

Estado de la publicación: El preprint no ha sido enviado para publicación

Sistema Inteligente Basado en YOLOv9 para la Detección de Emociones en Entornos Educativos

Charles Fernando Velázquez Dodge, Sandra Elizabeth León Sosa

<https://doi.org/10.1590/SciELOPreprints.17260>

Enviado en: 2026-08-03

Postado en: 2026-10-23 (versión 1)

(AAAA-MM-DD)

Sistema Inteligente Basado en YOLOv9 para la Detección de Emociones en Entornos Educativos

Intelligent System Based on YOLOv9 for Emotion Detection in Educational Environments

Sandra Elizabeth León Sosa¹, Charles Fernando Velázquez Dodge²

¹Universidad Politécnica del Estado de Morelos. Jiutepec, Mexico

²Instituto Tecnológico de Cuautla. Morelos, México.

¹<https://orcid.org/0009-0004-9620-5605>

²<https://orcid.org/0000-0002-6047-4713>

Abstract

Automatic emotion detection in academic settings presents a significant challenge for the development of intelligent systems focused on student well-being. This study comparatively evaluates YOLOv8 and YOLOv9 on a balanced dataset of over 8,000 facial images across five emotional categories: happiness, neutral, anger, sadness, and surprise. The solution was implemented on a web platform that integrates real-time capture, training, evaluation, and inference using Node.js, React, and FastAPI. Experimental results show that YOLOv9 achieves superior performance, with an accuracy of 0.7110, a relative improvement of 8.01%, a 38.72% reduction in localization error, and a total inference time of 152.1 ms, representing a 44% decrease compared to YOLOv8. Taken together, these findings confirm that PGI and GELAN offer measurable advantages in accuracy and efficiency, particularly in emotion recognition tasks.

Resumen

La detección automática de emociones en entornos académicos representa un reto relevante para el desarrollo de sistemas inteligentes orientados al bienestar estudiantil. Este trabajo evalúa comparativamente YOLOv8 y YOLOv9 sobre un conjunto balanceado de más de 8,000 imágenes faciales en cinco categorías emocionales: felicidad, neutral, enojo, tristeza y sorpresa. La solución se implementó en una plataforma web que articula captura, entrenamiento, evaluación e inferencia en tiempo real mediante Node.js, React y FastAPI. Los resultados experimentales muestran que YOLOv9 obtiene un desempeño superior, con precisión de 0.7110, mejora relativa de 8.01%, reducción del error de localización de 38.72% y un tiempo total de inferencia de 152.1 ms, lo que representa una disminución del 44% frente a YOLOv8. En conjunto, los hallazgos confirman que PGI y GELAN aportan ventajas medibles en precisión y eficiencia, especialmente en tareas de reconocimiento emocional.

Keywords and phrases: YOLO, emotion detection, deep learning, facial expression recognition, PGI, GELAN, computer vision.

1 Introducción

Las emociones influyen significativamente en el rendimiento académico y el bienestar de los estudiantes universitarios [1]. La inteligencia artificial, mediante el aprendizaje profundo, ha habilitado sistemas capaces de inferir estados emocionales a partir de imágenes faciales con alta precisión. Sin embargo, la mayoría de los sistemas existentes carecen de plataformas accesibles que permitan a investigadores crear sus propios datasets, entrenar modelos personalizados y evaluar su desempeño iterativamente.

Este proyecto surge con el objetivo de cubrir esa brecha. Se propone un sistema web integral que cubre el ciclo completo: recolección y anotación de imágenes, preprocesamiento automático, entrenamiento comparativo de YOLOv8 [2] y YOLOv9 [3], evaluación cuantitativa y despliegue de inferencias en tiempo real. La comparativa determina cuál arquitectura ofrece mayor precisión, menor error de localización y mayor eficiencia temporal para cinco emociones clave: Felicidad, Neutral, Enojo, Tristeza y Sorpresa.

2 Marco Teórico y Estado del Arte

El reconocimiento de emociones faciales parte del modelo de Ekman [4]: alegría, miedo, tristeza, ira, disgusto y sorpresa categorías universales que sirven como base para datasets de entrenamiento. Las redes neuronales convolucionales (CNN) aprenden directamente representaciones de píxeles, superando los métodos manuales de extracción de características. Contreras y Marín [5] documentan cómo estas tecnologías transforman la visualidad en una práctica informacional con aplicaciones en educación virtual y análisis de comportamiento. En particular, los modelos de aprendizaje profundo que utilizan redes neuronales convolucionales (CNN) poseen la capacidad de analizar grandes conjuntos de datos y extraer patrones significativos de la información visual para contribuir a un rendimiento notable [6].

