' interactions. The research developed by [20] proposes a hybrid framework where a non-cooperative competition dynamically selects policy update modes using Nash equilibrium, ensuring diversity in agent strategies, while a cooperative collaboration balances exploration and convergence. The system can overcome local optima and adapt more effectively to dynamic conditions by allowing evolutionary algorithms to drive environment-independent exploration. This approach highlights the potential of combining game-theoretic principles with reinforcement learning. It suggests pathways for enhancing models of regulatory and adaptive agents in socio-political and multi-agent systems, aligning well with the challenges and opportunities identified for this work.

5 Proposed Solution and Methodology

5.1 Description of the Simulated Environment

This simulated environment will be a graph-based representation of a facility location game. The base will consist of a weighted, undirected, connected tree graph $G = (V, D, E, W)$, where:

- V (nodes):
  - Each with an associated client demand.
  - Potential facility locations $F \subseteq V$.
- D (node weights): node weights represent the demand at each node, such that $\forall d \in D,\ d \in \mathbb{N}$.
- E (edges): connections between nodes (e.g., roads, transit links).
- W (edge weights): edge weights represent serving costs (e.g., distance, congestion, transportation fees), such that $\forall w \in W,\ w \in \mathbb{N}$.

The decision to use a tree was made to ensure convergence via BRD. There are three main reasons: the potential is bounded ($\phi \geq 0$); each best response reduces ϕ or leaves it unchanged if equilibrium is reached; and the absence of cycles prevents infinite loops.

The group of players acting under Best Response Dynamics can be defined as a set $N = \{player_1, player_2, \ldots, player_n\}$, where each player i chooses a location $f_i \in F$ to build its unique uncapacitated facility with no associated building cost, and this location becomes exclusive to i while i decides to keep it. We also define a distance function $d_G(x, y)$ giving the shortest-path distance between any two vertices $x, y \in V$ and a profit function $U_i(f_i, f_{-i})$ depending on the locations of all players, defined in equation (4):

$$U_i(f_i, f_{-i}) = \sum_{\substack{c \in V \\ f_i = \text{nearest}(c)}} D(c) \cdot \big(1 - d_G(c, f_i)\big). \tag{4}$$

Since this FLG is considered under the potential games framework, we use a global potential function ϕ that reflects system-wide efficiency:

$$\phi(f) = \sum_{c \in V} D(c) \cdot d_G\big(c, \text{nearest}(f)\big). \tag{5}$$

Here, ϕ represents the total weighted distance from all clients to their nearest facility. Thus, players' strategies directly impact ϕ.

The best-response update for player i at time step $t+1$ is described in equation (6):

$$f_i^{(t+1)} = \arg \max_{f \in F \setminus f_{-i}^{(t)}} U_i\big(f, f_{-i}^{(t)}\big). \tag{6}$$

where $f_{-i}^{(t)}$ are the locations of the other players at time t. When the distance calculation between customers and facilities results in a tie, it is broken by random assignment.

Given these definitions, the BRD process proceeds as follows:

1. Start with an initial random configuration $(x_1^{(0)}, x_2^{(0)}, \ldots, x_n^{(0)})$ with no overlap.
2. At each time step t, select a player at random to update their position to their best response given the current positions of others.
3. Halt when no player can improve their utility (PNE).

Under this design, a Nash equilibrium arises when no player can unilaterally move to capture more clients (i.e., $\nexists f_i' : U_i(f_i', f_{-i}) > U_i(f_i, f_{-i})$).

After success with this simple simulation, certain changes could improve the model's real-life applicability. For example, we could simulate a more dynamic graph where road conditions change over time, include traffic congestion, model evolving client demand with stochastic variations, etc.

