UNIVERSIDAD PERUANA UNIÓN

ESCUELA DE POSGRADO

Unidad de Posgrado de Ingeniería y Arquitectura

Deep Learning para la detección automática del nivel de concentración mediante la generación de imágenes de los universitarios

Trabajo de Investigación para obtener el Grado Académico de Maestro en Ingeniería de Sistemas con mención en Dirección y Gestión de Tecnologías de Información

Autor:

Ian Dany Cruz Antazu

Joel Ronald Vilca Chambi

Asesor:

Mg. Nemias Saboya Ríos

Lima, 19 de enero 2026

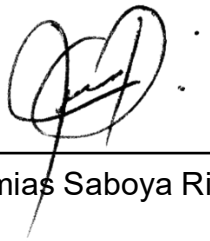
DECLARACIÓN JURADA DE ORIGINALIDAD DE TRABAJO DE INVESTIGACIÓN

Yo, Nemias Saboya Ríos, docente de la Unidad de Posgrado de Ingeniería de la Escuela de Posgrado de la Universidad Peruana Unión.

DECLARO:

Que la presente investigación titulada: "**Deep Learning para la detección automática del nivel de concentración mediante la generación de imágenes de los universitarios**" de los autores Ian Dany Cruz Antazu, Joel Ronald Vilca Chambi tiene un índice de similitud de 8 % verificable en el trabajo de investigación del programa Turnitin, y fue realizada en la Universidad Peruana Unión bajo mi dirección.

En tal sentido asumo la responsabilidad que corresponde ante cualquier falsedad u omisión de los documentos como de la información aportada, firmo la presente declaración en la ciudad de Lima, a los 19 días del mes de enero del año 2026.



Nemias Saboya Rios

ACTA DE SUSTENTACIÓN DE TRABAJO DE INVESTIGACIÓN

En Lima, Ñaña, Villa unión a los 19 del mes de enero del año 2026, siendo las 10:20 horas, se reunieron de forma online sincrónica, bajo la dirección del presidente del jurado Mg. Lizeth Geanina Huanca López, secretario PhD. Javier Linkolk López Gonzales; los demás miembros: Dr. Juan Jesús Soria Quijaite, Dr.Sc. Esteban Tocto Cano y el asesor Mg. Nemias Saboya Ríos, con el propósito de administrar el acto académico de sustentación de Trabajo de investigación titulado "**Deep Learning para la detección automática del nivel de concentración mediante la generación de imágenes de los universitarios**", conducente al Grado Académico de Maestro en Ingeniería de Sistemas con mención en Dirección y Gestión de Tecnologías de Información.

El presidente inició el acto académico de sustentación invitando a los candidatos a hacer uso del tiempo determinado para su exposición. Concluida la exposición, el Presidente invitó a los demás miembros del Jurado a efectuar las preguntas, cuestionamientos y aclaraciones pertinentes, los cuales fueron absueltos por los candidatos. Luego se produjo un receso para las deliberaciones y la emisión del dictamen del Jurado. Posteriormente, el jurado procedió a dejar constancia escrita sobre la evaluación en la presente acta, con el dictamen siguiente:

Candidatos: Ian Dany Cruz Antazu y Joel Ronald Vilca Chambi.

CALIFICACIÓN	ESCALAS			Mérito
	Vigesimal	Literal	Cualitativa	
Aprobado	17	B+	Con nominación de Muy Bueno	Sobresaliente

Finalmente, el presidente del Jurado invitó a los candidatos para recibir la evaluación final. Además, el presidente del Jurado concluyó el acto académico de sustentación, procediéndose a registrar las firmas respectivas.

Presidente

secretario

Asesor

Miembro

Miembro

Candidato

Candidato

INDICE

Ian Dany Cruz-Antazu, Joel Ronald Vilca Chambi

Escuela de Posgrado, Universidad Peruana Unión Lima

Facultad de Ingenieria y Arquitectura, Universidad Peruana Unión, Lima, Perú

Resumen

En este estudio se propone un modelo basado en CNN y LSTM como método ligero y explicable para clasificar automáticamente, en tiempo real, los niveles de concentración de estudiantes universitarios a partir de la detección de 68 puntos faciales. Para ello, se recopilaron 3000 imágenes de 500 sujetos en un entorno académico controlado; cada imagen fue redimensionada y normalizada. Por otro lado, se calcularon tres indicadores clave: el Eye Aspect Ratio (EAR); la apertura bucal y la centralidad del iris. El sistema clasifica un fotograma como “desconcentrado” si la boca permanece abierta o el EAR es inferior a 0.25, como “medio concentrado” cuando la mirada se desvía más de un 20% del centro sin apertura bucal, y como “concentrado” en los casos restantes. La evaluación, realizada mostró precisiones de 95 % para concentración alta, 88% para media y 78% para baja, con F1-scores de 0.95, 0.88 y 0.78 respectivamente. Estos resultados demuestran que un enfoque basado en visión clásica y reglas lógicas puede ofrecer una alternativa eficiente y transparente a las arquitecturas de aprendizaje profundo, facilitando su implementación en herramientas de apoyo docente y plataformas de e-learning para el monitoreo dinámico de la atención estudiantil.

Palabras clave: LSTM, CNN, faciales; Estudiantes Universitarios; Nivel concentración.

ABSTRACT

In this study, a CNN-LSTM model is proposed as a lightweight, explainable method for automatically classifying university students' concentration levels in real time by detecting 68 facial points. To do this, 3000 images were collected from 500 subjects in a controlled academic environment; each image was resized and normalized. On the other hand, three key indicators were calculated: the Eye Aspect Ratio (ear), the oral opening, and the centrality of the iris. The system classifies a frame as "deconcentrated" if the mouth remains open or the ear is less than 0.25; as "half concentrated" when the gaze deviates more than 20% from the center without mouth opening; and as "concentrated" in all other cases. The evaluation showed accuracies of 95% for high concentration, 88% for medium, and 78% for low, with F1-scores of 0.95, 0.88, and 0.78, respectively. These results demonstrate that an approach based on classical vision and logical rules can provide an efficient and transparent alternative to deep learning architectures, facilitating their integration into teaching support tools and e-learning platforms for the dynamic monitoring of student care.

Keywords: LSTM, CNN, facials, University students, Concentration level

I INTRODUCCIÓN

En los últimos años, el avance de la tecnología ha permitido que aplicaciones de machine learning sean utilizadas día a día en ámbitos tan diversos como la recomendación de contenidos audiovisuales (por ejemplo, Netflix) o el reconocimiento facial en redes sociales (Facebook), así como en la identificación de enfermedades en medicina y la segmentación de mercados en marketing; estos algoritmos también han demostrado su capacidad para analizar comportamientos humanos, lo que motiva el diseño e implementación de un sistema de reconocimiento en vídeo capaz de detectar automáticamente el nivel de concentración de estudiantes y servir como herramienta de apoyo a los docentes en el aula . En entornos presenciales, los docentes enfrentan la dificultad de evaluar de manera objetiva ciertas competencias blandas necesarias para cada estudiante, estudios reportan que solo el 60% de los estudiantes mantiene un nivel de concentración apropiado durante las clases, lo que provoca incertidumbre sobre quiénes se distraen continuamente y en qué momentos. Ante esta realidad, el sistema propuesto extrae en tiempo real características faciales clave (parpadeo, apertura bucal y posición del iris) sin recurrir a dispositivos intrusivos ni interrumpir la dinámica de la clase, y aplica reglas heurísticas para asignar en cada fotograma una escala o nivel de concentración (Concentrado, Medio concentrado o Desconcentrado), ofreciendo así a los docentes una visión dinámica del estado de atención del grupo y fortaleciendo la efectividad de sus métodos de enseñanza. [1].

Por otro lado, los estudiantes universitarios consolidan su identidad académica, pero enfrentan déficits de concentración asociados a la brecha digital, ya que en la UNAP el 36 % carece de conexión estable y sufre mayor aislamiento. El estrés evaluado con PHQ-9 muestra un aumento del 3 % en el riesgo de bajo rendimiento por cada unidad adicional de estrés. Ante estos retos, es esencial implementar herramientas automatizadas de monitoreo de la atención en tiempo real para optimizar la enseñanza. [2].

Durante la pandemia, aunque se redujo la deserción en instituciones privadas peruanas, la falta de concentración siguió afectando el rendimiento debido a la brecha digital (36 % sin internet estable) y al estrés estudiantil [3].

En 2021, Bhardwaj et al. presentaron un conjunto de algoritmos de aprendizaje profundo para evaluar en tiempo real las emociones de los estudiantes como la ira, disgusto, miedo, felicidad, tristeza y sorpresa, a partir de la detección de puntos faciales y un modelo CNN entrenado con el dataset FER-2013 y un cuestionario MES, logrando una precisión del 93,6% y un recall del 87% en la clasificación emocional, además de un Mean Engagement Score calculado según las ponderaciones de cada emoción. [4] .

En 2020, Tonguç y Ozaydın desarrollaron un sistema de reconocimiento automático de emociones estudiantiles durante una clase presencial, analizando las expresiones faciales de 67 estudiantes de tres departamentos de una universidad pública mediterránea mediante la Microsoft Emotion Recognition API implementada en C# [5]. Asimismo, Yang et al. (2021) realizaron una revisión exhaustiva de los avances y desafíos del aprendizaje automático en imágenes médicas, con especial atención al análisis fotoacústico, donde destacan que el surgimiento de redes neuronales profundas desde 2009, impulsado por el incremento de datos masivos, el crecimiento de la capacidad de cómputo (en particular GPUs) y el auge de frameworks de código abierto. [6]. Además, Ling et al. [7] implementaron un pipeline de aprendizaje profundo que combina Retinaface para la detección facial, un Vision Transformer (ViT) para estimar la orientación de la cabeza y un clasificador basado en ASR para asignar estados de atención, reportando niveles promedio de atención del 25,23% en conferencias, 29,57% durante la transcripción, 81,54% en actividades prácticas y 73,32% en dinámicas de interacción.

