Fecha de recepción: 06 de marzo de 2026, fecha de publicación en línea: junio de 2026.

Visualizando la Toma de Decisiones de una Red Convolucional Mediante Mapeo de Activación de Clases Ponderado por Gradiente

Carlos. Pérez-Corona[1], J. Federico Ramírez-Cruz[1], Rocio Ochoa-Montiel[2], B. Estela Pedroza-Méndez[1], y Yesenia N. González-Meneses[1].

[1] Tecnológico Nacional de México – Instituto Tecnológico de Apizaco. San Andrés Ahuashuatepec, Municipio de Tzompantepec, Tlaxcala, C.P. 90491, México.
[2] Facultad de Ingeniería y Tecnología de la Universidad Autónoma de Tlaxcala. Calzada Apizaquito S/N. Apizaco, Tlaxcala, C.P. 90300, México.

Autor de correspondencia: Autor. Carlos Pérez-Corona (correo electrónico: carlos.pc@apizaco.tecnm.mx).

Abstract- The success of deep learning in many kinds of applications has been observed; in the case of computer vision, there are applications such as classification, object detection, face recognition, etc; however, these kinds of artificial intelligence systems are not transparent to users, which means that we don´t know how they make their decisions to support their answers. The lack of interpretability and explainability in black-box models represents a significant challenge in modern artificial intelligence systems, such as convolutional neural networks (CNNs). This paper explores the behavior of Gradient-weighted Class Activation Mapping (Grad-CAM) on the MNIST dataset, whose content is manuscript digit images in gray scale. The objective is to identify which input image regions have the greatest influence when the CNN classifies that input image, providing a visual interpretation that facilitates understanding of the classifier's behavior. Experimental results show that Grad-CAM highlights the relevant stroke areas in each digit image, which helps to detect ambiguity and allows correct misclassification.

Keywords: CNN, Grad-CAM, MNIST, XAI.

I. INTRODUCCIÓN

La inteligencia artificial (IA) ha logrado avances significativos en los ´últimos años, especialmente en el campo del aprendizaje profundo (DL). Sin embargo, uno de los principales desafíos actuales radica en la falta de interpretabilidad de estos modelos, comúnmente considerados como "cajas negras". En aplicaciones criticas como la medicina, la seguridad o la educación, comprender cómo estos modelos toman decisiones para una predicción resulta importante de tal forma que un usuario mantenga su confianza en estos modelos [1, 2]. Así, por ejemplo, se desearía que un modelo DL tenga capacidad de explicabilidad de acuerdo con su atención a los rasgos de una célula para clasificarla como cancerígena o no.

De acuerdo con [2] el propósito de la Inteligencia Artificial Explicable (XAI) es diseñar métodos que permitan entender, visualizar y justificar las decisiones de los modelos de IA, brindando confianza a los usuarios o facilitando la validación de resultados.

En este trabajo se presenta un estudio sobre la explicabilidad de una red neuronal convolucional (CNN) entrenada con el conjunto de datos MNIST, utilizando la técnica de mapeo de activaciones de clases ponderado por gradiente (Grad-CAM). El objetivo principal es visualizar las regiones de las imágenes que más influyen en la clasificación, mostrando de forma intuitiva qué "observa" la CNN para realizar sus predicciones.

II. FUNDAMENTOS Y TRABAJOS RELACIONADOS

A. Redes Neuronales Convolucionales

Las redes neuronales convolucionales (CNN) se han consolidado como una de las arquitecturas más eficaces en el ámbito de la visión por computadora, destacando en tareas de clasificación, detección de objetos y descripción de escenarios, entre otros. Estos modelos han mostrado capacidad para aprender jerarquías de características visuales desde niveles bajos hasta representaciones abstractas, lo que los ha convertido en una herramienta fundamental en el análisis de datos visuales [3].