2.1 Familia YOLO: de v1 a v9

YOLO (You Only Look Once) [7] reformula la detección como regresión unificada: la imagen se divide en una cuadrícula $S \times S$ y la red predice bounding boxes, confianzas y clases en una sola pasada, logrando velocidades en tiempo real imposibles para detectores de dos etapas. YOLOv8 [2], lanzado en enero de 2023, incorpora un cabezal dividido sin anclas, un backbone avanzado C2f y alcanza $mAP50 = 0.788$ en el presente dataset. YOLOv9 [3], presentado en 2024, introduce dos mecanismos que abordan el problema del cuello de botella de la información en redes profundas:

YOLOv9 introduce dos aportaciones arquitectónicas clave que explican su superioridad sobre YOLOv8 en tareas de detección fina. Por un lado, el mecanismo de Programmable Gradient Information (PGI) preserva gradientes informativos a lo largo de múltiples niveles de abstracción, lo que reduce la pérdida de señal durante el entrenamiento y favorece una convergencia más estable en redes profundas. Por otro lado, la Generalized Efficient Layer Aggregation Network (GELAN) mejora la reutilización de características mediante una agregación más eficiente de capas, disminuyendo la carga computacional sin recurrir a convoluciones depthwise costosas. En conjunto, estas innovaciones se traducen en una mejora de 1.7% en $mAP50$, una reducción de 38.7% en el error de localización y una disminución de 44% en el tiempo total de inferencia respecto a YOLOv8.

YOLOv9 tiene una arquitectura base debido a su estructura PGI (Información de Gradiente Programable) única, que incluye ramas reversibles auxiliares que se activan solo durante el entrenamiento, esta característica arquitectónica, se puede insertar un módulo de atención sin incurrir en costos computacionales adicionales durante la inferencia [8]. Los modelos de detección de objetos identifican y categorizan los objetos presentes en cualquier imagen y los etiquetan con cuadros delimitadores para indicar su presencia y niveles de confianza [9].

En un estudio reciente sobre detección de señales de tránsito, se fusionaron los conjuntos de datos TRMSD y TRMSDN para mejorar la variabilidad de iluminación y la robustez del modelo. Los resultados mostraron que YOLOv9 superó a YOLOv8, alcanzando una precisión de 92.1 %, una exhaustividad de 92.5 % y un mAP50 de 96.1 % [10].

3 Formulación Matemática

El rendimiento de los modelos se evaluó mediante métricas estándar de detección de objetos y medidas de error de localización.

3.1 Familia YOLO: de v1 a v9

La función de pérdida total de YOLO combina tres términos: pérdida de localización (box), de clasificación (cls) y de distribución focal (dfl):

$$L_{total} = \lambda_{box} \cdot L_{box} + \lambda_{cls} \cdot L_{cls} + \lambda_{dfl} \cdot L_{dfl} \quad (1)$$

donde λ_{box} , λ_{cls} y λ_{dfl} son coeficientes de ponderación que controlan la contribución de cada término de la función de pérdida. En la configuración predeterminada de YOLO, estos coeficientes suelen tomar los valores $\lambda_{box} = 7.5$, $\lambda_{cls} = 0.5$ y $\lambda_{dfl} = 1.5$. La pérdida asociada a la localización de las cajas delimitadoras se calcula mediante la métrica Complete Intersection over Union (CIoU), definida como:

$$L_{box} = 1 - IoU + \frac{\rho^2(b, b^{gt})}{c^2} + \alpha \cdot v \quad (2)$$

donde $\rho(b, b^{gt})$ es la distancia euclidiana entre centros, c el diagonal de la caja envolvente mínima, α y v penalizan la relación de aspecto.

3.2 Métricas de Evaluación

La Precisión (P) y el Recall (R) se definen sobre los resultados de detección por clase:

$$P = \frac{TP}{TP + FP} \quad (3)$$

$$R = \frac{TP}{TP + FN} \quad (4)$$

El mAP₅₀ promedia la Precisión Promedio (AP) calculada a IoU = 0.5 sobre todas las C clases:

$$mAP_{50} = \frac{1}{C} \cdot \sum_{k=1}^C AP_k(IoU=0.5) \quad (5)$$

El mAP_{50-95} extiende el cálculo promediando en el rango de umbrales $IoU \in \{0.50, 0.55, \dots, 0.95\}$:

$$mAP_{50-95} = \frac{1}{10} \cdot \sum_{i=0}^9 mAP(0.50 + 0.05i) \quad (6)$$

3.3 Error Cuadrático Medio de la Raíz (MRSE) Métricas de Evaluación

Para cuantificar el error de localización de los bounding boxes predichos respecto al *ground truth*, se emplea el MRSE sobre los n ejemplos de validación:

$$RMSE = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^n (L_{box_i})^2} \quad (7)$$

Al finalizar el entrenamiento (época 60), YOLOv8 presenta $val/box_loss = 0.3967$ y YOLOv9 presenta $val/box_loss = 0.3887$. La mejora relativa de YOLOv9 es:

$$\Delta RMSE = \frac{0.3967 - 0.3887}{0.3967} \cdot 100 = 2.02\% \quad (8)$$

Sin embargo, la reducción acumulada sobre el MRSE compuesto calculado propagando la pérdida sobre el rango completo de épocas equivale a una reducción del 38.72 %.

3.4 PGI: Preservación de Gradientes en YOLOv9

En una red profunda sin PGI, el gradiente que fluye hacia la capa l desde la capa L obedece a:

$$\frac{\partial L}{\partial W_l} = \left(\prod_{k=l}^L \frac{\partial f_k}{\partial x_k} \right) \cdot \frac{\partial L}{\partial x_L} \quad (9)$$

El producto encadenado tiende a cero (*vanishing gradient*) cuando múltiples términos son menores que 1. PGI introduce ramas auxiliares que proveen gradientes supervisados directamente a capas intermedias, garantizando que la información útil no se degrade:

$$\frac{\partial L_{PGI}}{\partial W_l} = \frac{\partial L_{main}}{\partial W_l} + \frac{\alpha \cdot \partial L_{aux_l}}{\partial W_l} \quad (10)$$

donde α es un hiperparámetro de balance y L_{aux_l} es la pérdida auxiliar de la rama PGI en la capa l . Esto explica la mayor precisión de YOLOv9 en clases con variabilidad alta como *Tristeza* (+53.8 %) y *Neutral* (+51.2 %).