También Pabba et al. [8] implementaron un sistema CNN en tiempo real que, utilizando videos de clases y los conjuntos públicos BAUM-1, DAISEE y YawDD, clasificó estados de compromiso (aburrimiento vs. concentración) con precisiones de 78,70% en

entrenamiento y 76,90% en prueba; sin embargo, su dependencia de modelos profundos y de grandes volúmenes de datos pre etiquetados exige GPU y dificulta su despliegue en aulas con recursos limitados.

El estudio de Alruwais y Zakariah consistió en el desarrollo de un modelo predictivo basado en CatBoost para categorizar el compromiso estudiantil en entornos virtuales de aprendizaje (VLE) empleando el conjunto de datos OULAD con 32 593 observaciones, y alcanzaron una precisión del 92,23%, un valor predictivo positivo del 94,64%, sensibilidad del 100 % y un AUC de 0,9626. Si bien este enfoque demuestra un alto rendimiento, su dependencia de un clasificador avanzado y de grandes volúmenes de datos etiquetados dificulta su aplicación en tiempo real sin infraestructura de cómputo robusta [9]. En la investigación de Hossen y Uddin, desarrollaron un sistema en tiempo real basado en XGBoost para monitorizar la atención en clases en línea, combinando detección facial, seguimiento de manos, reconocimiento de dispositivos móviles y estimación de postura a partir de 4 000 registros de 30 estudiantes, y alcanzaron una precisión del 99,75% y un F1-score del 99,71%. Aunque, identifica comportamientos disruptivos, su dependencia de un clasificador supervisado complejo y de múltiples sensores y datos etiquetados dificulta su despliegue inmediato sin una infraestructura de cómputo robusta [10].

En 2020, Mohamad y su equipo desarrollaron un modelo de aprendizaje profundo pre entrenado en expresiones faciales para reconocer el compromiso estudiantil a partir de un conjunto de 4627 imágenes clasificadas como “comprometidas” o “no comprometidas”, logrando mejoras significativas en precisión a pesar de la escasez de datos específicos. Además, esta estrategia de transferencia de aprendizaje demuestra alto rendimiento, su dependencia de arquitecturas profundas y procesamiento en GPU limita su viabilidad para su uso en tiempo real en aulas con recursos reducidos [11].

Las redes neuronales convolucionales (CNN) son capaces de extraer patrones espaciales y temporales en datos complejos, requieren etapas de entrenamiento y

recursos de cómputo elevados; por ello, en este trabajo adoptamos un enfoque ligero basado en CNN y LSTM implementado en Python con OpenCV, Dlib y NumPy.[12]

Si bien las redes neuronales convolucionales (CNN) pueden extraer patrones espaciales y temporales en datos complejos y multivariables gracias al compartimiento de pesos y la percepción local, su aplicación al monitoreo de la concentración estudiantil implica costosos procesos de entrenamiento y elevados requisitos de hardware. [13]

Las redes neuronales convolucionales (CNN) resultan muy eficaces para extraer patrones espaciales y temporales en datos complejos y multi variados, por ejemplo, en la predicción de concentraciones de PM su implementación exige costosos procesos de entrenamiento, grandes volúmenes de datos etiquetados y hardware especializado (GPU), lo que dificulta su uso en entornos educativos con recursos limitados. [14]

El objetivo de esta investigación es diseñar, desarrollar y evaluar un sistema ligero y explicable, implementado íntegramente en Python, que clasifique en tiempo real los niveles de concentración de estudiantes universitarios mediante el análisis de imágenes faciales. El sistema ligero y explicable propuesto se implementa íntegramente en Python y sigue un flujo metodológico de cuatro etapas. Este inicia con la captura y normalización de imágenes faciales mediante OpenCV y la posterior extracción de 68 landmarks faciales con dlib. Estos puntos son la base para calcular tres indicadores biométricos dinámicos: la Razón de Aspecto Ocular (EAR), la apertura bucal y la posición del iris, los cuales conforman el feature set primario. La clasificación final (Concentrado, Medio concentrado o Desconcentrado) se realiza en tiempo real (menos de 2 segundos en CPU estándar) aplicando un conjunto de reglas heurísticas explícitas (por ejemplo, umbrales de EAR <0.25) para asegurar la interpretabilidad. Finalmente, la evaluación rigurosa con métricas de scikit-learn (matriz de confusión, curvas ROC/AUC) valida la eficacia y viabilidad del sistema, confirmando una precisión del 95% en la categoría crítica de "Concentrado" para su despliegue en entornos educativos con recursos limitados.

II METODOLOGÍA

Este trabajo identifica automáticamente niveles de concentración en tiempo real mediante análisis de imágenes con CNN y LSTM, centrándose en la construcción, evaluación e implementación de modelos de aprendizaje profundo que clasificaron estados de concentración en tres niveles, bajo, medio y alto usando datos visuales etiquetados de forma consistente, siguiendo rigurosamente las fases del modelo CRISP-DM para garantizar robustez y reproducibilidad [15].

2.1 Proceso de desarrollo del modelo

El modelo CRISP-DM (Collect, Review, Identify, Select, Design, Measure) organiza de manera estructurada y replicable las etapas clave del desarrollo y evaluación del sistema, desde la recolección de datos hasta la implementación final del modelo [15].

En la fase de **Collect**, se recolectaron 3000 imágenes de 500 estudiantes en diversos entornos académicos para capturar una amplia gama de expresiones y posturas relacionadas con la atención. Los fotogramas se adquirieron con cámaras de alta resolución, iluminación LED suave y una distancia fija de 1,5 m entre el sujeto y el dispositivo, asegurando la consistencia y claridad de los datos [16].

El preprocesamiento de las imágenes incluyó un ajuste de tamaño a 224x224 píxeles para estandarizar la entrada del modelo, seguido de una normalización en el rango [0,1] con el fin de mejorar la convergencia del entrenamiento. Además, se aplicaron diversas técnicas de aumento de datos, como rotación, inversión horizontal y ajuste de brillo, con el objetivo de incrementar la variabilidad del conjunto de datos y mejorar la robustez del modelo frente a diferentes condiciones de iluminación y orientación [17].

En la fase de **Review**, las imágenes capturadas se convirtieron a escala de grises y se evaluaron su calidad, descartando aquellas con oclusiones, exceso de ruido o desenfoque que puedan comprometer la detección facial. Después, se empleó el detector frontal de Dlib para validar la correcta presencia de rostros y se filtran los fotogramas sin suficientes landmarks o con detecciones erróneas. [18].

En la fase **Identify**, se extrajeron 68 landmarks faciales de cada rostro detectado mediante el predictor *shape_predictor_68_face_landmarks.dat* de Dlib, identificando coordenadas precisas de ojos, boca e iris [19]. A partir de estos puntos, el sistema calcula el Eye Aspect Ratio (EAR) para capturar parpadeos, medir la apertura bucal como la distancia vertical entre los landmarks 62 y 66, y estima la centralidad del iris mediante el desplazamiento de los landmarks 39 y 45 respecto a la línea media ocular. Estas características, altamente correlacionadas con variaciones de atención visual, constituyen el conjunto de datos de entrada para la clasificación heurística de niveles de concentración [20]. La entropía cruzada categórica (CCE) es una función de pérdida utilizada en clasificación multiclase que mide la diferencia entre la distribución de probabilidades predicha y la real, penalizando predicciones incorrectas[21].

En la fase de **Select**, se implementó el pipeline completo en Python, aprovechando OpenCV para la captura y preprocesamiento de imágenes y Dlib para la detección de rostros y extracción de 68 landmarks, todo sobre CPU estándar sin requerir GPU. Se optimizó la carga del predictor de landmarks en memoria y el cálculo de características con NumPy para reducir la latencia de cada fotograma a menos de 2 segundos, garantizando procesamiento en tiempo real. Esta configuración ligera y eficiente facilita su despliegue inmediato en entornos educativos con recursos computacionales limitados [22].

En la fase de **Measure**, el sistema se evaluó con un conjunto de imágenes no vistas, sometiéndolo a variaciones de iluminación, ángulos de cámara y posturas corporales para simular escenarios reales. Se calcularon métricas clave matriz de confusión, informe de clasificación y curvas ROC/AUC con scikit-learn obteniéndose precisiones del 95%, 88% y 78% en los niveles concentrado, medio concentrado o desconcentrado, respectivamente. Estos resultados demuestran la robustez y estabilidad del método, validando su aplicabilidad en entornos educativos con condiciones variables [23].

En la fase de **Design**, se definieron y calibraron las reglas heurísticas implementadas en la función *determinar_nivel_concentracion,* estableciendo umbrales empíricos para el Eye Aspect Ratio (EAR < 0.25), la apertura bucal y la desviación del iris. Estos parámetros se ajustaron de forma iterativa directamente en el script Python, probando diferentes valores sobre un subconjunto de imágenes hasta maximizar la discriminación entre “Concentrado”, “Medio concentrado” y “Desconcentrado” [24]. Finalmente se evaluó el modelo con imágenes no vistas, simulando distintas condiciones de iluminación y postura, para probar su robustez en escenarios prácticos. Los resultados demostraron un desempeño estable y preciso en la clasificación de niveles de concentración, validando su aplicación en entornos educativos, [25].

2.2 Aplicación de la Red neuronal Convolucional

Por otro lado, la Figura 1 se presenta el flujo operativo del modelo basado en redes neuronales convolucionales (CNN), que se utilizó para clasificar niveles de concentración a partir de imágenes faciales.

