Modelo de red neuronal convolucional para la detección precisa de emociones en imágenes faciales

Convolutional Neural Network Model for Accurate Emotion Detection in Facial Images

Universidad Nacional Micaela Bastidas de Apurimac – Perú
Riqchary, revista de investigación en ciencia y tecnología
ISSN: 2810-8124 (en línea) / ISSN: 2706-543x
Vol. 8 Núm. 1 (2026) - Publicado: 22/07/2026
doi.org/10.57166/riqchary/v8.n1.2026
Páginas: 45 - 54
Recibido 14/08/2025 ; Aceptado 13/07/2026
https://doi.org/10.57166/riqchary/v8.n1.2026.6

Max Plinior Zavala-Ayvar
Universidad Nacional Micaela Bastidas de Apurímac
https://orcid.org/0009-0001-6304-8840
201072@unamba.edu.pe

Rodrigo Espinoza-Sayago
Universidad Nacional Micaela Bastidas de Apurímac
https://orcid.org/0009-0001-7537-9356
192196@unamba.edu.pe

Mario Aquino-Cruz
Universidad Nacional Micaela Bastidas de Apurímac
https://orcid.org/0000-0002-2552-5669
maquino@unamba.edu.pe

Francisco Cari-Incahuanaco
Universidad Nacional Micaela Bastidas de Apurímac
https://orcid.org/0000-0002-2807-0495
fcari@unamba.edu.pe

Resumen— El objetivo de este trabajo es desarrollar una arquitectura de red neuronal convolucional (CNN) ligera para el reconocimiento de emociones faciales, utilizando el conjunto de datos FER2013. El modelo se compuso de cuatro bloques convolucionales con un número creciente de filtros, seguidos de capas totalmente conectadas para la clasificación multiclase. Dado que el conjunto FER2013 estuvo compuesto por imágenes faciales en escala de grises, no fue necesario aplicar transformaciones de color adicionales. El preprocesamiento empleado se limitó a la normalización de los valores de píxel y a técnicas de aumento de datos, en concordancia con prácticas comunes en modelos CNN modernos, con el objetivo de mejorar la capacidad de generalización del modelo. La arquitectura propuesta alcanzó una precisión del 67,49 % y un F1-score macro del 65,64 %, demostrando un rendimiento competitivo frente a enfoques previos basados en redes convolucionales. Se implementó un sistema de detección de emociones faciales en tiempo real mediante visión por computadora, permitiendo la identificación automática de emociones a partir de imágenes capturadas por cámara. El proyecto, incluyendo el código fuente, instrucciones de ejecución y soporte para Windows y Linux, quedó disponible públicamente en GitHub, promoviendo la reproducibilidad y la implementación práctica del modelo.

Palabras clave: FER2013, reconocimiento de emociones faciales, redes neuronales convolucionales, visión artificial.

Abstract— The objective of this work is to develop a lightweight convolutional neural network (CNN) architecture for facial emotion recognition using the FER2013 dataset. The model consisted of four convolutional blocks with an increasing number of filters, followed by fully connected layers for multi-class classification. Since the FER2013 dataset comprised grayscale facial images, no additional color transformations were required. Preprocessing was limited to pixel value normalization and data augmentation techniques—consistent with common practices in modern CNN models—aimed at improving the model's generalization capability. The proposed architecture achieved an accuracy of 67.49% and a macro F1-score of 65.64%, demonstrating competitive performance compared to previous approaches based on convolutional networks. A real-time facial emotion detection system was implemented using computer vision, enabling the automatic identification of emotions from camera-captured images. The project—including source code, execution instructions, and support for Windows and Linux—was made publicly available on GitHub, fostering reproducibility and the practical implementation of the model.

Keywords: computer vision, convolutional neural networks, facial emotion recognition, FER2013.