Las CNNs operan primero extrayendo características aplicando pequeños filtros (kernels) sobre la imagen, los cuales se desplazan por toda la superficie de la imagen para identificar patrones básicos como bordes, contornos y texturas, cada uno de estos filtros recaba una parte de la información de la imagen y las coloca en una siguiente capa filtrada (capa convolucional), conforme la información

avanza a través de cada una de sus capas convolucionales, las características resultan cada vez más abstractas y que pueden representar partes de un objeto. Posterior a una capa convolucional puede encontrarse otra capa convolucional o una capa de agrupación (pooling), esta última con la intención de reducir dimensionalidad y para evitar sobreajuste en el proceso de aprendizaje. Una variante de agrupación obtiene la activación máxima de un subgrupo (disjunto) de activaciones en la capa convolucional (max-pooling), otra variante obtiene la activación promedio de ese subconjunto (average-pooling). El conjunto de capas de entrada, convolucionales y de agrupación tiene la tarea de extraer características en forma jerárquica. Posteriormente, las características últimas extraídas son combinadas a través de capas totalmente conectadas (indicadas como dense en la figura 1) para formar representaciones más complejas, permitiendo reconocer formas u objetos presentes en la imagen o escena. Este proceso facilita que la CNN aprenda automáticamente las características relevantes, lo que ha sido una de las principales razones de su éxito en diversas aplicaciones. Finalmente, una CNN suele ser menos sensibles a variaciones menores como iluminación o transformaciones dentro de una imagen [5, 6].

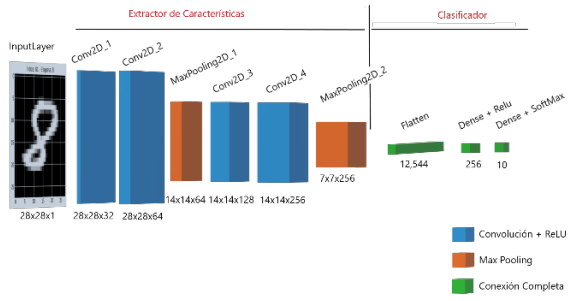

Figura 1. Arquitectura en bloques de la red CNN.

B. Inteligencia Artificial Explicable

De la preocupación de la XAI, antes ya explicada, surgen varios modelos, que según [2] pueden categorizarse de la siguiente forma: 1) los que modifican la arquitectura de la CNN, 2) los que simplifican la arquitectura de una CNN, 3) modelos XAI que se basan sobre la importancia de las características con respecto a la predicción, y 4) modelos XAI que interpretan a las CNN de manera visual. Los modelos XAI que realizan interpretaciones visuales permiten comprender qué partes de la imagen influyen más en las predicciones de una CNN, entre ellos se tienen los modelos CAM [7], Grad-CAM [8], Saliency Maps [9] o Deconv [10], entre otros varios. Estas herramientas permiten estudiar de manera más detallada el comportamiento interno de las CNN y evaluar la fiabilidad de sus decisiones en distintos contextos.

C. MNIST

Mediante herramientas como Keras y TensorFlow puede implementarse diversas arquitecturas de redes profundas y a la vez, acceder a diversos conjuntos de datos, tal como MNIST, propuesto por LeCun et al. [4]. La popularidad de MNIST se debe a que ofrece un problema relativamente sencillo para experimentar con arquitecturas iniciales, pero a la vez suficientemente representativo como para analizar el comportamiento de modelos más avanzados, convirtiéndose en un punto de referencia en el área de aprendizaje profundo (DL) y en visión por computadora (CV). En la siguiente sección se reportan las características más específicas de este conjunto de datos.

D. Grad-CAM

Se entiende por "Grad-CAM" como Mapa de Activación de Clases ponderado por Gradiente, por sus siglas en inglés (Gradient weighted Class Activation Map). El uso de técnicas basadas en gradientes es especialmente relevante porque el gradiente actúa como una medida de sensibilidad dentro del modelo de IA. En términos simples, el gradiente indica qué tan importante es cada parte de la entrada, es decir, qué píxeles son más importantes dada una imagen para la decisión final de la red. En métodos como Integrated-Gradients [11, 12], cuando un pequeño cambio ocurre en algún píxel puede producir una variación significativa en la predicción, el gradiente asociado a ese píxel será mayor, lo que significa que esa región tiene un impacto directo en la clasificación [8].