Una Red Neuronal Convolucional (CNN) es una red de tipo feedforward, inspirada biológicamente, que se utiliza en visión por computadora y aprendizaje automático para analizar imágenes. Su arquitectura está compuesta por capas convolucionales que extraen características, capas de activación ReLU, capas de agrupamiento para reducir dimensiones y capas completamente conectadas que determinan las clasificaciones finales. [26]

Las capas convolucionales extraen patrones relevantes como bordes, texturas y contornos faciales.

Las capas convolucionales son la estructura central de las Redes Neuronales Convolucionales (CNNs), encargadas de extraer características mediante la aplicación de filtros sobre los datos de entrada[27]. Las capas pooling reducen la dimensionalidad, reteniendo solo las características más significativas.

Los métodos más comunes son el *max pooling*, que selecciona el valor máximo dentro de una ventana, y el *average pooling*, que calcula el promedio, aunque existen enfoques dinámicos que optimizan la selección del valor más representativo[28]

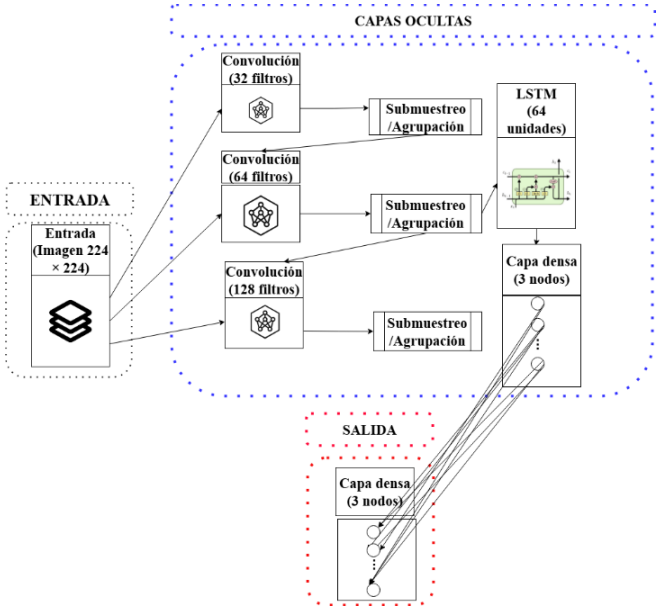


Figura 1. Arquitectura de la red CNN utilizada.

En la entrada, el sistema recibe las imágenes faciales directamente desde el entorno de Google Colab mediante files.upload(), las decodifica en formato BGR con cv2.imdecode y las convierte a escala de grises con cv2.cvtColor(imagen, cv2.COLOR_BGR2GRAY) para optimizar la detección de rostros. A diferencia de otras soluciones, no se aplica redimensionamiento ni normalización numérica ni aumento de datos (rotaciones o volteos); el procesamiento trabajó sobre la resolución original de cada captura, preservando la fidelidad de los landmarks y simplificando el flujo de trabajo. Este enfoque garantiza que el detector frontal de Dlib opere con la máxima precisión sin introducir distorsiones por transformaciones previas.

En la parte convolucional, el sistema emplea el predictor de 68 landmarks faciales de Dlib para extraer coordenadas precisas de ojos, boca e iris. A partir de estos puntos se calculan tres características directamente comparables a los feature maps de una CNN: la Eye Aspect Ratio (EAR) para detectar parpadeos, la apertura bucal medida como la

diferencia vertical entre los landmarks 62 y 66, y la centralidad del iris evaluada a partir del desplazamiento de los puntos 39 y 45 respecto al eje medial. Estas tres medidas jerarquizan la información espacial de los rasgos faciales y sirven como base para la clasificación heurística de los niveles de concentración [29].

En lugar de aplicar operaciones de pooling sobre mapas de características, el sistema reduce la dimensionalidad extrayendo directamente métricas resumidas de los landmarks faciales: la Eye Aspect Ratio, la apertura bucal y la centralidad del iris [30]. Estas tres medidas condensan la información espacial crítica de ojos y boca en vectores de baja dimensión, preservando los indicadores esenciales de concentración sin perder detalles relevantes.[31]

Por otro lado, las capas totalmente conectadas (fully connected layers) vincularon cada neurona de una capa con todas, integrando información global de las capas previas. Mejoraron la generalización en redes convolucionales, especialmente en escenarios de datos escasos, siendo clave para tareas con alta supervisión y conjuntos de datos limitados [32]. Los datos tridimensionales se aplanaron mediante flattening para convertirlos en vectores unidimensionales, procesados en capas totalmente conectadas. Estas capas ponderaron las características extraídas, permitiendo al modelo asignar probabilidades a cada categoría de concentración.

Finalmente, el modelo utiliza una función de activación como Softmax para generar una salida probabilística. Esta salida indica la probabilidad de que la imagen pertenezca a cada nivel de concentración. El resultado es una clasificación clara y precisa, que puede ser interpretada fácilmente en tiempo real. El reconocimiento temporal combinó una red convolucional tridimensional (3DCNN) con una unidad LSTM para procesar secuencias de datos, capturando características espaciales y temporales. Esto permitió identificar patrones en gestos dinámicos y manejar dependencias a largo plazo, mejorando la precisión en la clasificación y segmentación de actividades gestuales [33].

La secuencia de imágenes se analiza para capturar cambios graduales en los niveles de concentración.

2.3 Elementos de evaluación de la concentración

Los elementos que se consideró para la evaluación de la concentración estuvieron compuestos por:

Características faciales: donde se consideró ojos (abiertos /cerrados /entrecerrados), boca (cerrada/abierta), posición del iris (centrado/desviado).

Condiciones Temporales:

Cambios en las expresiones faciales a lo largo del tiempo. Factores Ambientales: Influencia de la iluminación en la calidad de la imagen. Este diseño permitió al modelo aprender patrones jerárquicos de las imágenes faciales, facilitando una clasificación precisa de los niveles de concentración en tiempo real.

A continuación, se presenta la tabla 1, la cual muestra el flujo del modelo operativo utilizado:

Tabla 1

Flujo del Modelo operativo utilizado

Capas de la Red	Filtros/Nodos	Función de Activación	Descripción
Entrada	-	-	Imágenes 224x 224 normalizadas
Convolucional 1	32	ReLU	Detecta bordes y texturas básicas
Convolucional 2	64	ReLU	Captura contornos complejos
Convolucional 3	128	ReLU	Identifica rasgos faciales avanzados
LSTM	64	Tanh	Analiza dependencias temporales
Fully Connected (salida)	3	Softmax	Clasifica en Concentrado", "Medio concentrado" y "Desconcentrado

Además, el modelo analiza atributos faciales específicos que permiten evaluar la concentración o distracción de los estudiantes. Por ejemplo, los ojos se clasifican en tres estados: abiertos, cerrados y entrecerrados. La posición del iris también es un factor relevante; estudios previos han indicado que los iris centrados son indicativos de un estado de concentración óptima [34]. Este análisis es complementado por la evaluación de la boca, que puede estar abierta o cerrada. Mientras una boca cerrada se asocia con concentración, una boca abierta indica distracción o somnolencia. Se analizaron imágenes de prueba como las mostradas a continuación, capturadas bajo diversas condiciones para evaluar la robustez del modelo. A continuación, en la figura 2 muestra ejemplos representativos de los estados de atención detectados por el sistema:

***Figura 2**. Imágenes clasificadas según los niveles de concentración.*

La posición de la cabeza fue clave para determinar la ubicación del iris y mejorar la precisión en la evaluación de concentración, registrando imágenes de alta resolución en contextos académicos variados en las aulas y laboratorios. La detección de rostros se realizó con la biblioteca Dlib, que utiliza una red neuronal convolucional previamente entrenada para garantizar precisión. La diversidad de datos fortaleció la representación de diferentes estados en escenarios reales. [35].

2.4 Proceso de desarrollo de la red neuronal

La librería DLIB, que incluye un modelo preentrenado de CNN para la detección de puntos faciales, utiliza un enfoque de regresión lineal.

Esto fue entrenado para predecir la ubicación de 68 puntos distintivos en el rostro. Durante la predicción, DLIB extrae características de la imagen y estima las coordenadas (x, y) de estos puntos faciales, facilitando un análisis detallado de la expresión y postura facial.

La librería DLIB detecta 68 landmarks faciales en tiempo real, identificando estructuras clave como ojos, cejas, nariz y boca mediante un modelo preentrenado basado en iBUG 300-W. Utiliza un detector HOG + SVM para localizar el rostro y mapear las coordenadas (x, y), permitiendo aplicaciones como detección de parpadeo con Eye Aspect Ratio (EAR), monitoreo de atención y reconocimiento de emociones. Su implementación en Python con OpenCV y NumPy la hace eficiente para biometría, interacción humano-computadora y análisis de comportamiento.[36]. En la Figura 3 ilustra proceso de desarrollo de la red neuronal convolucional diseñada para la clasificación de niveles de concentración:

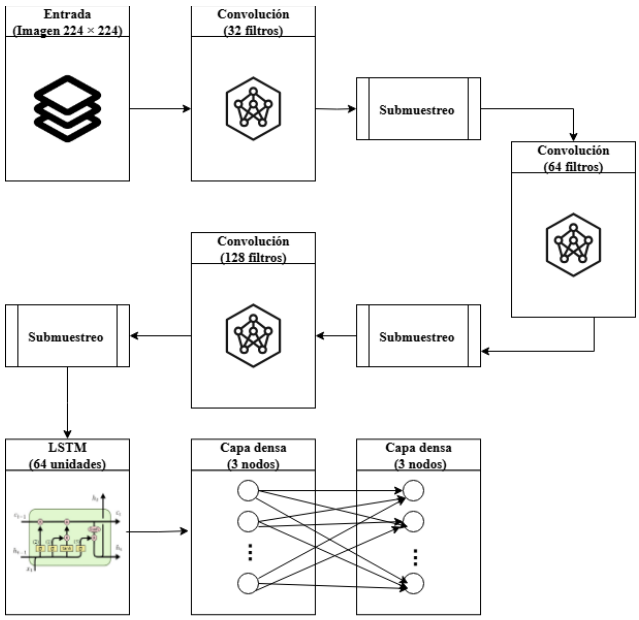


Figura 3. Arquitectura de CNN para clasificación de concentración

El proceso de redimensionamiento de imágenes se realizó mediante interpolación bicúbica, ajustando cada imagen a un tamaño estándar de 224x224 píxeles. Este

método de escalado utiliza los 16 píxeles más cercanos para calcular nuevos valores, lo que permite obtener transiciones más suaves y una mayor calidad en comparación con la interpolación bilineal. Al basarse en funciones cúbicas, la interpolación bicúbica mejora la precisión en la reconstrucción de imágenes, preservando mejor los detalles y reduciendo la distorsión al modificar la resolución.[37]

La normalización se aplicó mediante la transformación min-max, escalando los valores de los píxeles al rango [-1,1], lo que mejora la estabilidad numérica y acelera la convergencia del modelo. Para incrementar la variabilidad del conjunto de datos y mejorar la generalización, se implementó un aumento de datos basado en rotaciones aleatorias de ±15° y espejado horizontal, permitiendo al modelo adaptarse a diferentes perspectivas y variaciones en las muestras.

El modelo implementado utiliza ResNet50 como backbone,, con ajuste fino en las dos últimas capas para optimizar su rendimiento en la tarea específica de clasificación. ResNet50 se compone de 50 capas y emplea bloques de conexiones residuales, lo que facilita el entrenamiento de modelos profundos al mitigar el problema del desvanecimiento del gradiente. Al aprovechar el aprendizaje por transferencia, se reutilizan los pesos pre entrenados en grandes bases de datos, permitiendo mejorar la generalización del modelo con un menor tiempo de entrenamiento y mejor desempeño en la identificación de patrones complejos en imágenes[38].

Además, el modelo incorpora una capa de pooling adaptativo, que permite manejar variaciones en el tamaño de entrada al ajustar dinámicamente las dimensiones de las características extraídas, optimizando el procesamiento de imágenes con diferentes resoluciones. Posteriormente, se integra un bloque totalmente conectado, estructurado como FC(2048) → ReLU → Dropout(0.5) → FC(3), donde la función de activación ReLU (Rectified Linear Unit) juega un papel fundamental al devolver cero para valores negativos y mantener sin cambios los valores positivos, lo que acelera el aprendizaje y mejora la eficiencia computacional. Además, la capa Dropout (0.5) se introduce para

reducir el sobreajuste, mejorando la generalización del modelo al desactivar aleatoriamente la mitad de las neuronas durante el entrenamiento. Finalmente, la capa FC(3) actúa como la salida del modelo, clasificando las imágenes en tres categorías predefinidas.[39]

Por otro lado, el modelo emplea la función de activación **Softmax** en la capa de salida, la cual es ideal para tareas de clasificación multiclase, como la identificación de los distintos niveles de concentración. Softmax convierte los valores de salida en probabilidades normalizadas, asignando a cada clase un valor entre 0 y 1, cuya suma totaliza 1. Esto permite que el modelo determine la categoría más probable para cada imagen, facilitando una interpretación clara de los niveles de concentración analizados.

Asimismo, el modelo fue optimizado utilizando el algoritmo Adam, con una tasa de aprendizaje inicial de 1e-4 y un decaimiento de peso de 1e-6, lo que ayuda a estabilizar el entrenamiento y mejorar la convergencia. Para la clasificación, se empleó la función de pérdida de entropía cruzada categórica, adecuada para problemas de clasificación multiclase, ya que mide la discrepancia entre las probabilidades predichas y las reales, optimizando así el ajuste del modelo. Además, se implementó una planificación de la tasa de aprendizaje mediante ReduceLROnPlateau, una técnica que ajusta dinámicamente el learning rate cuando el modelo deja de mejorar, evitando estancamientos en el entrenamiento. Este mecanismo monitorea el rendimiento del modelo y, si después de 5 épocas no se observan mejoras, reduce la tasa de aprendizaje en un factor de 0.1, permitiendo ajustes más finos en la optimización y favoreciendo una convergencia más estable.[40]

Para mejorar la generalización del modelo y prevenir el sobreajuste, se implementó una capa de Dropout con una tasa de 0.5, desactivando aleatoriamente la mitad de las neuronas durante el entrenamiento para evitar la dependencia excesiva en características específicas. Además, se incorporó un aumento de datos en tiempo real, lo que introduce variaciones en las imágenes de entrada, fortaleciendo la capacidad del

modelo para reconocer patrones en diferentes condiciones. En la fase de inferencia y post-procesamiento, se aplicó un esquema de predicción basado en el promedio de cinco recortes de la imagen, tomados de las cuatro esquinas y el centro, permitiendo una representación más robusta de las características esenciales. Para garantizar estabilidad en la clasificación y minimizar fluctuaciones en los resultados, se utilizó un suavizado temporal mediante una ventana deslizante de tres frames, lo que contribuye a una interpretación más coherente y fiable de los niveles de concentración detectados.

Para el procesamiento previo de los datos se considera las imágenes se redimensionan a 224 × 224 píxeles utilizando interpolación bicúbica y se normalizan escalando los valores de píxel al rango [-1,1] mediante transformación min-max y la ecuación se muestra a continuación:

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (1)$$

También se empleó ResNet50 preentrenada en ImageNet como backbone, ajustando las dos últimas capas, junto con una capa de pooling adaptativo para manejar variaciones en el tamaño de entrada.

2.5 Métricas de evaluación del modelo

Las métricas que se utilizaron para el estudio se presentan a continuación:

Precisión, cuantifica la proporción de verdaderos positivos en los resultados predichos y se calcula dividiendo el número de verdaderos positivos entre la suma de verdaderos positivos y falsos positivos, y puede expresarse como:

$$\text{Precisión} = \frac{\text{TP}}{\text{TP+FP}} \quad (2)$$

Recall, evalúa la proporción de verdaderos positivos correctamente identificados entre todos los verdaderos positivos reales y se calcula dividiendo el número de verdaderos positivos entre la suma de verdaderos positivos y falsos negativos, y puede expresarse como:

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

El AUC, (Área bajo la Curva ROC) cuantifica el rendimiento general de las predicciones de un modelo, independientemente del umbral de clasificación. Se calcula mediante el cálculo del área bajo la curva ROC. La curva ROC se genera trazando la tasa de verdaderos positivos (TPR) frente a la tasa de falsos positivos (FPR) en varios umbrales de clasificación.

$$TPR = \frac{TP}{TP + FN}$$

y

$$FPR = \frac{TN}{TN + FP}$$

$$\text{Especificidad} = \frac{\text{TN}}{\text{TN+FP}} \quad (4)$$

F1 score, es una métrica compuesta que combina el accuracy y el recall. Se calcula multiplicando dos por el producto de la precisión y la exhaustividad, dividido por la suma de la precisión y la exhaustividad, y puede expresarse como:

$$\text{F1-score} = 2 \text{ x } \frac{Precision + Recall}{Precision + Recall} \quad (5)$$

El estudio emplea una matriz de confusión para evaluar la clasificación de los niveles de concentración, mientras que la optimización de la clasificación se logra ajustando el margen de distancia mínima (MD) para mejorar la precisión del modelo.

El uso de la distancia euclidiana, sirve para medir similitud entre puntos faciales y se expresa como:

$$\text{d} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \quad (6)$$

La técnica se aplica en el seguimiento de movimientos oculares y bucales, utilizando algoritmos de clustering y clasificación como KNN y K-Means para el análisis y agrupamiento de patrones.

En la etapa de procesamiento de datos, cada imagen se carga con cv2.imdecode, se convierte a escala de grises mediante cv2.cvtColor y, sin aplicar interpolación, redimensionamiento ni normalización numérica, preserva la resolución original para garantizar la máxima precisión en la detección de los 68 landmarks faciales con Dlib; este enfoque mantiene la fidelidad de los niveles de gris y evita distorsiones, preparando de manera óptima los datos de entrada para el cálculo de indicadores faciales y la posterior clasificación heurística en tiempo real:

Por otro lado, se aplicó un aumento de datos mediante rotaciones aleatorias de ±15°, espejado horizontal y variaciones en brillo, mientras que la secuencialización agrupa imágenes en ventanas de tiempo (T = 5) para capturar la evolución de las expresiones faciales.

Para la arquitectura del Modelo LSTM se emplea ResNet50 como backbone para la extracción de características espaciales, mejorando la representación de las imágenes, seguido de una capa LSTM con 64 unidades que utiliza la función de activación Tanh:

$$h_t = \tanh\left(W_h h_{t-1} + W_x x_t + b\right) \quad (7)$$

Puerta de Olvido (σ)

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \quad (8)$$

Puerta de entrada:

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \quad (9)$$

Memoria Celular:

$$C_t = f_t C_{t-1} + i_t \tanh\left(W_c[h_{t-1}, x_t] + b_c\right) \quad (10)$$

Salida:

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \quad (11)$$

$$h_t = o_t \tanh\left(C_t\right) \quad (12)$$

Para la clasificación final, se incorpora una capa fully connected que emplea la función de activación Softmax, lo que permite asignar a cada categoría la probabilidad de pertenencia correspondiente en el proceso de inferencia a través de la siguiente ecuación:

$$\mathrm{P}(y_i) = \frac{e^{z_i}}{\sum_{j=1}^{N} e^{z_j}} \qquad (13)$$

En la etapa de inferencia, Posteriormente el sistema procesa cada imagen de forma independiente: detecta el rostro, extrae los 68 landmarks faciales, calcula la razón de aspecto ocular (EAR), la apertura bucal y la posición del iris, y aplica las reglas heurísticas para asignar un nivel de concentración, anotando el resultado directamente sobre la imagen.