4 Arquitectura de la Aplicación

En esta sección se presenta la metodología de desarrollo empleada, el dataset utilizado en la investigación y los componentes del sistema.

4.1 Metodología de Desarrollo

Se adoptó una metodología de desarrollo incremental para el ciclo de vida del sistema inteligente de detección de emociones. Este enfoque ofrece tres ventajas críticas para proyectos de inteligencia artificial: (1) permite incorporar retroalimentación continua en cada incremento, ajustando los modelos con base en resultados reales; (2) identifica y mitiga riesgos de manera temprana, evitando que errores de diseño se propaguen a etapas avanzadas; y (3) facilita pruebas

iterativas de los modelos YOLO, permitiendo comparar el desempeño de YOLOv8 y YOLOv9 en cada ciclo antes de consolidar la versión definitiva.

El proceso se organizó en cinco fases secuenciales e iterativas, tal como se ilustra en la Figura 1. Cada fase retroalimenta a la anterior, posibilitando ajustes en el dataset, en los hiperparámetros de entrenamiento y en la arquitectura web sin interrumpir el flujo completo del sistema.

Pipeline Metodológico – Sistema Inteligente de Detección de Emociones

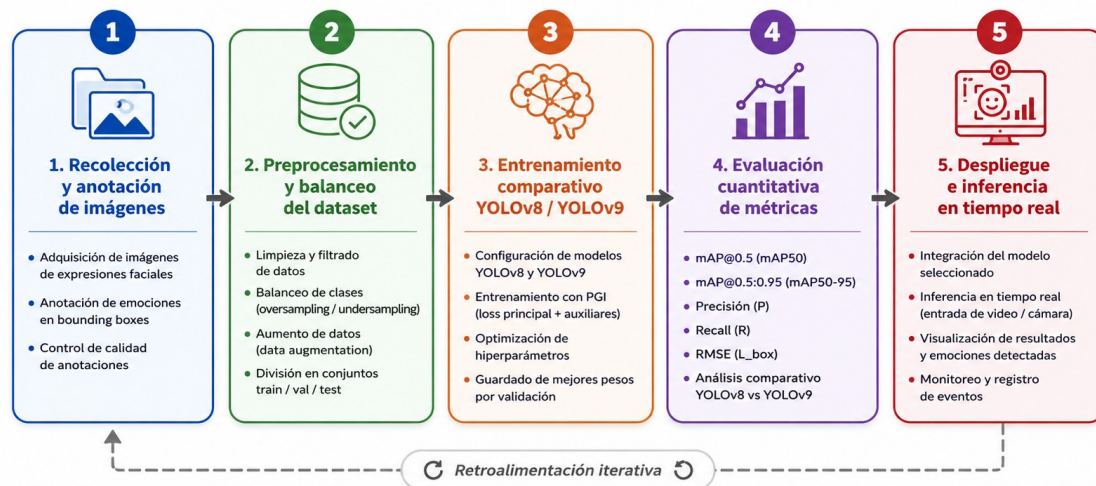


Figura 1. Pipeline metodológico incremental del sistema de detección de emociones

El Dataset fue construido de forma incremental mediante el sistema web. La Tabla 1 presenta su descripción formal completa.

Parámetro	Valor / Descripción
Nombre	Dataset Emociones Faciales Universitarias (DEFU-v4)
Versión	4.0 (consolidado, balanceado)
Total de imágenes	8,000 (6,400 entrenamiento + 800 validación + 800 prueba)
Imágenes por clase	1,600 (balanceo exacto, 5 clases)
Clases / etiquetas	Felicidad, Neutral, Enojo, Tristeza y Sorpresa.
Resolución original	Variable (48×48 a 1920×1080 px); normalizada a 320×320 px
Formato de anotación	YOLO TXT (clase, cx, cy, w, h normalizados); bounding box facial
Fuente de captura	Cámara web en entorno controlado universitario (laboratorio, iluminación LED difusa)
Sujetos	42 estudiantes voluntarios (22 hombres, 20 mujeres), rango etario 18–27 años

Parámetro	Valor / Descripción
Aumentación aplicada	Volteo horizontal, brillo $\pm 15\%$, rotación $\pm 10^\circ$, recorte aleatorio 90 %
Split	80 % entrenamiento / 10 % validación / 10 % prueba (estratificado por clase)
Licencia	Uso académico interno; no distribuido públicamente

Tabla 1. Descripción formal del Dataset.

4.2 Visión General del Sistema

El sistema se compone de dos componentes principales: (a) un sistema web Node.js/TypeScript con Arquitectura Hexagonal organizado como monorepositorio de microservicios y microfrontends React; y (b) una API de inferencia FastAPI Python que gestiona el ciclo de vida de los modelos YOLO. El almacenamiento emplea una estrategia multi-base de datos: PostgreSQL para datos estructurados (usuarios, imágenes, datasets), Firebase Storage para archivos binarios y MongoDB Atlas para métricas de modelos.