Durante el entrenamiento de una CNN, estos gradientes se utilizan para ajustar los pesos mediante el proceso de retropropagación (BackPropagation), permitiendo que el modelo aprenda patrones cada vez más precisos. Sin embargo, en tareas de interpretabilidad, el gradiente también se aprovecha con un objetivo diferente: analizar qué partes de la imagen de entrada influyen más en la salida del modelo. Técnicas como Grad-CAM [8] e Integrated-Gradients [11, 12] utilizan esta información para generar mapas visuales que resaltan las zonas más relevantes para la decisión de la red, proporcionando una explicación intuitiva del comportamiento interno de una CNN.

En [8] se describe la teoría que soporta a Crad-CAM y se resume a continuación. Para obtener un mapa de localización discriminante de clase Grad-CAM, $L^c_{Grad-CAM} \in \mathbb{R}^{u \times v}$, de ancho u y alto v para cualquier clase c, primero se calculan los gradientes de las puntuaciones para la clase c, y^c, con respecto a las activaciones del mapa de características A^k de una capa convolucional: $(\partial y^c)/(\partial A^k)$. Estos gradientes (que fluyen de regreso) se agrupan mediante un promedio global considerando las dimensiones alto y ancho (indexadas por i, j respectivamente) para obtener las ponderaciones de "importancia" del mapa de características k, para una clase objetivo c, α_k^c :

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{1}$$

Posteriormente se realiza una combinación ponderada de mapas de activación hacia adelante seguida por la aplicación de una función ReLU ($f(x) = \max(0, x)$) para obtener:

$$L^c_{Grad-CAM} = ReLU\left(\sum_k \alpha^c_k A^k\right) \quad (2)$$

Según [8], la aplicación de ReLU a la combinación lineal de mapas ayuda a obtener las características que influyen positivamente en la clase de interés, es decir, pixeles cuya intensidad debería ser incrementada con fin de incrementar también a y^c. Es probable que los pixeles negativos pertenezcan a otras categorías de la imagen de entrada.

III. METODOLOGÍA

A. Obtención del Conjunto de Datos

MNIST está conformado por 60,000 imágenes para entrenamiento y 10,000 para pruebas. Cada imagen está en escala de grises con tamaño 28x28 pixeles, y representa un dígito manuscrito junto con etiquetas del 0 al 9. Este conjunto de datos puede ser accedido mediante la librería Keras.

B. Arquitectura de la Red CNN

Para el modelado, compilación y entrenamiento de la CNN se utilizó las librerías de TensorFlow y Keras. La arquitectura de la CNN se puede observar en la figura 1 y está conformada por las siguientes capas en detalle:

- Capa de entrada con forma 28×28×1.
- Primer bloque: dos capas Conv2D con 32 y 64 filtros de tamaño 3×3 y activación ReLU, seguidas de una capa MaxPooling2D de 2×2.
- Segundo bloque: dos capas Conv2D con 128 y 256 filtros de tamaño 3×3 y activación ReLU, seguidas de una capa MaxPooling2D de 2×2.
- Capa Flatten para aplanar las características extraídas.
- Capa Dense de 256 neuronas con activación ReLU.
- Capa de salida Dense con 10 neuronas y activación Softmax, correspondiente a las diez clases de dígitos.

El modelo se compiló con el optimizador Adam y la función de pérdida categorical cross-entropy. Fue entrenado durante 50 épocas, alcanzando una precisión promedio del 99% sobre el conjunto de prueba; en la figura 2, se puede observar la convergencia de la exactitud y la función de pérdida, durante el entrenamiento y validación.

A. Implementación de Grad-CAM

Cómo se detalló anteriormente, Grad-CAM calcula la importancia de cada mapa de características usando los gradientes del modelo. Los gradientes se promedian espacialmente y se utilizan como pesos sobre las activaciones, generando un mapa de calor que muestra las regiones que más influyeron en la decisión del modelo. Posteriormente, este mapa se reescala al tamaño original de la imagen y se superpone con ella para su visualización.