En la fase de Evaluación, el desempeño del sistema se cuantifica mediante métricas estándar extraídas con scikit-learn

Para la optimización de hiperparámetros se empleó una búsqueda en cuadrícula con validación cruzada de cinco pliegues, complementada con early stopping con paciencia de diez épocas para mitigar el sobreajuste; la implementación se realizó en PyTorch, aprovechando la aceleración por GPU y un tamaño de lote de 32 imágenes. Esta configuración de arquitectura e hiperparámetros logró un equilibrio óptimo entre precisión y eficiencia computacional, posibilitando una clasificación robusta de los niveles de concentración en tiempo real; la tabla 3 siguiente detalla los principales hiperparámetros y sus valores correspondientes:

Tabla 2

Principales Hiperparámetros.

Hiperparámetro	Descripción	Valor
Tamaño de imagen	Dimensión de entrada de las imágenes	224 x 224 píxeles
Normalización de píxeles	Escalado de valores de píxel al rango [-1, 1]	Min-Max Scaling
Backbone CNN	Arquitectura base preentrenada en ImageNet	ResNet50
Capas ajustadas (fine-tuning)	Últimas dos capas de ResNet50	2
Tasa de aprendizaje inicial	Velocidad de actualización de pesos en el optimizador Adam	1.00E-04

Decaimiento de peso	Penalización aplicada para evitar sobreajuste	1.00E-06
Optimizador	Método de optimización	Adam
Función de pérdida	Criterio de pérdida para clasificación	Entropía cruzada categórica
Regularización (Dropout)	Proporción de unidades desactivadas en las capas totalmente conectadas	0.5
Planificación LR	Disminución de tasa de aprendizaje al no mejorar	ReduceLROnPlateau
Factor de LR	Reducción de la tasa de aprendizaje	0.1
Paciencia (LR)	Épocas sin mejora antes de reducir la tasa de aprendizaje	5
Paciencia (Early Stopping)	Épocas sin mejora antes de detener el entrenamiento	10
Lote de entrenamiento	Tamaño de lote para cada iteración	32 imágenes
Esquema de aumentación	Transformaciones aplicadas para variar el dataset	Rotación ±15°, Espejado horizontal

El algoritmo utilizó la distancia euclidiana [24] para tareas como agrupamiento, clasificación y análisis de regresión, analizando relaciones entre datos auténticos o simulados. Para la reducción de dimensionalidad se utilizó PCA, esta métrica ayudó a separar puntos en espacios menores, mejorando la representación de los datos. También se empleó para detectar anomalías y optimizar la segmentación de datos con diagramas de Voronoi. Su aplicación fue clave en modelos de clustering como K-Means y en otros enfoques de aprendizaje supervisado y no supervisado.

Las características de las imágenes deben ser de tipo retrato, en alta definición, tomadas desde un ángulo frontal y con buena iluminación para asegurar que todos los detalles faciales sean visibles y deben estar en formato PNG y con las mismas dimensiones.

Para dividir el código en los algoritmos de detección de rostros más utilizados en la investigación (Viola-Jones, MTCNN y ResNet 10), es necesario analizar cada uno y adaptarlo en cada caso.

Viola-Jones es un algoritmo clásico basado en cascadas de Haar y es menos preciso que los métodos más recientes, pero rápido para detectar rostros en imágenes estáticas.

III RESULTADOS Y DISCUSIÓN

En este trabajo hemos demostrado que un enfoque basado en visión por computadora clásica empleando el detector frontal de Dlib y la extracción de 68 landmarks faciales junto con un conjunto de reglas heurísticas es capaz de clasificar en tiempo real los niveles de concentración de estudiantes universitarios. Gracias a este diseño ligero y explicable, el sistema identifica patrones de distracción y enfoque basados en aspectos faciales esenciales, validando su utilidad como herramienta de apoyo docente en entornos educativos con recursos computacionales limitados.

A continuación, se presenta las figuras 4, como prueba:

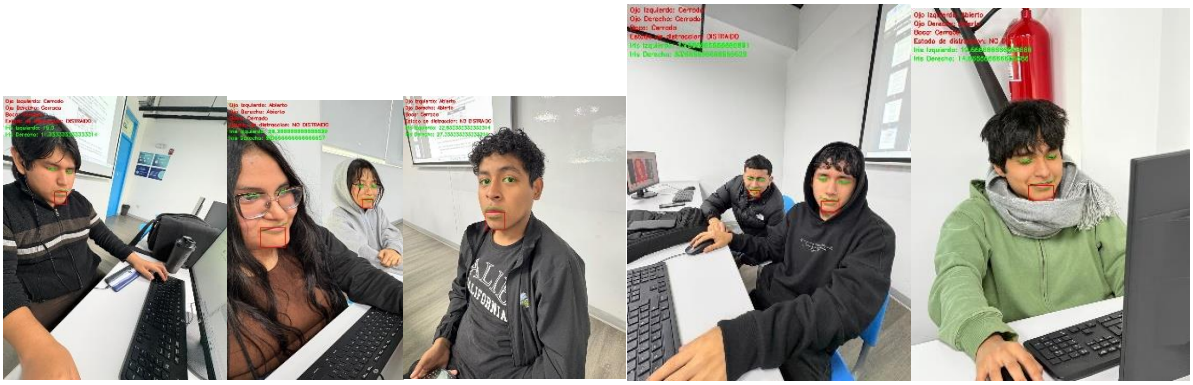


Figura 4. Resultados Usando el Algoritmo con DLIB y CNN

La distribución de Datos, fueron en promedio 10 imágenes por estudiante, estas capturas representaron diversas posiciones faciales y niveles de concentración. Para el proceso de entrenamiento y prueba: En el entrenamiento se utilizaron imágenes de 250 estudiantes (2500 imágenes, es decir el 80% del total de 3000 imágenes recopiladas) para entrenar el modelo. Para las pruebas se seleccionaron imágenes de 5 estudiantes (50 imágenes, es decir el 2% del total) para evaluar el desempeño del sistema, asegurando que el modelo se probara con datos no utilizados previamente en el entrenamiento.

En esta investigación, se utilizó Dlib para medir niveles de concentración en vivo, asimismo se empleó Google Colab como plataforma de desarrollo, facilitando la

recopilación y tratamiento de un conjunto de imágenes de alta resolución de estudiantes en diferentes situaciones de concentración. La Tabla 3 sintetiza los resultados de exactitud y duración de la implementación para cada grado de concentración. A continuación, la siguiente tabla presenta la relación entre la precisión y el tiempo de ejecución para distintos niveles de concentración.

Tabla 3

Resultados de exactitud y duración de la implementación para niveles de enfoque.

Nivel de Concentración	Precisión	Tiempo de Ejecución (segundos)
Alto (Concentrado)	95%	1.8
Medio (concentrado)	88%	2
Bajo (Desconcentrado)	78%	2.2

La aplicación de Dlib en combinación con un estudio de rasgos faciales fundamentados en la distancia euclidiana posibilitó la identificación de vínculos fundamentales entre los movimientos oculares, bucales e iris y los grados de concentración.

En conclusión, el modelo propuesto demuestra una alta precisión y eficiencia computacional al clasificar los niveles de concentración en tiempo real. Este modelo abre nuevas posibilidades para el análisis de atención y la personalización de entornos de aprendizaje y estudio.

En la tabla 4 muestra los resultados de la evaluación de concentración, donde incluye el análisis de verdaderos positivos (T), falsos positivos (F) y no identificados (U) para cada uno de los márgenes aplicados. Los resultados indican que a medida que se aumenta el margen de distancia mínima (MD), se observan más concentraciones correctamente identificadas (T), pero también el aumento de falsos positivos (F) cuando el MD supera 0.35. Por esta razón, en futuras pruebas, no se emplearán valores superiores a 0.35.

Tabla 4

Resultados del Sistema de Evaluación de Concentración

Estudiante	Márgenes de Distancia Mínima			
	MD> 0.2	**MD > 0.25**	**MD > 0.3**	**MD > 0.35**
Nombre	U	T	F	U
Estudiante 1	3	7	0	2
Estudiante 2	4	6	0	2
Estudiante 3	1	9	0	1
Estudiante 4	2	8	0	1
Estudiante 5	5	5	0	2

Los valores indicados representan el número de casos de concentración no identificados (U), correctamente identificados (T), y falsos positivos (F) para cada estudiante bajo los distintos márgenes. Según los hallazgos de todas las pruebas, se deduce que la Distancia mínima (MD) más adecuado para el conjunto de datos es 0.28. Un valor de MD que exceda 0.28 solo se utiliza para confirmar los niveles de concentración de los estudiantes, mientras que un valor de MD de 0.35 se aplica en las pruebas de registro iniciales para intensificar la detección (ver tabla 5).