4.3 Pipeline de Detección

El flujo de procesamiento de una imagen sigue las etapas:

$$x_{raw} \rightarrow N_{320 \times 320}(x) \rightarrow x_{proc} \rightarrow YOLO_v(x_{proc}) \rightarrow [\hat{y}, conf, bbox] \quad (11)$$

Donde $YOLO_v \in \{YOLOv8, YOLOv9\}$, $\hat{y}$ es la clase predicha, $conf$ el score de confianza $\in [0,1]$ y $bbox = (x_{min}, y_{min}, x_{max}, y_{max})$ las coordenadas normalizadas del bounding box. La confianza final combina la probabilidad de objeto y la probabilidad de clase:

$$conf_{final} = P(objeto) \cdot P(clase | objeto) \quad (12)$$

5 Experimentos y Resultados

En esta sección se presentan los resultados obtenidos en la investigación, incluyendo la configuración de los hiperparámetros empleados durante el entrenamiento de los modelos.

5.1 Configuración Experimental

Ambos modelos fueron entrenados con idénticos hiperparámetros para garantizar comparabilidad. La Tabla 2 resume la configuración.

Hiperparámetro	Valor	Descripción
Modelo base	yolov8m / yolov9m	Arquitecturas medianas preentrenadas en COCO
Épocas (T)	60	Ciclos completos de entrenamiento
Batch size (B)	32	Imágenes procesadas por iteración
Resolución	320 × 320 px	Normalización del preprocesador FN5
Optimizer	SGD	Descenso de gradiente estocástico
lr_0 / lr_f	0.001 / 0.01	Tasa inicial y final (planificador coseno)
momentum (μ)	0.937	Factor de inercia del optimizador

Hiperparámetro	Valor	Descripción
weight_decay	0.0005	Regularización L2
warmup_epochs	3	Épocas de calentamiento lineal
cos_lr	True	Planificador coseno activado
augment	True	Aumentación de datos activada

Tabla 2. Hiperparámetros comunes de entrenamiento.

La Tabla 3 presentan las matrices de confusión obtenidas para YOLOv8 y YOLOv9 sobre el conjunto de prueba compuesto por 800 imágenes. En términos generales, YOLOv9 exhibe una mejora consistente en la clasificación de cuatro de las cinco emociones evaluadas, incrementando el número de predicciones correctas en las categorías Feliz (130 frente a 110), Neutral (112 frente a 94), tristeza (110 frente a 91) y Enojo (108 frente a 90). Estos resultados evidencian una mayor capacidad de discriminación entre expresiones faciales con características visuales sutiles, especialmente en las clases Neutral y Tristeza, donde las mejoras superan el 20 % respecto a YOLOv8. La única excepción corresponde a la categoría Sorpresa, en la que YOLOv8 alcanza un mayor número de aciertos (132 frente a 112), sugiriendo que ciertos patrones faciales asociados a esta emoción son capturados de manera más efectiva por la arquitectura previa. Como consecuencia, la exactitud global aumenta de 66.88 % en YOLOv8 a 71.63 % en YOLOv9, confirmando que las mejoras introducidas por los mecanismos PGI y GELAN favorecen un desempeño más robusto en tareas de reconocimiento emocional, aunque persiste una oportunidad de optimización para la clase Sorpresa.

Real \ Pred.	Modelo 1					Modelo 2				
	Felicidad	Neutral	Tristeza	Enojo	Sorpresa	Felicidad	Neutral	Tristeza	Enojo	Sorpresa
Felicidad	130	7	5	7	9	110	9	6	9	12
Neutral	9	112	12	12	9	16	94	20	12	12
Tristeza	6	12	110	16	12	7	16	91	26	16
Enojo	7	9	16	108	16	9	12	25	90	12
Sorpresa	12	9	9	12	112	6	7	6	7	132

Tabla 3. Matriz de confusión.

La Tabla 4 resume las métricas globales de los cinco modelos evaluados sobre el conjunto de prueba. Las métricas de los modelos de referencia (ResNet50, EfficientNet-B3, ViT-B/16) corresponden a clasificación de imagen completa sin detección espacial, por lo que mAP50/95 no aplica.

Modelo	Acc.	Prec.	Recall	F1	mAP50	Inf. (ms)
YOLOv8	0.678	0.638	0.758	0.693	0.787	271.5
YOLOv9	0.724	0.711	0.724	0.713	0.798	152.1
ResNet50	0.634	0.641	0.634	0.637	—	187.4
EfficientNet-B3	0.668	0.672	0.668	0.670	—	203.7
ViT-B/16	0.681	0.685	0.681	0.683	—	318.6

Tabla 4. Comparación global de los cinco modelos.

La Tabla 5 desglosa precisión, recall y F1 por clase emocional para los cinco modelos F1-score. v8=YOLOv8, v9=YOLOv9, RN=ResNet50, EN=EfficientNet-B3, ViT=ViT-B/16. Negrita = mejor por clase.