1 INTRODUCCIÓN

El reconocimiento de emociones faciales (Facial Emotion Recognition, FER) se ha convertido en una herramienta clave dentro del campo de la visión por computadora, especialmente en aplicaciones que requieren interacción afectiva entre humanos y sistemas inteligentes. Entre sus principales ámbitos de aplicación destacan los asistentes virtuales, sistemas de vigilancia inteligente, entornos educativos personalizados y herramientas de apoyo en salud mental [1], [2], [3], [4], [5]. Esta tarea consiste en identificar automáticamente emociones humanas a partir de expresiones faciales, ya sea en imágenes estáticas o secuencias de vídeo [6].

A pesar de los avances recientes impulsados por el aprendizaje profundo, el FER continúa enfrentando desafíos importantes relacionados con la variabilidad de las expresiones faciales. Factores como edad, género, cultura, condiciones de iluminación, ángulos de visión y oclusiones pueden dificultar la generalización de los modelos [7], [8], [9], [10]. Estos problemas son especialmente críticos en ambientes no controlados, donde las emociones pueden manifestarse de forma ambigua o combinada.

En este estudio se aborda esta problemática mediante el diseño de una arquitectura ligera basada en redes neuronales convolucionales (CNN), entrenada específicamente sobre el conjunto de datos FER2013. Este dataset, ampliamente utilizado en la comunidad de investigación, contiene 35 775 imágenes faciales en escala de grises (48×48 píxeles), clasificadas en siete categorías emocionales: enojo, disgusto, miedo, felicidad, tristeza, sorpresa y neutral [11].

La arquitectura propuesta está compuesta por cuatro bloques convolucionales con un número creciente de filtros, seguidos de capas densas para la clasificación. Con el objetivo de mejorar la extracción de características, el preprocesamiento se limitó a la normalización de los valores de píxel al rango [0, 1] y a técnicas de aumentación de datos en tiempo real, dado que las imágenes del conjunto FER2013 ya se encuentran en escala de grises y no requieren conversión adicional.

El modelo fue entrenado y validado sobre FER2013, alcanzando una precisión del 67,49 % y un F1-score macro de 65,64 %, resultados que lo posicionan competitivamente frente a modelos similares existentes [12], [13], [14]. Adicionalmente, se implementó un sistema de detección de emociones en tiempo real mediante cámara, con soporte para entornos Linux y Windows, lo cual permite su despliegue práctico en múltiples contextos.

Todo el proyecto, incluyendo el código fuente, scripts de ejecución, documentación técnica y ejemplos visuales, quedó disponible públicamente en un repositorio de GitHub [23], promoviendo así la transparencia, reproducibilidad y reutilización por parte de la comunidad científica y desarrolladores interesados.

A pesar de los avances recientes en el reconocimiento de emociones faciales mediante aprendizaje profundo, se identifica un vacío relevante en la literatura relacionado con la combinación de múltiples conjuntos de datos para mejorar la capacidad de generalización de los modelos, manteniendo al mismo tiempo arquitecturas ligeras y eficientes. Si bien diversos estudios han demostrado que el entrenamiento inter-dataset —integrando bases como FER2013, RAF-DB, AffectNet o CK+— puede mejorar el rendimiento en escenarios no controlados, estas aproximaciones suelen apoyarse en modelos complejos y computacionalmente costosos, lo que limita su aplicación en sistemas en tiempo real o en entornos con recursos restringidos. En este contexto, existe una carencia de investigaciones que exploren arquitecturas convolucionales livianas como base para esquemas de entrenamiento multi-dataset o aprendizaje transferido. El presente trabajo se centra deliberadamente en el conjunto de datos FER2013 y sienta las bases para abordar dicho vacío, proponiendo una CNN eficiente y de bajo costo computacional cuyo diseño modular la hace adecuada para futuras extensiones orientadas a la combinación de múltiples datasets y a la mejora progresiva del reconocimiento emocional.