IV. RESULTADOS

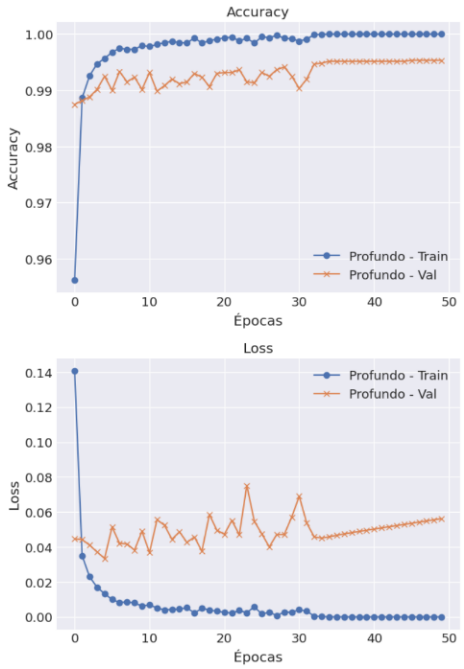

Figura 2. Exactitud y pérdida durante el entrenamiento.

Los principales resultados los podemos resumir como sigue: El modelo logró un desempeño general del 99% de precisión y la función de perdida disminuyó de forma estable, sin indicios de sobreajuste (ver figura 2).

Las imágenes generadas mediante Grad-CAM mostraron que la red presta atención principalmente a los trazos centrales de los dígitos. Por ejemplo, el dígito ”1”, el mapa se concentró en el trazo vertical (ver figura 3), en el dígito ”7” en la línea horizontal y vertical (figura 6)

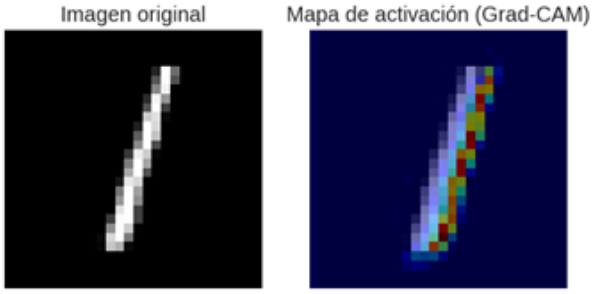

Figura 3. Resultado de Grad-CAM con la imagen del dígito “1”.

Estas observaciones permiten confirmar que el modelo basa sus predicciones en patrones visualmente coherentes con los que usaría un ser humano. Sin embargo, Grad-CAM, si bien es intuitiva tiene limitaciones, si contrastamos las figuras 6 y 7 puede observarse que su resolución espacial depende de la capa convolucional elegida y no siempre refleja con precisión la contribución de cada píxel, esto último puede verse en la figura 4 donde existe algo de ruido alrededor del dígito ”8”.

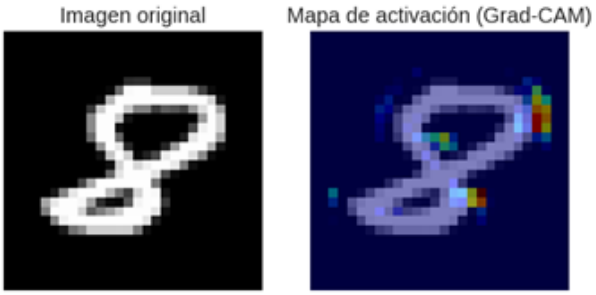

Figura 4. Resultado de Grad-CAM con la imagen del dígito "8".

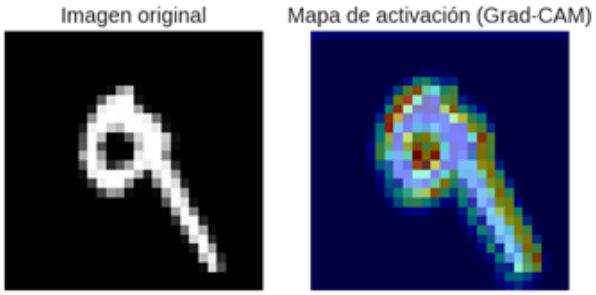

Figura 5. Resultado de Grad-CAM con la imagen del dígito "9".