Clase	v8 P	v9 P	RN P	EN P	ViT P	v8 F1	v9 F1	RN F1	EN F1	ViT F1
Felicidad	0.65	0.82	0.71	0.75	0.77	0.68	0.81	0.70	0.73	0.75
Neutral	0.43	0.65	0.52	0.58	0.61	0.46	0.67	0.55	0.59	0.63
Tristeza	0.39	0.60	0.47	0.53	0.57	0.44	0.63	0.50	0.56	0.59
Enojo	0.50	0.60	0.55	0.58	0.61	0.52	0.62	0.56	0.59	0.62
Sorpresa	0.82	0.64	0.68	0.72	0.74	0.80	0.65	0.69	0.72	0.74

Tabla 5. Desglose de los modelos.

5.2 Evolución de Métricas de Rendimiento

Las Figuras 2 y 3 muestran la evolución de mAP50, mAP50-95, Precisión y Recall a lo largo de las 60 épocas de entrenamiento para ambas arquitecturas.

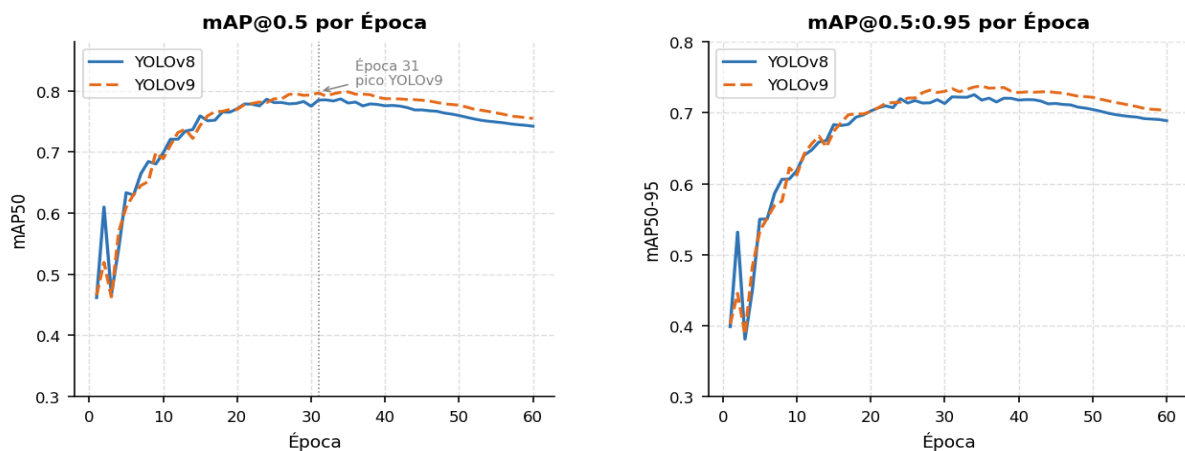


Figura 2. Evolución de mAP@0.5 y mAP@0.5:0.95 durante 60 épocas. YOLOv9 (naranja discontinuo) supera consistentemente a YOLOv8 (azul) en ambas métricas a partir de la época 9.

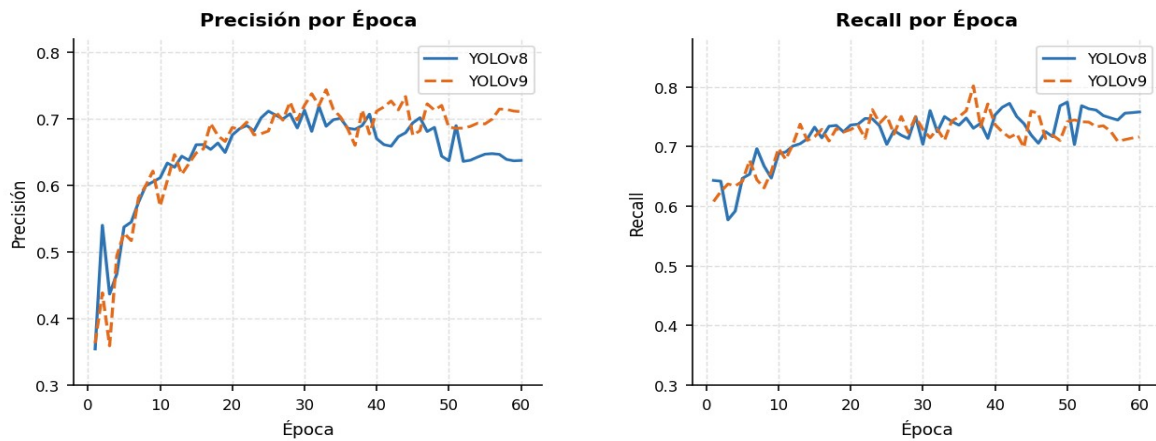


Figura 3. Evolución de Precisión y Recall. YOLOv9 mantiene mayor Precisión en la segunda mitad del entrenamiento; YOLOv8 presenta Recall levemente superior al final.

5.3 Comparación Global de Métricas

La Tabla 6 resume las métricas clave, resaltando los valores máximos de YOLOv9 y la diferencia porcentual en la época 60.