La relevancia de esta investigación radica en que los modelos de mayor precisión suelen ser computacionalmente costosos, lo que dificulta su implementación en entornos con recursos limitados o en tiempo real; por ello, resulta necesario desarrollar arquitecturas que equilibren precisión y eficiencia. En consecuencia, el objetivo principal de este estudio es diseñar y validar una arquitectura de red neuronal convolucional (CNN) liviana para el reconocimiento preciso de emociones faciales sobre el conjunto de datos FER2013, manteniendo una eficiencia computacional adecuada para aplicaciones de detección de emociones en tiempo real.

En este contexto, se plantea la siguiente hipótesis de investigación: el diseño de una arquitectura de red neuronal convolucional (CNN) liviana, que integre bloques progresivos y una capa de regularización espacial (BlurLayer), permite alcanzar una precisión superior al 65 % en el conjunto de datos FER2013, manteniendo una eficiencia computacional adecuada para aplicaciones de detección de emociones en tiempo real.

El resto del presente artículo se encuentra organizado de la siguiente manera: la Sección 2 presenta una revisión de los trabajos relacionados y el estado del arte en el reconocimiento emocional; la Sección 3 describe la metodología, el conjunto de datos FER2013 y la arquitectura de red neuronal propuesta; la Sección 4 detalla los experimentos realizados y los resultados obtenidos en términos de precisión y F1-score; la Sección 5 ofrece una discusión comparativa de los hallazgos frente a otros modelos; y finalmente, la Sección 6 expone las conclusiones y líneas de investigación futuras.

2 TRABAJOS RELACIONADOS

El reconocimiento automático de emociones faciales ha

sido ampliamente explorado en la literatura, destacando particularmente el impacto de las técnicas de aprendizaje profundo en la mejora del rendimiento frente a enfoques tradicionales basados en extracción manual de características, como HOG, SIFT o LBP [15]. Estas técnicas convencionales, aunque útiles en sus inicios, han sido progresivamente superadas por modelos capaces de aprender representaciones jerárquicas directamente de los datos de entrada, tal como señalan Zeng et al. [1].

Las redes neuronales convolucionales (CNN) han demostrado una alta eficacia en tareas de clasificación emocional. Por ejemplo, Mollahosseini et al. [15] propusieron una arquitectura CNN optimizada sobre el conjunto de datos AffectNet, logrando mejoras significativas frente a modelos clásicos. Goodfellow et al. [11] aplicaron redes neuronales profundas (DNN) en escenarios realistas, haciendo hincapié en la importancia de un preprocesamiento robusto para contrarrestar variaciones de iluminación y pose, elementos críticos en entornos no controlados[10].

En años recientes, la investigación ha evolucionado hacia modelos que integran mecanismos de atención espacial y de canal, con el objetivo de focalizar el aprendizaje en regiones faciales más relevantes. Tang [17] desarrolló una arquitectura CNN simplificada que alcanzó resultados sobresalientes en FER2013 gracias al uso de dropout, data augmentation y batch normalization. De forma paralela, la arquitectura LHC-Net propuesta por Pecoraro et al. [18] integró atención canalizada sobre una red ResNet34, mostrando mejoras consistentes en conjuntos como RAF-DB y AffectNet.

Otros enfoques recientes incluyen el uso de Vision Transformers (ViTs), como en el trabajo de Li et al. [19], que emplearon módulos de atención y promotores complementarios multivista, logrando resultados sobresalientes en tareas de reconocimiento emocional. Asimismo, redes generativas adversariales (GANs) como GroupGAN [20] han sido utilizadas para mejorar la diferenciación de emociones ambiguas o con pocas muestras, mediante generación de datos sintéticos y realce de características afectivas.

Recientemente, también se han desarrollado modelos optimizados para su implementación en dispositivos móviles o de bajo costo, como los propuestos por Pascual et al. [14], quienes exploraron módulos convolucionales eficientes y ligeros para FER2013.

Particularmente, el conjunto de datos FER2013 ha sido objeto de múltiples estudios debido a su accesibilidad y complejidad. Diversos autores han ajustado arquitecturas profundas al tamaño reducido de las imágenes (48×48 px), utilizando desde CNN básicas [12] hasta modelos con bloques residuales y técnicas avanzadas de aumento de datos [20], [21].