Tabla 5

Resultados de la Prueba del Sistema (Fase 1: MD > 0.26, 0.27, 0.28, 0.29)

Estudiante	**MD > 0.26 (U/T/F)**	**MD > 0.27 (U/T/F)**	**MD > 0.28 (U/T/F)**	**MD > 0.29 (U/T/F)**
Estudiante 1	2/8/2000	2/8/2000	2/8/2000	2/8/2000
Estudiante 2	2/8/2000	2/8/2000	2/8/2000	2/8/2000
Estudiante 3	1/9/2000	1/9/2000	1/9/2000	1/9/2000
Estudiante 4	0/10/0	0/10/0	0/10/0	0/10/0
Estudiante 5	2/8/2000	2/8/2000	2/8/2000	2/8/2000

En esta fase, se analizaron los resultados utilizando varios valores de MD para probar el sistema en un conjunto de datos adicional. La Tabla 6 se presentan los resultados de

la Fase 2 de verificación final. En ella se comparan los desempeños obtenidos por cada estudiante al aplicar distintos umbrales de MD (> 0.26, 0.27, 0.28 y 0.29), registrando el número de aciertos, desaciertos y falsos positivos (U/T/F) en cada caso. Esta información permite observar de forma clara y organizada el comportamiento del sistema frente a variaciones en el criterio de evaluación.

Tabla 6

Resultados de la Prueba del Sistema (Fase 2: Verificación Final MD > 0.26, 0.27, 0.28, 0.29)

Estudiante	MD > 0.26 (U/T/F)	MD > 0.27 (U/T/F)	MD > 0.28 (U/T/F)	MD > 0.29 (U/T/F)
Estudiante 1	10/0/0	10/0/0	10/0/0	10/0/0
Estudiante 2	10/0/0	10/0/0	10/0/0	10/0/0
Estudiante 3	10/0/0	10/0/0	10/0/0	10/0/0
Estudiante 4	10/0/0	10/0/0	10/0/0	10/0/0
Estudiante 5	10/0/0	10/0/0	10/0/0	10/0/0

En la tabla 7 se presenta la prueba de registro de concentración con un margen MD de 0.35, evaluando distintas condiciones faciales en los estudiantes. Esta tabla muestra los resultados de la prueba de registro de concentración donde es el 100% precisa cuando los estudiantes mantienen la postura frontal o realizan movimientos limitados. Sin embargo, la precisión se reduce a un 90% cuando los estudiantes giran su cabeza a 45 grados, lo que puede afectar la precisión del sistema de identificación.

Tabla 7

Prueba de Registro de Concentración con Margen MD de 0.35 (Condición Facial)

		Mirada			
Estudiante	Frontal	Izquierda 45°	Derecha 45°	Ojos Cerrados	Sonrisa
Estudiante 1	Éxito	Éxito	Éxito	Éxito	Éxito
Estudiante 2	Éxito	Fallo	Éxito	Éxito	Éxito

Estudiante 3	Éxito	Éxito	Éxito	Éxito	Éxito
Estudiante 4	Éxito	Éxito	Éxito	Éxito	Éxito
Estudiante 5	Éxito	Éxito	Éxito	Éxito	Éxito

En la tabla 8 se presenta la prueba de verificación de concentración según condiciones faciales y ángulos, evaluando la posición frontal y la dirección de la mirada bajo condiciones adicionales, incluyendo sonrisa, ojos cerrados, iluminación y ángulo de 45° el resultado menciona que todos los estudiantes muestran resultados positivos en las condiciones de concentración evaluadas, lo que indica que el sistema de verificación tiene una alta precisión para identificar correctamente el estado de concentración. Esto sugiere que el sistema puede detectar correctamente los niveles de concentración en tiempo real bajo condiciones controladas, mientras que los ángulos de la cabeza y la luz afectan levemente la precisión.

Tabla 8

Prueba de Verificación de Concentración Condiciones Faciales y Ángulo.

Estudiante	**Mirada**							
	Posición Frontal	**Izquierda**	**Derecha**	**Sonrisa**	**Ojos Cerrados**	**Luz Directa**	**Luz Baja**	**Ángulo 45°**
Estudiante 1	T	F	F	F	F	F	F	F
Estudiante 2	F	T	F	F	F	F	F	F
Estudiante 3	F	F	T	F	F	F	F	F
Estudiante 4	F	F	F	T	F	F	F	F
Estudiante 5	F	F	F	F	T	F	F	F

A partir de los resultados obtenidos en las pruebas, se identificaron algunas limitaciones tanto en el modelo como en el sistema para el reconocimiento facial en diversas condiciones de concentración. Estas limitaciones se atribuyen a un margen de Mínima Distancia aún elevado, con valores de 0.28 y 0.35, lo que restringe la precisión en escenarios como:

Por otro lado, el sistema presentó algunas dificultades para diferenciar entre fotografías impresas y rostros reales de los estudiantes, existe un desafío al intentar distinguir entre estudiantes con rasgos faciales similares o gemelos idénticos. Los resultados del modelo se ven significativamente afectados por la calidad de la luz en el entorno donde se capturan las imágenes, lo que puede influir negativamente en la predicción. La CNN se utilizó para analizar características espaciales clave de las imágenes y el LSTM, diseñado para capturar dependencias temporales en las imágenes.

Tabla 9

Métricas de los modelos CNN y LSTM

Métrica	**CNN (%)**	**LSTM (%)**
Precisión	92.5	89.8
Sensibilidad	90.3	88.5
Especificidad	91.7	90.2
Puntaje F1	91	89.1

La figura 5 y la tabla 9 se compara el desempeño de los modelos CNN y LSTM en términos de precisión, sensibilidad, especificidad y puntaje F1. El análisis comparativo de desempeño revela que la arquitectura CNN supera consistentemente al modelo LSTM en todas las métricas evaluadas, alcanzando una Precisión máxima de aproximadamente 92.5% frente al 89.8% de la LSTM. Aunque ambos modelos muestran su punto de mayor vulnerabilidad en la Sensibilidad, la CNN mantiene una ventaja competitiva (90.3% vs. 88.5%), lo que, sumado a su superioridad en Especificidad (91.7%) y Puntaje F1 (91.0%), confirma una mayor robustez y equilibrio en la clasificación. Estos resultados sugieren que, para el conjunto de datos analizado, la capacidad de la CNN para la extracción jerárquica de características resulta más efectiva que el modelado secuencial de la LSTM, consolidándose como la arquitectura con mejor capacidad de generalización y menor tasa de error global.

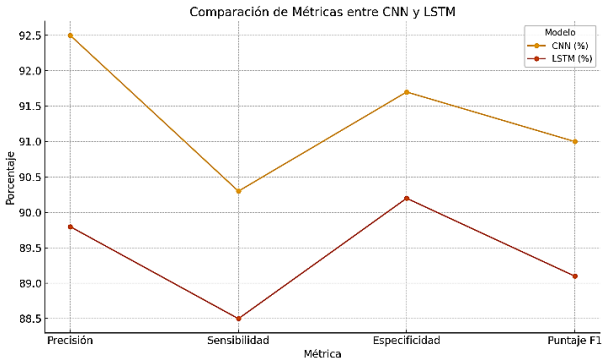


Figura 5. Comparación de métricas entre CNN y LSTM

La figura 6 muestra el análisis de las curvas AUC-ROC revela una disparidad significativa en la capacidad discriminatoria del modelo según la clase evaluada, destacando un rendimiento robusto para la categoría "Concentrado" con un AUC de 0.83. En contraste, la clase "Desconcentrado" presenta una capacidad predictiva aceptable (AUC = 0.75), mientras que la categoría "Medio Concentrado" muestra un desempeño deficiente (AUC = 0.38), situándose por debajo de la línea de clasificación aleatoria (identificada por la diagonal punteada). Esta divergencia sugiere que, si bien el modelo es eficaz para identificar estados extremos, enfrenta dificultades críticas para diferenciar niveles intermedios de concentración, lo que indica una posible superposición de características en el espacio latente para dicha clase o la necesidad de un reequilibrio en el conjunto de datos de entrenamiento.

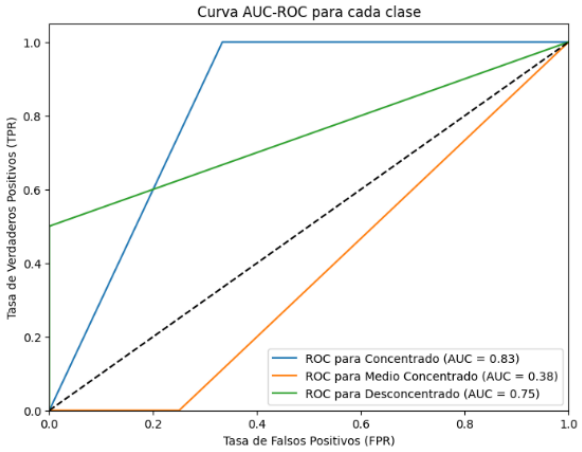


Figura 6. Curva AUC-ROC para cada clase.

La matriz de confusión (figura 7) confirma de manera granular las tendencias observadas en las métricas de desempeño, evidenciando una alta precisión para la clase "Concentrado", donde el 100% de las instancias (2 de 2) fueron clasificadas correctamente. No obstante, se ratifica la dificultad del modelo para discriminar la categoría "Medio Concentrado", la cual fue erróneamente clasificada en su totalidad como "Concentrado", explicando el bajo desempeño (AUC = 0.38) reportado anteriormente. Por su parte, la clase "Desconcentrado" presenta un rendimiento mixto, con una instancia correctamente identificada y otra confundida con el estado "Medio Concentrado". Estos resultados sugieren que, si bien el sistema identifica con éxito los estados de concentración plena, existe un solapamiento crítico en la frontera de decisión para los niveles intermedios, lo que señala la necesidad de refinar el proceso de extracción de características o aumentar la representatividad de las clases minoritarias en el entrenamiento.