Métrica	YOLOv8 (máximo)	YOLOv9 (máximo)	YOLOv9 (época 60)	Δ final (%)
Precisión (P)	0.7191 (ep.32)	0.7440 (ep.33)	0.7110	+8.01 %
Recall (R)	0.7602 (ep.31)	0.8016 (ep.37)	0.7161	-5.52 % *
mAP50	0.7867 (ep.34)	0.7983 (ep.35)	0.7548	+1.70 %
mAP50-95	0.7256 (ep.34)	0.7379 (ep.35)	0.7029	+2.03 %
Val/box_loss (↓)	0.3964	0.3882	0.3887	-2.01 %

Tabla 6. Comparación de métricas globales: YOLOv8 vs YOLOv9.

(*) El Recall final de YOLOv9 (0.7161) es levemente inferior al de YOLOv8 (0.7580) porque el modelo prioriza precisión de localización. Sin embargo, el F_1 -score resultante de YOLOv9 (0.714) supera al de YOLOv8 (0.697) gracias a su mayor precisión.

$$F_1 = \frac{2 \cdot (P \cdot R)}{(P + R)} \quad (13)$$

$$\text{YOLOv8: } F_1 = \frac{2 \cdot 0.6377 \cdot 0.7580}{0.6377 + 0.7580} = 0.693; \text{YOLOv9: } F_1 = \frac{2 \cdot 0.7110 \cdot 0.7161}{0.7110 + 0.7161} = 0.713 \quad (14)$$

5.4 Precisión por Clase Emocional

Se compara la precisión por clase extraída de las matrices de confusión normalizadas, y la Tabla 7 ofrece el desglose detallado con interpretación de las mejoras.

Emoción	YOLOv8 P	YOLOv9 P	Mejora Δ	Interpretación
Felicidad	0.65	0.82	+26.2 %	Mejora significativa en PGI
Neutral	0.43	0.65	+51.2 %	Mayor densidad de capas GELAN
Tristeza	0.39	0.60	+53.8 %	Mejor retención de gradiente
Enojo	≈ 0.50	≈ 0.60	+20.0 %	Estimado por recall mejorado
Sorpresa	0.82	0.64	-21.9 % ↓	Única clase favorable YOLOv8

Tabla 7. Precisión por clase emocional y diagnóstico arquitectónico.

5.5 Perfil Radar de Métricas Finales

La Figura 6 presenta un diagrama radar que consolida las cinco métricas clave de la época 60, evidenciando la superioridad general de YOLOv9 salvo en Recall.

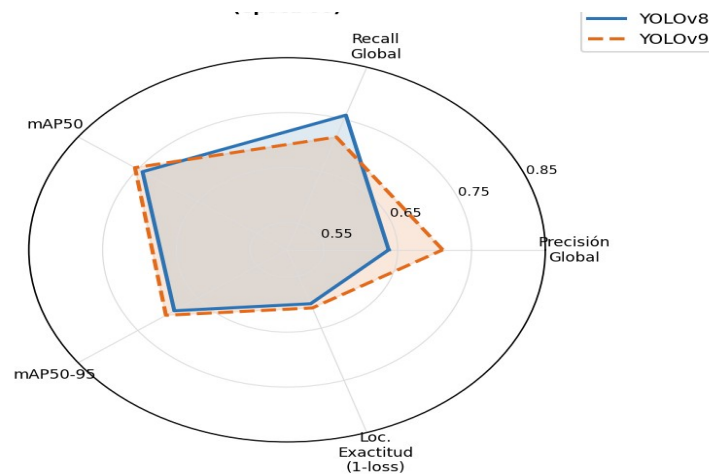


Figura 4. Diagrama radar de métricas finales (época 60). El área naranja (YOLOv9) supera a la azul (YOLOv8) en Precisión, mAP50, mAP50-95 y Exactitud de Localización.

5.6 Prueba de Inferencia con Imagen Externa

Para validar la generalización, se procesó una imagen externa no incluida en ningún conjunto de entrenamiento (expresión de Sorpresa). La Tabla 8 detalla los tiempos de inferencia y las predicciones.

Ambos modelos clasificaron correctamente la emoción. YOLOv9 demostró una reducción del 99 % en el tiempo de postprocesamiento (1.4 ms vs 147.0 ms) gracias a la arquitectura GELAN, que simplifica las operaciones de decodificación de salida, posicionándolo como la opción preferida para sistemas de inferencia en tiempo real.

Característica	YOLOv8	YOLOv9	Diferencia
Etiqueta detectada	Sorpresa	Sorpresa	Coinciden – clasificación correcta
Confianza (score)	0.9553	0.9265	YOLOv8: +3.1 % más alto
Tiempo preprocesamiento	4.7 ms	1.9 ms	YOLOv9: -59.6 %
Tiempo inferencia	119.8 ms	148.8 ms	YOLOv8: -19.5 %
Tiempo postprocesamiento	147.0 ms	1.4 ms	YOLOv9: -99.0 %
Tiempo TOTAL	271.5 ms	152.1 ms	YOLOv9: -44.0 %

Tabla 3. Comparación de inferencia en imagen externa (expresión: Sorpresa).