En este contexto, el presente trabajo se alinea con las estrategias basadas en redes convolucionales eficientes, proponiendo una arquitectura optimizada para FER2013. Se enfatiza el uso de un preprocesamiento adecuado y se evalúa exhaustivamente su desempeño. Además, se acompaña de una implementación completa y pública [23], con soporte multiplataforma (Windows y Linux), lo cual contribuye a su replicabilidad, extensión práctica y posible uso en contextos educativos o emocionales asistidos por inteligencia artificial.

A pesar de los avances recientes en el reconocimiento de emociones faciales, una limitación recurrente en la literatura es la dependencia de modelos entrenados y evaluados sobre un único conjunto de datos. Estudios recientes han demostrado que la combinación de múltiples datasets, como FER2013, RAF-DB, AffectNet y CK+, puede mejorar significativamente la capacidad de generalización de los modelos, especialmente en escenarios no controlados y con variabilidad cultural y demográfica. Sin embargo, este enfoque suele implicar arquitecturas más complejas y un mayor costo computacional, lo que dificulta su adopción en entornos con recursos limitados.

3 MÉTODOS

3.1 Conjunto de datos de entrenamiento

En esta etapa se empleó el conjunto de datos FER2013 (Facial Expression Recognition 2013), ampliamente utilizado en tareas de reconocimiento de emociones faciales. Este dataset está compuesto por 35 775 imágenes faciales en escala de grises, con una resolución de 48×48 píxeles, clasificadas en siete categorías emocionales: enojo (*angry*), asco (*disgust*), miedo (*fear*), felicidad (*happy*), tristeza (*sad*), sorpresa (*surprise*) y neutral (*neutral*) [11]. Las imágenes fueron recolectadas originalmente para competiciones de visión por computadora y están disponibles públicamente a través de la plataforma Kaggle.

Cada imagen de entrada se representó como un tensor bidimensional $X \in \mathbb{R}^{48x48}$, donde cada valor $X_{i,j}$ corresponde a la intensidad de píxel en la posición (i, j). Dado que las imágenes del conjunto FER2013 están en escala de grises, no se requiere un canal adicional para color, y por tanto, la representación tensorial se mantiene en dos dimensiones.

Para su procesamiento, las imágenes fueron organizadas en carpetas separadas por clase, facilitando su carga mediante generadores automáticos en Keras. Posteriormente, el conjunto fue dividido en dos subconjuntos: 80% para entrenamiento y 20% para validación, conservando la distribución original por clase.

Durante el preprocesamiento, los valores de píxel fueron reescalados al rango [0,1], y se aplicaron técnicas de aumento de datos en tiempo real. Estas incluyeron transformaciones como rotación, desplazamiento, escalado (*zoom*) y volteo horizontal, con el objetivo de mejorar la robustez y generalización del modelo ante variaciones en las expresiones faciales.

El volteo horizontal fue definido mediante la siguiente expresión:

$$V_{flip}(m,n) = V_I(m,(n_{max}-n)+1)$$

Donde V_{flip} representa la matriz de salida, VI la imagen original, m y n son los índices de fila y columna respectivamente, y n_{max} es el número total de columnas.

3.2 División del conjunto de datos.

Posteriormente, el conjunto de datos fue dividido en dos particiones: 80% para entrenamiento y 20% para validación. Esta división se realizó respetando la proporción original de muestras por clase, con el fin de mantener el equilibrio de clases y garantizar una evaluación justa del modelo durante el proceso de validación.

3.3 Configuración del entrenamiento

El modelo CNN fue entrenado utilizando el entorno gratuito de Google Colab, haciendo uso de una GPU NVIDIA Tesla T4, sobre un sistema operativo Ubuntu 20.04 de 64 bits, e implementado en TensorFlow 2.15 con soporte para CUDA 12.4.