5.2 Preliminary description of the proposed solution

As we discussed earlier, there are two main problems with agents that act selfishly in any given game: the time complexity to achieve Nash Equilibrium is usually exponential, and many states are not socially optimal since the algorithms they follow usually lead to local optimums; this is the case as well for BRD [4]. We decided to address both of these issues by implementing an agent with a different nature, one that can alter certain game conditions by modifying the rules, applying incentives, and learning how to obtain the optimal values for the variables it can control using reinforcement learning. Fig 1 illustrates this idea. Fig 1 represents the process through which the agent obtains information of the environment given by the Facility Location Game with BRD to decide how to act through regulations and incentives (with policy π), and also takes rewards that help it improve its policy. Reinforcement learning depends on the agent-environment framework since it pays attention to the current state of the

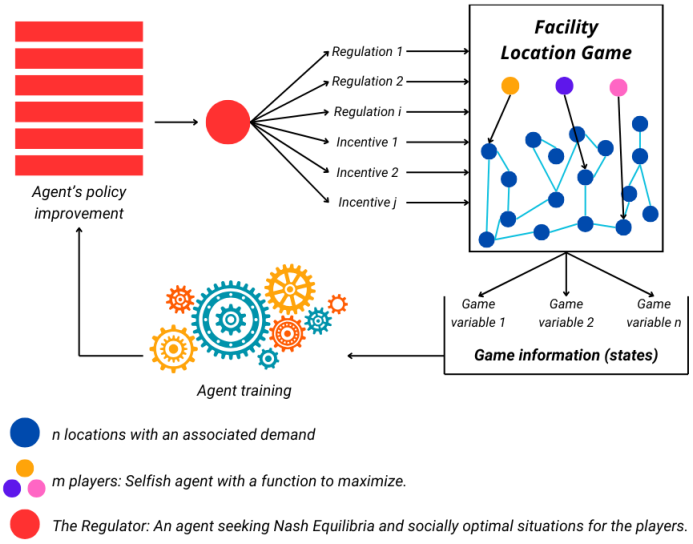


Fig. 1. Proposed architecture.

situation and considers its space of action. The first step is creating a simulation of the Facility Location Game, described in the previous section. Once the simulation is operating, we can extract the set of states corresponding to the environment. These states will be encoded to capture spatial relationships with a set of variables such as the positions of facilities and clients, the distribution of demand, the number of players, the players' decisions, costs, etc, as well as the potential function, which is the function to maximize.

This paper is centered on the objective of reducing convergence time. The process of balancing both objectives, given that we proposed a Multi-Objective Reinforcement Learning framework, is still being developed.

The regulatory agent (RA) will take actions a_t from a predefined action space A ($a_t \in A = \{\text{tax on locations}, \text{incentives for players}, \dots\}$) depending on the state S_t; from this space, it will select a policy to pursue its two main objectives:

1. Accelerate convergence to NE, which we formally consider as the reduction to polynomial time ($O(n)$).
2. Obtaining socially optimal situations, even if it means escaping Nash Equilibria

This policy will be iteratively improved with the reward function assigned to the environment, which will give the necessary incentives to the RA to achieve an optimal policy, which is the policy that secures the best positions for the players, and, therefore, the best rewards for the RA. A policy π can be defined as:

$$\pi : S_t \rightarrow A.$$

Each state S_t at time t consists of:

- Graph structure G_t: The adjacency matrix of the current network.
- Facility locations (Both potential and taken) F_t: A binary vector indicating which nodes have facilities.
- Demand distribution D_t: A vector assigning demand values to client nodes.
- Cost matrix C_t: Transportation costs from each client to its assigned facility.

The final reward signal will most likely be based on arguments like: cost minimization (e.g., transportation or setup costs), Stagnation (when there are no changes in players' positions or utilities), and/or convergence incentives (e.g., penalizing large deviations from equilibrium solutions). A proposed reward function $R(S_t)$ is:

$$R(S_t) = \Delta\phi_t - I. \tag{7}$$

Where I is the number of iterations, this way, the reward is proportional to how much the general utility is improved and is reduced by the iterations it takes to achieve NE.

The final implementation of the RA will make use of different variables to achieve its goal, which will be its *action space*, like:

- Conveniently changing the player selection order (instead of it being random) to prioritize players with higher potential to improve system efficiency.
- Utility function weighting: Dynamically adjusting α and β weights corresponding to the *sum_ demands* and *sum_ costs* variables in the utility function.
- Altering the tie-breaking rules
- Facility activation/deactivation

It will use a model-free approach called Proximal Policy Optimization, which is suitable for discrete action spaces.