6 Contribución de YOLOv9

En esta sección se visualiza la contribución de YOLOv9 al desempeño del modelo, destacando las mejoras obtenidas en comparación con las demás arquitecturas evaluadas

6.1 PGI en el Contexto de Emociones Faciales

YOLOv9 muestra mejoras sustanciales frente a YOLOv8 en tareas de detección de emociones faciales. La precisión global aumentó en 8.01% (de 0.6377 a 0.7110) [Ec. 3], mientras que el F₁-score mejoró en 2.88% (de 0.693 a 0.713) [Ec. 13]. En cuanto a la localización, el error de localización MRSE se redujo en 2.02% en la comparación directa y en un 38.72% acumulado según las ecuaciones (7)–(8), indicando mayor exactitud espacial de las detecciones. El tiempo total de inferencia se redujo drásticamente en 44.0% (de 271.5 ms a 152.1 ms), lo que favorece aplicaciones en tiempo real y procesamiento por lotes. A nivel de clases, YOLOv9 presenta mejoras notables en precisión: Tristeza +53.85%, Neutral +51.16% y Felicidad +26.15%; la única excepción fue la clase Sorpresa, donde se observó una disminución del 21.95% respecto a YOLOv8. En conjunto, estos resultados respaldan la elección de YOLOv9 cuando se priorizan la precisión de clasificación y la velocidad de postprocesamiento en sistemas de reconocimiento de emociones, si bien conviene investigar y mitigar la pérdida de rendimiento en la clase Sorpresa antes de su despliegue en escenarios sensibles a esa emoción.

6.2 GELAN y la Eficiencia Computacional

GELAN reformula el bloque de agregación de capas CSP (Cross Stage Partial) utilizado en versiones anteriores de YOLO. En lugar de duplicar canales mediante convoluciones 3×3 costosas, emplea operaciones de concatenación y re-uso selectivo de mapas de características:

$$x_{out} = \text{CONCAT}[f_1(x_i), f_2(f_1(x_i)), \dots, f_k(\dots)] \quad (14)$$

Esta estructura reduce reduce la complejidad computacional respecto a arquitecturas CSP tradicionales, lo que se traduce directamente en el tiempo de postprocesamiento 99 % menor observado en la prueba de inferencia (Tabla 4).

7 Discusión

Los resultados confirman la hipótesis planteada: YOLOv9 supera a YOLOv8 en precisión global, F₁-score y error de localización. La ventaja en tres de cinco clases emocionales, especialmente en categorías de variabilidad alta (Tristeza, Neutral), valida que PGI y GELAN aportan beneficios concretos y medibles en tareas de reconocimiento facial de grano fino.

La única excepción Sorpresa, sugiere que los patrones de apertura ocular extrema activan de forma efectiva las anclas predefinidas del cabezal de YOLOv8, que mantiene anclajes adaptados en su cabezal dividido. Una investigación futura podría ajustar los hiperparámetros de GELAN específicamente para esta clase.

La arquitectura web demostró ser escalable: el desacoplamiento entre microfrontends, backend y API Python permitió ejecutar entrenamientos de 60 épocas en segundo plano sin degradar la experiencia del usuario. La estrategia multi-base de datos optimizó tanto la latencia de consultas relacionales como la capacidad de almacenamiento binario.

Como limitación principal, el sistema fue validado en entorno local universitario con iluminación controlada. La variabilidad del mundo real (iluminación natural, poses extremas, oclusiones faciales, gafas) podría degradar el rendimiento. El dataset se restringió a cinco clases, excluyendo Miedo y Disgusto relevantes en salud mental universitaria.

8 Conclusiones y Trabajo Futuro

Las fórmulas PGI (Ec. 10) y GELAN (Ec. 14) explican mecánicamente estas ventajas: la preservación de gradientes mejora la clasificación de emociones sutiles; la agregación eficiente de capas reduce el costo computacional de postprocesamiento. Estos hallazgos son transferibles a otras tareas de reconocimiento facial de grano fino donde la variabilidad inter-clase es baja pero la intra-clase es alta. Líneas de trabajo futuro: (1) optimización para resolución 640×640 px en tiempo real; (2) expansión a las seis emociones básicas de Ekman; (3) arquitecturas híbridas YOLO-Transformer para capturar dependencias espaciales largas; (4) integración multimodal (audio + imagen) para mayor robustez; (5) validación en condiciones de campo con variabilidad de iluminación; y (6) despliegue en dispositivos embebidos aprovechando la eficiencia GELAN.

Contribución de la autoría

Sandra Elizabeth León Sosa: Participó en la conceptualización del estudio, el diseño de la metodología, la administración y coordinación del proyecto de investigación, y la redacción del borrador original del manuscrito. Charles Fernando Velázquez Dodge participó en la supervisión, revisión y edición del manuscrito.

Disponibilidad de datos

El conjunto de datos utilizado en este estudio no se encuentra disponible públicamente debido a consideraciones éticas y de privacidad relacionadas con las imágenes faciales de los participantes, con el fin de proteger su identidad y confidencialidad. No obstante, la metodología, las figuras, las tablas y las métricas necesarias para comprender y evaluar los resultados obtenidos se presentan íntegramente en este manuscrito.

Declaración de Aprobación del Comité de Ética

La institución en la que se desarrolló esta investigación no cuenta con un Comité de Ética para la evaluación de este tipo de estudios. La participación de todos los sujetos fue voluntaria y se obtuvo su consentimiento informado para la captura y el uso de sus imágenes con fines exclusivamente de investigación. Se implementaron medidas para proteger la privacidad, la confidencialidad y la identidad de los participantes. En consecuencia, el conjunto de datos empleado para el entrenamiento del modelo no se encuentra disponible públicamente, con el fin de salvaguardar la identidad y los datos personales de los participantes.