Para el desarrollo se utilizaron las siguientes bibliotecas y entornos: Python 3.8, NumPy 1.24, Pandas 2.0, Matplotlib 3.7, Scikit-Learn 1.3, OpenCV-Python 4.8, entre otras dependencias necesarias para el preprocesamiento y la visualización de datos.

El entrenamiento se llevó a cabo durante 1000 épocas, con un tamaño de lote (*batch size*) de 32. Se utilizó el optimizador Adam con una tasa de aprendizaje de $1\,x\,10^{-4}$. Como función de pérdida se empleó la entropía cruzada categórica (*categorical cross-entropy*), y la precisión (accuracy) fue utilizada como métrica principal de evaluación.

El proceso completo de entrenamiento tomó aproximadamente 21 horas, considerando los recursos computacionales proporcionados por el entorno de ejecución.

3.4 Arquitectura del modelo propuesta

La arquitectura diseñada en este estudio está compuesta por cuatro bloques convolucionales, seguidos de capas densas, con el objetivo de extraer características discriminativas de imágenes faciales en escala de grises $X \in R^{48x48}$.

Cada bloque convolucional integra las siguientes operaciones: convolución 2D, normalización por lotes (*Batch Normalization*), función de activación ReLU, una capa personalizada de desenfoque (BlurLayer) implementada como un filtro pasa-bajos fijo, utilizada como mecanismo de regularización espacial sea $X \in R^{HXWXC}$, y agrupamiento espacial (*Pooling*, ya sea máximo o promedio).

Posteriormente, las capas densas procesan las representaciones extraídas y conducen a una capa de salida Softmax, la cual clasifica las entradas en una de las siete categorías emocionales: enojo, disgusto, miedo, felicidad, tristeza, sorpresa y neutral.

Para ofrecer una representación formal del flujo de operaciones, en la siguiente sección se detallan las formulaciones matemáticas que describen las funciones principales empleadas en el modelo.

Operación de convolución

$$y[i,j] = \sum_{m=0}^{k_h-1} \sum_{n=0}^{K_w-1} x[i+m, j+n] \cdot w[m,n]$$

Donde: x: imagen de entrada, w: núcleo de convolución, k_h: y K_w: alto y ancho del kernel, y $y[i,j]$: mapa de activación de salida en la posición $[i,j]$.

Función de activación ReLU

$$f(x) = max(0,x)$$

Donde: $f(x)$ es la salida activada y x representa la entrada. Esta función introduce no linealidad en la red neuronal al anular los valores negativos, manteniendo únicamente los positivos, lo que permite una propagación más eficiente del gradiente y mejora la capacidad de aprendizaje del modelo.

Agrupamiento máximo (Max Pooling)

$$MaxPool(f,s)_{i,j,c} = \left(x_{is:is+f-1,js:js+f-1,c}\right)$$

Donde: x: es el mapa de características de entrada, f: tamaño del filtro, s: stride (paso del filtro), i,j: posición espacial en la salida, c: canal correspondiente y el tamaño es el valor máximo en la región de tamaño fxf.

$$AvgPool(f,s)_{i,j,c} = \frac{1}{f^2} \sum_{k\epsilon is}^{is+f-1} \sum_{js}^{js+f-1} x_{k,l,c}$$

Donde: x: es el mapa de características de entrada (tensor tridimensional), f: representa el tamaño del filtro o ventana de agrupamiento s: es el stride o paso con el que se aplica el filtro, i,j: indican la posición espacial del elemento en el mapa de salida, c: representa el canal o profundidad del mapa de características y la operación consiste en calcular el promedio aritmético de los valores contenidos en la ventana de tamaño fxf ubicada en la región correspondiente del canal c.

Capa densa (Fully Connected)

$$z = wx + b$$

Donde: x: representa el vector de entrada proveniente de la etapa de aplanamiento (flatten) de las capas anteriores, w: es la matriz de pesos que parametriza las conexiones entre las n entradas y las m unidades de salida, b: corresponde al vector de sesgos que se suma de forma aditiva a cada neurona y z: representa la salida lineal del modelo antes de aplicar una función de activación.