In the Preliminary Results section, we implement a lightweight regulatory agent based on a tabular reinforcement-learning scheme with an ε-greedy Monte Carlo policy as a proof of concept. At each time step, the agent "intervenes" by selecting exactly one of the currently non-converged players—dynamically prioritizing those whose move is estimated to yield the greatest improvement in global efficiency. Concretely, the action space at state S_t has size equal to the number of active players n. To keep the state representation compact, we discretize each player's individual utility $u_i^{(t)}$ and the overall potential function $\phi^{(t)}$ into a small number of bins (from "very low" to "very high"), which were estimated using statistical data from the simulations run using only FLG + BRD; the resulting tuple

$$s_t = \big(\mathrm{bin}(u_1^{(t)}), \ldots, \mathrm{bin}(u_n^{(t)}), \mathrm{bin}(\phi^{(t)})\big). \tag{8}$$

serves as the index into a Q-table $Q(s, a)$, which is initialized with a small positive constant (10^{-5}) to encourage early exploration. At each decision point, with

probability ε the agent picks a random valid player (exploration), and otherwise it exploits by choosing

$$a_t \;=\; \arg\max_{a \in \mathcal{A}(s_t)} Q(s_t, a)\,. \tag{9}$$

Immediately after the chosen player best-responds and the game state advances, we compute a scalar reward

$$R(S_t) \;=\; \phi^{(t)} \;-\; \phi^{(t-1)} \;-\; I \;\times pw, \tag{10}$$

where I is the current iteration count (to penalize long trajectories) and pw is a tunable penalty weight. During learning, we perform the incremental Bellman update

$$Q(s_t, a_t) \;\leftarrow\; Q(s_t, a_t) \;+\; \alpha\Big[R(S_t) \;+\; \gamma \max_{a'} Q(s_{t+1}, a') \;-\; Q(s_t, a_t)\Big], \tag{11}$$

with learning rate $\alpha = 0.1$ and discount factor $\gamma = 0.99$. Once the Facility Location Game converges, a Monte Carlo end-of-episode pass backpropagates the final reward through the entire episode history: computing the return G backward and adjusting each visited Q-entry by

$$Q(s_t, a_t) \;\leftarrow\; Q(s_t, a_t) \;+\; \alpha\,\big(G - Q(s_t, a_t)\big). \tag{12}$$

To balance exploration and exploitation over successive episodes, the exploration rate is decayed multiplicatively:

$$\varepsilon \;\leftarrow\; 0.995\,\varepsilon. \tag{13}$$

Preliminary experiments show that this simple regulatory intervention significantly accelerates convergence of best-response dynamics, reducing both the number of iterations and the variance of the potential-function trajectory.

6 Preliminary Results

In fig 2 we show how the previously defined simulation results look using a specific seed (66). The simulation was implemented in Python, and the source code (including the Facility Location Game environment, Best Response Dynamics logic, and visualization tools) is publicly available on GitHub under an open-source license. The repository can be accessed at: https://github.com/Burjand/facility_location_game.git. These were the used hyperparameters:

- Number of nodes: 100
- Number of potential facilities: 80
- Number of BRD players: 10
- Seed: 66