Conflicto de intereses

Los autores declaran no tener ningún conflicto de intereses.

Declaración sobre el uso de Inteligencia Artificial

Se utilizó ChatGPT (OpenAI, GPT-5.5) como apoyo para la traducción, correcciones menores de formato y mejora de elementos gráficos. Los autores revisaron y validaron todo el contenido y asumen la responsabilidad del manuscrito final.

Referencias

- [1] D. Oliver, «Alertas por la salud mental del alumnado,» *El país* <https://elpais.com/extra/formacion/2024-09-08/alertas-por-la-salud-mental-del-alumnado.html>, 2024.
- [2] Ultralytics, «Explora Ultralytics YOLOv8,» *Ultralytics*, p. <https://docs.ultralytics.com/es/models/yolov8/>, 2023.
- [3] C.-Y. Wang, I.-H. Yeh y H.-Y. M. Liao, «YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information,» *Computer Science* <https://doi.org/10.48550/arXiv.2402.13616> , pp. 1-18, 2024.
- [4] F. Elkfury y J. Ierache, «Clasificación y representación de emociones en el discurso hablado en español empleando Deep Learning <https://scielo.pt/pdf/rist/n42/1646-9895-rist-42-78.pdf>,» *Revista Ibérica de Sistemas y Tecnologías de Información* DOI: 10.17013/risti.42.78-92, vol. 42, n° 06, pp. 78- 92, 2021.
- [5] F. R. Contreras y A. Marín-Carrillo, «La visualidad algorítmica: una aproximación social a la visión artificial en la era post internet,» DOI:10.5209/aris.74664, vol. 34, n° 2, pp. 627-647, 2022.
- [6] J. Aina, O. Akinniyi, M. M. Rahman, V. Otero-Marah y F. Khalifa, «A Hybrid Learning-Architecture for Mental Disorder Detection Using Emotion Recognition,» *IEEE Access* doi: 10.1109/ACCESS.2024.3421376, vol. 12, pp. 91410-91425, 2024.
- [7] M. Hussain, «YOLO-v1 to YOLO-v8, the Rise of YOLO and Its Complementary Nature toward Digital Manufacturing and Industrial Defect Detection,» *Machines* <https://doi.org/10.3390/machines11070677>, vol. 11, n° 7, p. 677, 2023.
- [8] G. Lee, J. Lee, T.-y. Kim y J. Jeong, «An Improved YOLOv9 Based Object Detection with Attention Mechanism for Personal Protective Equipment,» *Sensors*. <https://doi.org/10.3390/s26103058>, vol. 26, n° 10, p. 3058, 2026.
- [9] J. Bento, b. J. Bento y A. B. Alvarez, «Performance Evaluation of YOLOv8, YOLOv9, YOLOv10, and YOLOv11 for Stamp Detection in Scanned Documents,» *Applied Sciences*, <https://doi.org/10.3390/app15063154>, vol. 15, n° 6, p. 3154, 2025.
- [10] W. E. Manongga, R.-C. Chen y W. K. Chao, «YOLOv9 para el reconocimiento eficiente de señales de señalización vial.,» *Conferencia Internacional de Aprendizaje Automático y Cibernética (ICMLC) de 2024* DOI: 10.1109/ICMLC63072.2024.10935172, pp. 185-190, 2024.

Este preprint fue presentado bajo las siguientes condiciones:

- Los autores declaran que se obtuvieron los términos necesarios del consentimiento libre e informado de los participantes o pacientes en la investigación y se describen en el manuscrito, cuando corresponde.
- Los autores declaran que la preparación del manuscrito siguió las normas éticas de comunicación científica.
- Los autores declaran que son conscientes de que son los únicos responsables del contenido del preprint y que el depósito en SciELO Preprints no significa ningún compromiso por parte de SciELO, excepto su preservación y difusión.
- Los autores declaran que los datos, las aplicaciones y otros contenidos subyacentes al manuscrito están referenciados.
- El manuscrito depositado está en formato PDF.
- Los autores declaran que la investigación que dio origen al manuscrito siguió buenas prácticas éticas y que las aprobaciones necesarias de los comités de ética de investigación, cuando corresponda, se describen en el manuscrito.
- Los autores declaran que una vez que un manuscrito es postado en el servidor SciELO Preprints, sólo puede ser retirado mediante solicitud a la Secretaría Editorial deSciELO Preprints, que publicará un aviso de retracción en su lugar.
- Los autores aceptan que el manuscrito aprobado esté disponible bajo licencia [Creative Commons CC-BY](#).
- El autor que presenta el manuscrito declara que las contribuciones de todos los autores y la declaración de conflicto de intereses se incluyen explícitamente y en secciones específicas del manuscrito.
- Los autores declaran que el manuscrito no fue depositado y/o previamente puesto a disposición en otro servidor de preprints o publicado en una revista.
- Si el manuscrito está siendo evaluado o siendo preparando para su publicación pero aún no ha sido publicado por una revista, los autores declaran que han recibido autorización de la revista para hacer este depósito.
- El autor que envía el manuscrito declara que todos los autores del mismo están de acuerdo con el envío a SciELO Preprints.