Función de activación softmax para clasificación multiclase

$$\hat{y}_i = \frac{e^{z_i}}{\sum_{j=1}^{C} e^{z_j}}$$

Donde: $\hat{y}_i$: probabilidad predicha para la clase i, z_i: salida lineal (logit) correspondiente a la clase i, C: número total de clases y el denominador asegura que la suma de todas la salidas $\hat{y}_i$ sea igual a 1 (distribución de

probabilidad).

Arquitectura del modelo CNN propuesto

La Figura 1 resume la configuración estructural de la red neuronal convolucional diseñada en este estudio. El modelo consta de cuatro bloques convolucionales progresivos que integran operaciones de convolución, normalización por lotes, activación ReLU, capas de desenfoque (BlurLayer) y agrupamiento espacial (MaxPooling o AveragePooling). Posteriormente, se incorporan capas densas con regularización por dropout para reducir el sobreajuste.

Esta arquitectura se diseñó procurando un equilibrio entre la eficiencia en la extracción de características y una baja carga computacional, con el fin de facilitar su implementación en entornos embebidos o aplicaciones educativas móviles.

Justificación del diseño arquitectónico:

La arquitectura propuesta fue concebida con un enfoque progresivo en la extracción jerárquica de características faciales, considerando la baja resolución de las imágenes (48×48 píxeles) del conjunto FER2013. La red se compone de cuatro bloques convolucionales con un número creciente de filtros (32, 64, 128 y 256), siguiendo el principio de profundización progresiva mediante el apilamiento de filtros pequeños propio de las redes tipo VGG [16], lo cual permite capturar progresivamente patrones simples como bordes y texturas, hasta representaciones abstractas asociadas con emociones faciales.

Cada bloque integra operaciones de convolución, activación ReLU y normalización por lotes (Batch Normalization), lo que estabiliza el entrenamiento y mejora la velocidad de convergencia. Asimismo, se incorporó una capa personalizada de desenfoque (BlurLayer) diseñada para reducir el ruido local y suavizar las variaciones abruptas en la textura facial, lo cual es especialmente útil en entornos con condiciones de iluminación o calidad subóptimas. En su implementación, la BlurLayer aplica un filtro de suavizado (paso bajo) fijo y no entrenable, de tipo binomial, sobre el mapa de características antes del submuestreo con paso 2 (blur pooling); de este modo atenúa las componentes de alta frecuencia y reduce el aliasing introducido durante el downsampling, mejorando la invarianza del modelo ante pequeños desplazamientos de la entrada, siguiendo el principio de anti-aliasing propuesto por Zhang [24].

La combinación estratégica de capas de MaxPooling y AveragePooling busca equilibrar la preservación de estructuras dominantes (como cejas o labios) y una representación homogénea de regiones más difusas del rostro. Adicionalmente, se utilizó Dropout de 0.25 en los bloques convolucionales y de 0.50 en la capa densa, con el objetivo de mitigar el sobreajuste, particularmente en capas profundas que tienden a memorizar patrones del conjunto de entrenamiento.

La red culmina en una capa totalmente conectada (fully connected) con activación Softmax, que realiza la clasificación multiclase en siete categorías emocionales. Se trata de una elección estándar para clasificación multiclase, que proporciona distribuciones de probabilidad sobre las clases predichas.

En conjunto, se diseñó una arquitectura liviana, eficiente y reproducible, apta para su implementación en dispositivos de bajo consumo o aplicaciones educativas móviles, sin comprometer significativamente la precisión del modelo.

Además, se consideró la optimización de parámetros y pesos mediante inicialización He Normal, lo que favorece la propagación estable de gradientes en redes profundas. Esta técnica, junto con el uso de Batch Normalization, minimiza la posibilidad de desvanecimiento o explosión del gradiente, permitiendo entrenamientos más rápidos y consistentes.