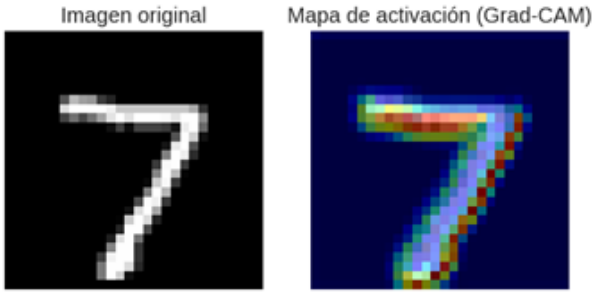

Figura 6. Resultado de Grad-CAM con la imagen del dígito "7" en la segunda capa convolucional.

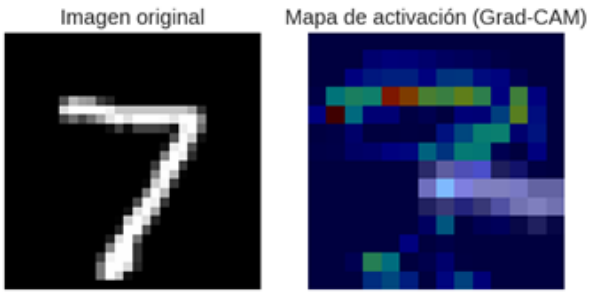

Figura 7. Resultado de Grad-CAM con la imagen del dígito "7" en la cuarta capa convolucional.

V. CONCLUSIONES

Nuestra experimentación muestra que Grad-CAM es una herramienta útil para dotar de explicabilidad a redes neuronales convolucionales. Las visualizaciones generadas permiten comprender de manera cualitativa cómo la red identifica los rasgos más representativos de cada dígito, ayudando a detectar errores de reconocimiento y poder mejorar la interpretabilidad del modelo. En trabajos futuros se planea aplicar la misma metodología con **a)** conjuntos de datos más complejos tal como Fashion-MNIST o CIFAR-10, **b)** utilizar modelos convolucionales pre-entrenados bien conocidos en el estado del arte, tales como VGG o ResNet, **c)** realizar comparaciones entre diferentes técnicas de explicabilidad para evaluar su fidelidad y coherencia visual, por ejemplo, y **d)** queda pendiente analizar e instrumentar las métricas cuantitativas de evaluación para los modelos XAI, pues existen trabajos con diferentes propósitos a evaluar según la aplicabilidad del modelo profundo; algunas métricas, según la literatura [13, 14], están orientadas a la fidelidad, la localización, conteo de falsos positivos o comprobación de sensibilidad.

REFERENCIAS

[1] A. B. Arrieta et al., "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI", arXiv.org, 2019, https://arxiv.org/abs/1910.10045.

[2] Rami Ibrahim and M. Omair Shafiq. 2023, Explainable Convolutional Neural Networks: A Taxonomy, Review, and Future Directions. ACM Comput. Surv. 55, 10, Article 206 (October 2023), p. 37.

[3] Q. Zhang, X. Wang, Y. N. Wu, H. Zhou, and S.-C. Zhu, "Interpretable CNNs for Object Classification", arXiv.org, 2019. https://arxiv.org/abs/1901.02413.

[4] Y. LeCun et al., "Gradient-based learning applied to document recognition", Proc. IEEE, vol. 86, no. 11, pp. 2278–2324, 1998.

[5] F. Chollet, "Visualizing what convnets learn", Keras Documentation, 2023. [En línea]. Disponible en: https://keras.io/examples/vion/visualizing/hat/convnets/earn/.

[6] J. Brownlee, "How to Visualize Filters and Feature Maps in Convolutional Neural Networks", Machine Learning Mastery, 2021. [En línea]. Disponible en: https://machinelearningmastery.com/how-to-visualize-filters-and-feature-maps-in-convolutional-neural-networks/.

[7] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva and A. Torralba, "Learning Deep Features for Discriminative Localization", 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 2921-2929, doi: 10.1109/CVPR.2016.319.

[8] R. R. Selvaraju et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization", Proc. IEEE Int. Conf. Computer Vision (ICCV), 2017, pp. 618–626.

[9] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps", arXiv.org, 2013. https://arxiv.org/abs/1312.6034.

[10] M. D. Zeiler, D. Krishnan, G. W. Taylor and R. Fergus, "Deconvolutional networks", 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 2010, pp. 2528-2535, doi: 10.1109/CVPR.2010.5539957.