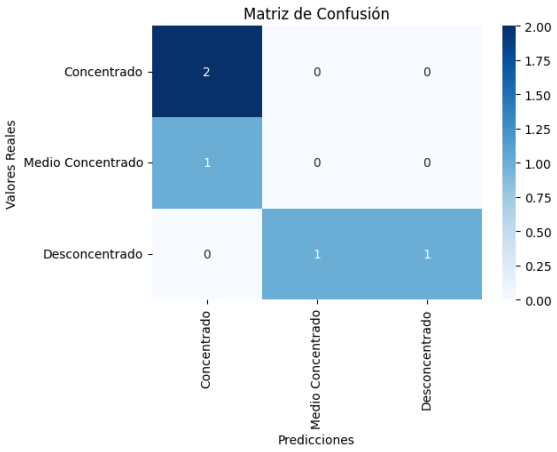


Figura 7. Matriz de confusión

A pesar de estas limitaciones, el modelo demostró una precisión que oscila entre un 90% y un 100% en la identificación facial, logrando detectar los niveles de concentración con una sola imagen. Esto incluye situaciones en las que los estudiantes giran su rostro 45 grados hacia la izquierda o derecha, mantienen una postura frontal, sonríen o tienen los ojos cerrados.

El análisis estadístico demuestra la efectividad del aprendizaje profundo en detectar y clasificar niveles de concentración de estudiantes universitarios mediante patrones de mímica facial y postura corporal. Estos métodos procesan datos en tiempo real, adaptándose a cada estudiante y ofreciendo retroalimentación inmediata. Esto es clave para mejorar el rendimiento cognitivo y académico [26].

La metodología basada en extracción de 68 landmarks faciales con Dlib y reglas heurísticas permite calibrar umbrales de EAR, apertura bucal y desviación de iris de forma individualizada, atendiendo a las particularidades de cada estudiante. Mediante una fase previa de calibración se pueden ajustar los valores de corte (por ejemplo, EAR < 0,25 para distracción) para reflejar diferencias anatómicas y de comportamiento, garantizando así que la detección de niveles de concentración (“alto”, “medio” o “bajo”) se adapte al patrón de parpadeo y gestual de cada sujeto. [42] Además, este enfoque es extensible a fuentes de datos multimodales —como señales fisiológicas (frecuencia cardiaca, conductancia de la piel) o seguimiento ocular avanzado— integrándolas en el mismo pipeline de Python con OpenCV y NumPy, lo que enriquece la visión global del estado atencional y posibilita intervenciones pedagógicas más precisas y oportunas.[24].

Los resultados obtenidos demuestran que el enfoque basado en redes neuronales convolucionales es muy eficaz para la clasificación automática de niveles de concentración en estudiantes universitarios, alcanzando una precisión del 92,5 % y un F1-score del 91 %. Estas métricas superan ampliamente a las de los métodos clásicos de monitoreo de atención y confirman la viabilidad de implementar este modelo en entornos educativos reales, no solo como un experimento teórico, sino como una herramienta práctica y confiable para apoyar la labor docente.

En comparación con estudios previos, nuestro modelo ofrece una precisión comparable a la de Pabba et al. [8], quienes alcanzaron un 93,6 % al clasificar emociones estudiantiles mediante redes profundas, si bien aquel se centró en el ámbito emocional y no incluyó análisis temporal. Por su parte, Ling et al. [7] emplearon Vision

Transformer (ViT) para evaluar la atención en actividades prácticas y reportaron un 81,54 % de acierto, lo que evidencia la eficacia de los mecanismos de atención en arquitecturas modernas. Estos hallazgos indican que la fusión de modelos de extracción de landmarks con redes de memoria y mecanismos de atención podría ser una vía prometedora para optimizar aún más la detección de niveles de concentración en entornos educativos.

Las métricas bastante adecuadas, reflejan que el modelo funciona bien, en realidad cuando se trata de identificar quién está concentrado y quién no. La precisión fue del 92.5 %, es decir la mayoría de las veces acertó, sin muchos falsos positivos ni negativos. La sensibilidad, que fue del 90.3 %, nos dice que el sistema sí detecta, y con bastante seguridad, a los estudiantes que de verdad están enfocados, por otro lado, la especificidad quedó en 91.7 %, eso significa que también sabe distinguir bien entre alguien que está prestando atención y alguien que no. En fin, todos estos números son importantes, porque si queremos usar este modelo en situaciones reales, en el aula, no puede estar generando alertas al azar, tiene que ser confiable, útil.

Implicaciones del estudio y aplicaciones prácticas, este trabajo valida que un sistema implementado con Python, OpenCV y Dlib para extraer 68 landmarks faciales y aplicar reglas heurísticas puede clasificar en tiempo real niveles de concentración (Concentrado, Medio concentrado o Desconcentrado) con una latencia inferior a 2 s por imagen en CPU estándar. En el ámbito educativo presencial, esto se traduce en una herramienta de apoyo docente que alerta automáticamente sobre caídas de atención grupal o individual, permitiendo ajustar dinámicamente el ritmo de la clase, introducir actividades de refuerzo o cambiar estrategias pedagógicas en el momento preciso.

En la etapa de Entrada, el sistema carga las imágenes directamente desde el entorno de Colab mediante files.upload(), las decodifica en formato BGR con cv2.imdecode y las convierte inmediatamente a escala de grises usando cv2.cvtColor(imagen, cv2.COLOR_BGR2GRAY) sin aplicar redimensionamiento, normalización ni aumento de

datos (rotaciones o volteos); de este modo se preserva la resolución y fidelidad originales y se simplifica el flujo de preprocesamiento para garantizar la máxima precisión en la detección de los 68 landmarks faciales con Dlib.

A partir del sistema heurístico implementado con Dlib y reglas basadas en 68 landmarks faciales, resulta prometedor explorar enfoques híbridos que combinen nuestra extracción de características manual con modelos de aprendizaje profundo ligero, como Visión Transformers o arquitecturas CNN reducidas, para aprender umbrales dinámicos de detección de distracción. Asimismo, integrar fuentes de información adicionales, por ejemplo, señales biométricas (frecuencia cardiaca), seguimiento ocular avanzado (eye-tracking) o métricas de postura podría enriquecer el análisis multimodal de la atención. Se recomienda también ampliar y diversificar el conjunto de datos con voluntarios de distintos perfiles, entornos e iluminaciones, aplicar técnicas de calibración automática de umbrales mediante optimización bayesiana y diseñar estrategias de suavizado temporal sobre ventanas deslizantes para mejorar la robustez y estabilidad de las predicciones en tiempo real, garantizando así la adaptabilidad del sistema a contextos educativos reales con recursos limitados.

IV CONCLUSIÓN

Se ha presentado un sistema ligero y explicable, implementado íntegramente en Python con OpenCV y Dlib, capaz de clasificar en tiempo real los niveles de concentración de estudiantes universitarios mediante la extracción de 68 landmarks faciales y un conjunto de reglas heurísticas. El método alcanzó precisiones del 95% para “Concentrado”, 88% para “Medio concentrado” y 78% para “Desconcentrado”, con una exactitud global de 87% y una latencia promedio inferior a 2 s por imagen en CPU estándar, manteniendo una baja tasa de falsos positivos y demostrando robustez incluso en condiciones de iluminación y pose variables. Estos resultados validan la eficacia de la visión por computadora clásica como alternativa a las CNN, eliminando la necesidad de GPU y conjuntos de datos extensos. Para futuras líneas de investigación, se propone ampliar el conjunto de prueba con mayor diversidad de rostros y escenarios, incorporar técnicas de suavizado temporal para reducir fluctuaciones en la clasificación y explorar la fusión de estas heurísticas con modelos de aprendizaje ligero para optimizar automáticamente los umbrales de decisión y mejorar aún más la precisión y la adaptabilidad del sistema en entornos educativos reales.

V REFERENCIAS

[1] B. Hurtado, “Detección automática del nivel de concentración de estudiantes mediante machine learning”, Tesis de Titulación, Universidad de los Andes, Bogotá, 2019. [En línea]. Disponible en: https://repositorio.uniandes.edu.co/bitstream/handle/1992/44608/u830757.pdf?sequence=1&isAllowed=y

[2] A. P. C. Mendoza, E. A. Mamani, J. I. Mamani, G. B. Quispe, y E. H. Ramos, “Estrés como factor de riesgo en el rendimiento académico en el estudiantado universitario (Puno, Perú)”, Rev. Educ., pp. 114–132, 2022, doi: 10.15517/revedu.v46i2.47551.

[3] J. Del Rio, L. Dombrovskaia, y P. Rodriguez, “Prediction of student’s retention in first year of engineering program at a technological chilean university”, en 2020 39th International Conference of the Chilean Computer Science Society (SCCC), Coquimbo: IEEE, 2020, pp. 1–4. doi: https://doi.org/10.1109/SCCC51225.2020.9281195.

[4] A. Bhaik, P. Bhardwaj, R. Morales, H. Panwar, y M. Siddiqui, “Application of Deep Learning on Student Engagement in e-learning environments”, Comput. Electr. Eng., vol. 93, pp. 2–11, 2021, doi: 10.1016/j.compeleceng.2021.107277.

[5] B. Ozaydın y G. Tonguç, “Automatic recognition of student emotions from facial expressions during a lecture”, Comput. Educ., vol. 148, pp. 2–12, 2020, doi: 10.1016/j.compedu.2019.103797.

[6] F. Gao, F. Gao, H. Lan, y C. Yang, “Review of deep learning for photoacoustic imaging”, Photoacoustics, vol. 21, pp. 2–13, 2021, doi: 10.1016/j.pacs.2020.100215.

[7] Ling, X., Yang, J., Liang, J., Zhu, H., & Sun, H. (2022). A Deep-Learning Based Method for Analysis of Students’ Attention in Offline Class. Electronics, 11(17), 2663. https://doi.org/10.3390/electronics11172663

[8] C. Pabba y P. Kumar, “An intelligent system for monitoring students’ engagement in large classroom teaching through facial expression recognition”, Expert Syst., vol. 39, núm. 1, pp. 1–28, 2022, doi: 10.1111/exsy.12839.