Un aspecto relevante es que el diseño modular de la arquitectura facilita su adaptación a otros conjuntos de datos con diferentes resoluciones o variaciones culturales en la expresión facial. Gracias a su estructura escalable, es posible ajustar el número de filtros o capas para adecuarse a requerimientos de mayor precisión o a entornos con restricciones computacionales más estrictas.

Por su diseño, la capa BlurLayer está orientada a reducir el ruido local y actuar como una forma de regularización espacial [24], aunque su aporte cuantitativo individual no se evaluó mediante ablación en este trabajo. Esta motivación se alinea con el comportamiento esperado en escenarios reales de captura, como imágenes comprimidas o cámaras de baja calidad.

En conjunto, la arquitectura propuesta comprende 686 215 parámetros, de los cuales 684 999 son entrenables y 1 216 corresponden a estadísticas móviles no entrenables de las capas de normalización por lotes (las tres BlurLayer no aportan parámetros entrenables, al aplicar un filtro binomial fijo). El modelo ocupa aproximadamente 2,6 MB en memoria, una cifra notablemente inferior a la de arquitecturas referentes como VGG-16 (~138 M de parámetros) o ResNet-50 (~25 M). Esta eficiencia se logró mediante decisiones de diseño concretas: cuatro bloques convolucionales en lugar de arquitecturas más profundas (≥13 en VGG), filtros progresivos de 32 a 256 canales con kernels 3×3 y una única capa densa intermedia de 128 unidades —frente a las dos capas de 4 096 unidades de VGG—, lo cual confirma su naturaleza liviana y la viabilidad de su despliegue en dispositivos con recursos computacionales limitados.

Otro factor clave es que el balance entre MaxPooling y AveragePooling permite mantener la riqueza de las características discriminativas mientras se reduce la dimensionalidad de forma eficiente. Esto no solo disminuye el costo computacional, sino que también preserva la coherencia espacial necesaria para identificar emociones con sutilezas en la expresión.

Finalmente, la arquitectura se concibió pensando en la compatibilidad con técnicas de compresión y

cuantización, de manera que pueda integrarse fácilmente en frameworks de despliegue en dispositivos móviles o sistemas embebidos. Este enfoque garantiza que el modelo no solo sea preciso en entornos de laboratorio, sino también viable en aplicaciones reales donde la latencia y el consumo energético son factores determinantes.

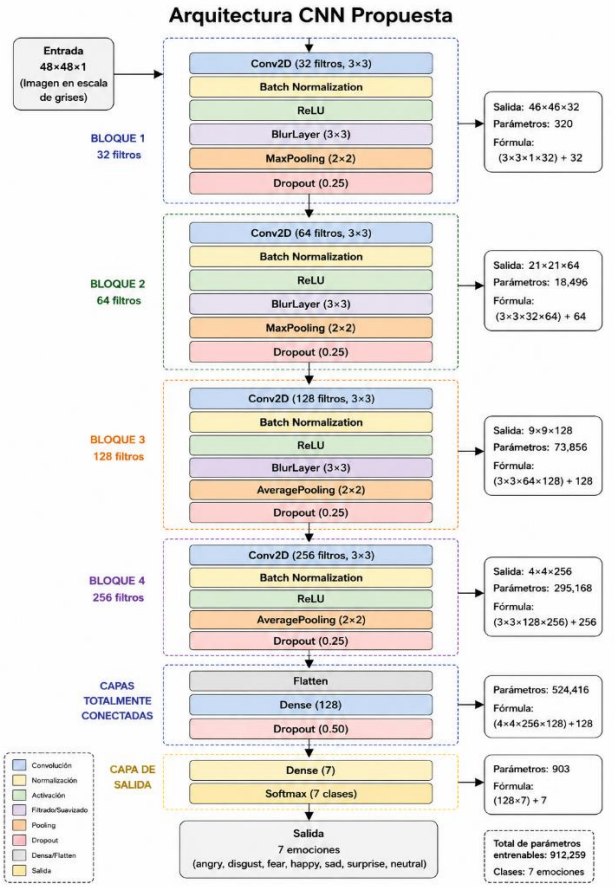

Fig. 1. Resumen estructural del modelo CNN propuesto, generado mediante Keras.