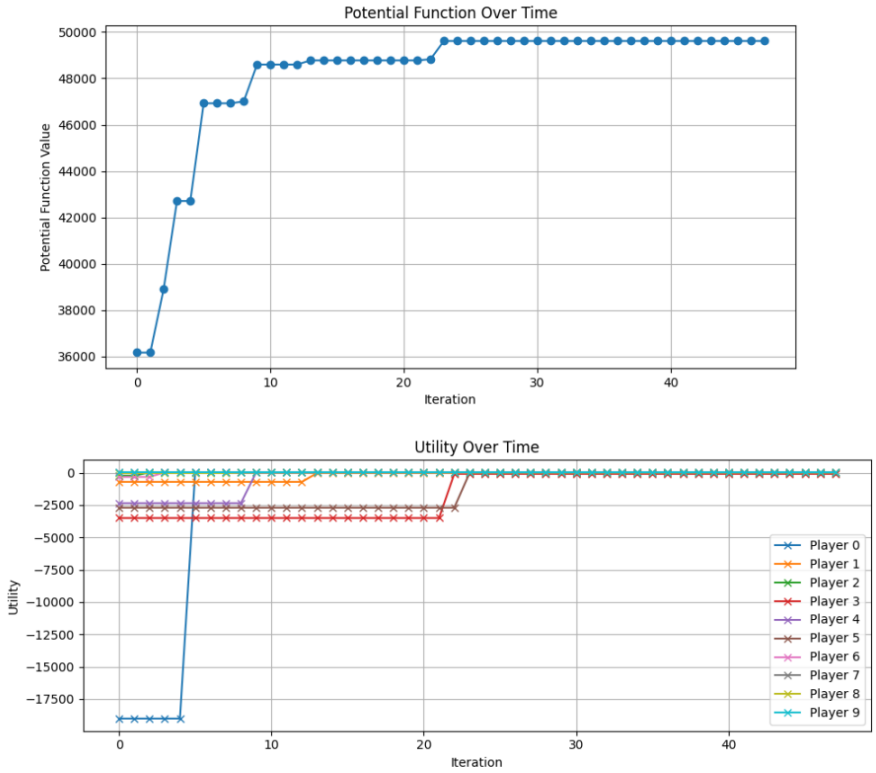


Fig. 2. FLG with BRD simulation results.

Running 1000 simulations with different seeds chosen randomly, 100 nodes, 80 potential facilities, and 10 players, the average number of iterations to achieve Nash Equilibrium was 58.515, and the average potential function value at the end was 60597.483. After implementing the regulatory agent with the conditions stated before, we ran again 1000 simulations with different seeds chosen randomly and the same parameters, and obtained that the average number of iterations to achieve Nash Equilibrium was 43.817, and the average potential function value at the end was 50949.478.

7 Scientific Novelty

The majority of studies involving multi-agent modeling have two main approaches regarding their adaptability to achieve objectives: The first is that agents act selfishly and non-cooperatively, maximizing their own utility function; the second is that each agent can learn, allowing them to adapt to situations in a more "intelligent" manner. The novelty of this work lies in a variation of the first modeling approach, introducing an agent capable of

modifying the system's rules so that agents can reach Nash Equilibria more quickly despite lacking the adaptability provided by learning and cooperating in the second approach.

8 Limitations and Future Work

Despite these promising benefits, the modeling technique has notable limitations. One of the most significant is its inability to incorporate the agents' capacity for cooperative behavior, long-term adaptation, and learning from past actions. This limitation arises because the framework assumes purely rational and selfish players. A potential area for future improvement —either in this work or in subsequent studies— would be enhancing the model to simulate agents' learning abilities and collaborative behavior, which can turn this into a Multi-agent Reinforcement Learning problem. Some other future works that could be developed based on this one could be developing computational methods to scale the model for larger, more complex systems and testing the framework in real-world scenarios to validate its assumptions and refine its applicability.

Acknowledgments. This research was funded in part by the "Secretaría de Ciencia, Humanidades, Tecnología e Innovación" (Secihti) of Mexico and the "Instituto Politécnico Nacional" under grant SIP 20253428.

References

1. Albaba, B. M., Yildiz, Y.: Modeling cyber-physical human systems via an interplay between reinforcement learning and game theory. Annual Reviews in Control, vol. 48, pp. 1–21 (1 2019) doi: 10.1016/j.arcontrol.2019.10.002
2. Barret, H., Ortmann, W., Dawson, L., Saenz, C., Reis, A.: Resource allocation and priority setting, vol. 3. Springer (2016), https://www.ncbi.nlm.nih.gov/books/NBK435776/
3. Bayer, P., Kozics, G., Szőke, N. G.: Best-response dynamics in directed network games. Journal of Economic Theory, vol. 213, pp. 105720 (8 2023) doi: 10.1016/j.jet.2023.105720
4. Cao, W., Zhang, H., Kou, G., Zhang, B.: Discrete opinion dynamics in social networks with stubborn agents and limited information. Information Fusion, vol. 109, pp. 102410 (4 2024) doi: 10.1016/j.inffus.2024.102410
5. Eremina, L., Mamoiko, A., Aohua, G.: Application of distributed and decentralized technologies in the management of intelligent transport systems. Intelligence Robotics, vol. 3, no. 2, pp. 149–61 (1 2023) doi: 10.20517/ir.2023.09
6. Feldman, M., Snappir, Y., Tamir, T.: The Efficiency of Best-Response Dynamics. Springer (1 2017), https://doi.org/10.1007/978-3-319-66700-3_15
7. Hara, K.: Coalitional strategic games. Journal of Economic Theory, vol. 204, pp. 105512 (7 2022) doi: 10.1016/j.jet.2022.105512
8. Iloglu, S., Albert, L. A., Michini, C.: Facility location and restoration games. Computers Operations Research, vol. 174, pp. 106896 (2025) doi: https://doi.org/10.1016/j.cor.2024.106896