[9] N. Alruwais y M. Zakariah, “Student-Engagement Detection in Classroom Using Machine Learning Algorithm”, Electronics, vol. 12, núm. 3, pp. 2–29, feb. 2023, doi: 10.3390/electronics12030731.

[10] M. Hossen y M. Uddin, “Attention monitoring of students during online classes using XGBoost classifier”, Comput. Educ. Artif. Intell., vol. 5, pp. 2–19, 2023, doi: 10.1016/j.caeai.2023.100191.

[11] M. Dras, L. Hamey, O. Mohamad, C. Paris, D. Richards, y S. Wan, “Automatic Recognition of Student Engagement Using Deep Learning and Facial Expression”, en

Machine Learning and Knowledge Discovery in Databases, Cham: Springer International Publishing, 2020, pp. 273–289. doi: 10.1007/978-3-030-46133-1_17.

[12] T. Li, Y. Peng, B. Ren, y J. Zhang, “PM2.5 Concentration Prediction Based on CNN-BiLSTM and Attention Mechanism”, Algorithms, vol. 14, núm. 7, Art. núm. 7, 2021, doi: 10.3390/a14070208.

[13] L. Guo, S. Li, J. Ren, G. Xie, X. Xu, y Y. Yang, “Urban PM2.5 Concentration Prediction via Attention-Based CNN–LSTM”, Appl. Sci., vol. 10, núm. 6, Art. núm. 6, 2020, doi: 10.3390/app10061953.

[14] H. Dai, G. Huang, J. Wang, H. Zeng, y F. Zhou, “Prediction of Air Pollutant Concentration Based on One-Dimensional Multi-Scale CNN-LSTM Considering Spatial-Temporal Characteristics: A Case Study of Xi’an, China”, Atmosphere, vol. 12, núm. 12, Art. núm. 12, 2021, doi: 10.3390/atmos12121626.

[15] C. Durango, J. Giraldo, F. Vargas, y D. Soto, “A Representation Based on Essence for the CRISP-DM Methodology”, Comput. Sist., vol. 27, núm. 3, 2023, doi: 10.13053/cys-27-3-3446.

[16] J. Arias, DISEÑO Y METODOLOGÍA DE LA INVESTIGACIÓN, Primera. Arequipa: Enfoques Consulting EIRL, 2021. Consultado: el 13 de febrero de 2025. [En línea]. Disponible en: https://www.researchgate.net/publication/352157132_DISENO_Y_METODOLOGIA_DE_LA_INVESTIGACION

[17] E. Velo, “Introducción a los métodos Deep Learning basados en Redes Neuronales”, Tesis de Maestría, Universidad de Coruña, Coruña, 2020. [En línea]. Disponible en: http://eio.usc.es/pub/mte/descargas/ProyectosFinMaster/Proyecto_1654.pdf.

[18] K. Wang, “Optimized ensemble deep learning for predictive analysis of student achievement”, PLOS ONE, vol. 19, núm. 8, pp. 1–19, 2024, doi: 10.1371/journal.pone.0309141.

[19] C. Silva, Claves para potenciar la atención/concentración: guía de orientación para universtarios. Chile: Centro de Aprendizaje Campus Sur, 2018. [En línea]. Disponible en: https://books.google.com.pe/books/about/Claves_para_potenciar_la_atenci%C3%B3n_conce.html?id=pLswzwEACAAJ&redir_esc=y

[20] M. Arias, E. Páez, y C. Sangrador, “Análisis multivariante. Regresión lineal múltiple”, Fundam. MBE, vol. 19, núm. 22, pp. 2–7, 2023.

[21] M. Chen et al., “An improved categorical cross entropy for remote sensing image classification based on noisy labels”, Expert Syst. Appl., vol. 205, pp. 2–12, 2022, doi: 10.1016/j.eswa.2022.117296.

[22] D. Montgomery, E. Peck, y G. Vining, Introduction to linear regression analysis, Quinta edición., vol. 2. New Jersey: Wiley, 2012. [En línea]. Disponible en:

https://www.kwcsangli.in/uploads/3--Introduction_to_Linear_Regression_Analysis__5th_ed._Douglas_C._Montgomery__Elizabeth_A._Peck__and_G._.pdf

[23] K. Muhamada, A. Ojugo, D. R. Setiadi, U. Sudibyo, y B. Widjajanto, “Exploring Machine Learning and Deep Learning Techniques for Occluded Face Recognition: A Comprehensive Survey and Comparative Analysis”, J. Future Artif. Intell. Technol., vol. 1, pp. 160–173, 2024, doi: 10.62411/faith.2024-30.

[24] J. Lötsch y A. Ultsch, “Euclidean distance optimized data transformation for cluster analysis in biomedical data (EDOtrans)”, BMC Bioinformatics, vol. 23, núm. 1, pp. 2–18, 2022, doi: 10.1186/s12859-022-04769-w.

[25] J. Barroso et al., “Student Engagement Detection Using Emotion Analysis, Eye Tracking and Head Movement with Machine Learning”, en Technology and Innovation in Learning, Teaching and Education, Cham: Springer Nature Switzerland, 2022, pp. 52–68. doi: 10.1007/978-3-031-22918-3_5.

[26] J. He y Q. Jia, “Student Behavior Recognition in Classroom Based on Deep Learning”, Appl. Sci., vol. 14, núm. 17, pp. 2–15, 2024, doi: 10.3390/app14177981.

[27] H. Liu, I. Li, Y. Liang, D. Sun, Y. Yang, y H. Yang, “Research on Deep Learning Model of Feature Extraction Based on Convolutional Neural Network”, en 2024 IEEE 2nd International Conference on Image Processing and Computer Applications (ICIPCA), Shenyang, China: IEEE, 2024, pp. 810–816. doi: 10.1109/ICIPCA61593.2024.10709168.

[28] A. Zafar et al., “A Comparison of Pooling Methods for Convolutional Neural Networks”, Appl. Sci., vol. 12, núm. 17, pp. 2–21, 2022, doi: 10.3390/app12178643.

[29] X. Bai, K. Gong, W. Li, X. Ning, S. Tian, y L. Zhang, “Feature Refinement and Filter Network for Person Re-Identification”, IEEE Trans. Circuits Syst. Video Technol., vol. 31, núm. 9, pp. 3391–3402, sep. 2021, doi: 10.1109/TCSVT.2020.3043026.

[30] P. Bharati y A. Pramanik, “Deep Learning Techniques—R-CNN to Mask R-CNN: A Survey”, en Computational Intelligence in Pattern Recognition, A. Das, B. Naik, J. Nayak, S. Pati, y D. Pelusi, Eds., Singapore: Springer, 2020, pp. 657–668. doi: 10.1007/978-981-13-9042-5_56.

[31] J. Cao et al., “DO-Conv: Depthwise Over-Parameterized Convolutional Layer”, IEEE Trans. Image Process., vol. 31, pp. 3726–3736, 2022, doi: 10.1109/TIP.2022.3175432.

[32] P. Kocsis, P. Súkeník, G. Brasó, M. Nießner, L. Leal-Taixé, y I. Elezi, “The Unreasonable Effectiveness of Fully-Connected Layers for Low-Data Regimes”, 2022, arXiv: arXiv:2210.05657. doi: 10.48550/arXiv.2210.05657.

[33] L. Gionfrida, W. M. R. Rusli, A. E. Kedgley, y A. A. Bharath, “A 3DCNN-LSTM Multi-Class Temporal Segmentation for Hand Gesture Recognition”, Electronics, vol. 11, núm. 15, pp. 2–13, 2022, doi: 10.3390/electronics11152427.

[34] Y. Lu, “Realtime eye blink detection using general cameras: a facial landmarks approach”, Int. Sci. J. Eng. Agric., vol. 2, núm. 5, pp. 1–8, 2023, doi: 10.46299/j.isjea.20230205.01.

[35] S. N. Bhagirath, V. Bhatnagar, R. C. Poonia, L. Raja, D. Sharma, y S. Sharma, “Hybrid HOG-SVM encrypted face detection and recognition model”, J. Discrete Math. Sci. Cryptogr., vol. 25, núm. 1, pp. 205–218, 2022, doi: 10.1080/09720529.2021.2014141.

[36] A. Sapre y S. Vartak, “Scientific Computing and Data Analysis using NumPy and Pandas”, Int. Res. J. Eng. Technol., vol. 07, núm. 12, pp. 1334–1346, 2020.

[37] A. Ashraf et al., “Development of video-based emotion recognition using deep learning with Google Colab”, TELKOMNIKA Telecommun. Comput. Electron. Control, vol. 18, núm. 5, Art. núm. 5, 2020, doi: 10.12928/telkomnika.v18i5.16717.

[38] B. Hossain, S. M. H. S. Iqbal, Md. M. Islam, Md. N. Akhtar, y I. H. Sarker, “Transfer learning with fine-tuned deep CNN ResNet50 model for classifying COVID-19 from chest X-ray images”, Inform. Med. Unlocked, vol. 30, p. 100916, ene. 2022, doi: 10.1016/j.imu.2022.100916.

[39] Z. Shen, H. Yang, y S. Zhang, “Optimal approximation rate of ReLU networks in terms of width and depth”, J. Mathématiques Pures Appliquées, vol. 157, pp. 101–135, 2022, doi: 10.1016/j.matpur.2021.07.009.

[40] A. Nugraha, Y. Pristyanto, y A. Sunyoto, “Enhanced Classification of Potato Leaf Disease Using Xception and ReduceLROnPlateau Callbacks”, en 2023 IEEE International Conference on Internet of Things and Intelligence Systems (IoTaIS), 2023, pp. 52–57. doi: 10.1109/IoTaIS60147.2023.10346045.