[11] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic Attribution for Deep Networks," arXiv:1703.01365, Enero. 2017, Disponible en: *https://arxiv.org/abs/1703.01365.*

[12] F. Chollet, “Integrated Gradients”, Keras Documentation, 2023. [En línea]. Disponible en: *https://keras.io/examples/vion/integratedgradients/.*

[13] X.-H. Li *et al.*, “Quantitative Evaluations on Saliency Methods: An Experimental Study”, 2020. [En línea]. Disponible en: *https://arxiv.org/abs/2012.15616.*

[14] A. Dugăeșescu y A. M. Florea, "Evaluation and analysis of visual methods for CNN explainability: a novel approach and experimental study," Neural Computing and Applications, vol. 37, pp. 14935–14970, Mayo 2025. [En línea]. Disponible en: *https://link.springer.com/article/10.1007/s00521-025-11282-7.*

CARLOS PÉREZ-CORONA Profesor en el Instituto Tecnológico Nacional de México (TecNM) – Campus Apizaco. Recibió el título de licenciatura en informática por parte del mismo instituto en 1994. El título de maestro en inteligencia artificial le fue otorgado por la Universidad Veracruzana, en Xalapa, Veracruz en 1999, obteniendo mención honorifica. Actualmente realiza su doctorado en ciencias de la ingeniería en el TecNM/Instituto Tecnológico de Apizaco. Su área de interés se ha centrado en el aprendizaje máquina, las redes neuronales artificiales, aprendizaje profundo y en la inteligencia artificial explicable (XAI).

J. FEDERICO RAMÍREZ-CRUZ Recibió su grado de Licenciatura en Ingeniería Industrial en Electrónica en el Instituto Tecnológico de Puebla en 1993, la Maestría en Ciencias con Especialidad en Electrónica en el Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE) en 1994, y el grado de Doctorado en Ciencias en el Área de Ciencias Computacionales en el (INAOE) en 2003. Actualmente, es Profesor-Investigador en el TecNM/Instituto Tecnológico de Apizaco. Sus intereses de investigación se centran en el área de aplicaciones del aprendizaje automático, computación evolutiva, visión por computadora y robótica inteligente.

ROCIO OCHOA-MONTIEL Es Ing. en Computación y Maestra en Ciencias en Ingeniería en Computación por la Universidad Autónoma de Tlaxcala. En 2022 obtuvo el grado de Doctora por el Centro de Investigación en Computación del Instituto Politécnico Nacional, y pertenece al Sistema Nacional de Investigadores en el área de Ingenierías y Desarrollo Tecnológico. Es docente-investigador en la Facultad de Ciencias Básicas, Ingeniería y Tecnología de la Universidad Autónoma de Tlaxcala, es miembro de la red Temática de Inteligencia Computacional Aplicada de CONACYT desde 2016, es integrante de la Academia Mexicana de Computación (AMEXCOMP), y de la Asociación Mexicana de Sistemas Dinámicos y Complejidad (AMESDYC). Sus principales líneas de investigación son la visión por computadora, el cómputo evolutivo y bio-inspirado; el reconocimiento de patrones y aprendizaje automático aplicado al área médica.

ESTELA PEDROZA-MENDEZ obtuvo el grado de doctor en la Facultad de Ciencias de la Electrónica de la Benemérita Universidad Autónoma de Puebla (BUAP) en 2018. Actualmente es docente investigador del área de posgrado y de licenciatura en el Tecnológico Nacional de México, campus Apizaco. Tiene el reconocimiento como Perfil PRODEP, hasta el 2026 y el reconocimiento en el SNII, en el nivel 1, hasta diciembre del 2029. Sus áreas de investigación incluyen modelos de lógica difusa, tutoriales inteligentes y aplicaciones de internet de las cosas.

YESENIA N. GONZÁLEZ-MENESES obtuvo el grado de doctor en la Facultad de Ciencias de la Electrónica de la Benemérita Universidad Autónoma de Puebla (BUAP) en 2020. Actualmente es docente investigador del área de posgrado y de licenciatura en el Tecnológico Nacional de México, campus Apizaco. Sus áreas de interés se centran en los tutoriales inteligentes y aplicaciones.