9. Lammers, I., Diestelmeier, L.: Experimenting with Law and Governance for Decentralized Electricity Systems: Adjusting Regulation to Reality? Sustainability, vol. 9, no. 2, pp. 212 (2 2017) doi: 10.3390/su9020212
10. Li, C.-M., Zhang, B.-L., Cao, Y.-L., Yin, B.: Reinforcement learning-based backstepping sliding mode heading control for unmanned sailboats. Ocean Engineering, vol. 327, pp. 120936 (2025) doi: https://doi.org/10.1016/j.oceaneng.2025.120936
11. Li, X., Lu, X.: An approximate cost recovery scheme for the k-product facility location game with penalties. Theoretical Computer Science, vol. 1021, pp. 114933 (2024) doi: https://doi.org/10.1016/j.tcs.2024.114933
12. Mimun, H. A., Quattropani, M., Scarsini, M.: Best-response dynamics in two-person random games with correlated payoffs. Games and Economic Behavior, vol. 145, pp. 239–262 (3 2024) doi: 10.1016/j.geb.2024.03.011
13. Monderer, D., Shapley, L. S.: Potential games. Games and Economic Behavior, vol. 14, no. 1, pp. 124–143 (5 1996) doi: 10.1006/game.1996.0044
14. Nisan, N., Roughgarden, T., Tardos, E., Vazirani, V. V.: Algorithmic Game Theory: Computing in Games. Cambridge University Press (1 2007), http://ebooks.cambridge.org/chapter.jsf?bid= CBO9780511800481cid= CBO9780511800481A011
15. Razaque, A., Beishenaly, A., Kalpeyeva, Z., Uskenbayeva, R., Nikolaevna, M. A.: A reinforcement learning and predictive analytics approach for enhancing credit assessment in manufacturing. Decision Analytics Journal, vol. 15, pp. 100560 (2025) doi: https://doi.org/10.1016/j.dajour.2025.100560
16. Sutton, R. S., Barto, A. G.: Reinforcement learning: An introduction. The MIT Press (2020)
17. Swenson, B., Murray, R., Kar, S.: On Best-Response Dynamics in Potential Games. SIAM Journal on Control and Optimization, vol. 56, no. 4, pp. 2734–2767 (1 2018) doi: 10.1137/17m1139461
18. Swenson, B., Murray, R., Kar, S., Poor, H. V.: Best-response dynamics in continuous potential games: Non-convergence to saddle points. 2018 52nd Asilomar Conference on Signals, Systems, and Computers, pp. 310–315 (Oct 2018) doi: 10.1109/acssc.2018.8645541
19. Yang, K., Huang, C., Dai, Q., Yang, J.: The effects of attribute persistence on cooperation in evolutionary games. Chaos Solitons and Fractals, vol. 115, pp. 23–28 (8 2018) doi: 10.1016/j.chaos.2018.08.018
20. Yu, J., Zhang, Y., Sun, C.: Balance of exploration and exploitation: Non-cooperative game-driven evolutionary reinforcement learning. Swarm and Evolutionary Computation, vol. 91, pp. 101759 (11 2024) doi: 10.1016/j.swevo.2024.101759
21. Yu, P., Liu, S., Jin, C., Gu, R., Gong, X.: Optimization-based spectral end-to-end deep reinforcement learning for equity portfolio management. Pacific-Basin Finance Journal, vol. 91, pp. 102746 (2025) doi: https://doi.org/10.1016/j.pacfin.2025.102746
22. Zhu, Y., Xia, C.: Asynchronous best-response dynamics of networked anti-coordination game with payoff incentives. Chaos Solitons and Fractals, vol. 172, pp. 113503 (5 2023) doi: 10.1016/j.chaos.2023.113503

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

http://rcs.cic.ipn.mx

Centro de Investigación
en Computación