3.5 Distribución de clases

La Tabla I presenta la distribución del conjunto de datos FER2013 por clase emocional, detallando la cantidad de imágenes asignadas a los subconjuntos de entrenamiento y validación. Esta división respeta la proporción original de muestras por clase, lo que garantiza una evaluación equitativa del modelo.

TABLA 1
División de datos del conjunto FER2013 por clase emocional

N° clases	Clase Emocional	Entrenamiento	Validación	Total
0	Angry	3995	960	4955
1	Disgusted	436	111	547
2	Fearful	4097	1018	5115
3	Happy	7215	1825	9040
4	Neutral	4965	1216	6181
5	Sad	4830	1139	5969
6	Surprised	3171	797	3968
	Total	**28709**	**7066**	**35775**

Finalmente, las pruebas se realizaron sobre el conjunto de validación. La elección de una arquitectura convolucional liviana obedece al objetivo de facilitar su implementación en contextos reales, tales como aplicaciones móviles o plataformas educativas interactivas. Gracias a su tamaño compacto y baja carga computacional, el modelo propuesto resulta adecuado para entornos con recursos limitados, ofreciendo así una herramienta viable para el monitoreo emocional en aulas universitarias o sistemas de aprendizaje virtual.

4 EXPERIMENTACIÓN Y RESULTADOS

El modelo propuesto fue entrenado durante 1000 épocas utilizando el optimizador Adam con una tasa de aprendizaje de 1×10^{-4}, integrando mecanismos de regularización mediante Dropout y una arquitectura CNN optimizada con filtros progresivos y capas de desenfoque (BlurLayer). Durante el proceso de entrenamiento, se monitorizaron la precisión y la pérdida tanto en el conjunto de entrenamiento como en el de validación, empleando la herramienta TensorBoard para seguimiento en tiempo real.

Esta sección presenta y analiza los resultados obtenidos tras el entrenamiento y validación del modelo propuesto para el reconocimiento automático de emociones faciales. Las métricas utilizadas permiten evaluar su rendimiento tanto a nivel de clase individual como de forma global, proporcionando una perspectiva detallada del comportamiento del sistema.

4.1 Resultado por clase emocional

La Tabla II resume las métricas de clasificación obtenidas para cada una de las siete clases emocionales. Se observa un desempeño sobresaliente en las categorías Happy y Surprise, con valores de *F1-score* superiores a 0.87 y 0.79 respectivamente, lo que indica una alta capacidad de generalización del modelo para emociones de expresión distintiva. Por el contrario, las clases Fear y Sad muestran puntuaciones más bajas, posiblemente debido a la similitud de sus patrones faciales con otras emociones o a un menor soporte en el conjunto de entrenamiento.

TABLA 2
Métricas de clasificación por clase emocional

Emoción	Precisión	Recall	F1-score	Soporte (n)
Angry	0.5936	0.6010	0.5973	960
Disgust	0.8072	0.6036	0.6907	111
Fear	0.6542	0.3605	0.4649	1018
Happy	0.8795	0.8756	0.8775	1825
Neutral	0.5536	0.7344	0.6313	1216
Sad	0.5177	0.5663	0.5409	1139
Surprise	0.8036	0.7804	0.7919	797

4.2 Evaluación global del modelo

La Tabla III presenta las métricas agregadas, calculadas mediante promedios macro y ponderados (*weighted*) El modelo alcanzó una exactitud de 67.49% (IC 95%: 66.37%–68.58%) y un F1-score macro de 65.64% (IC 95%: 64.10%–67.05%).

TABLA 3
Promedios globales de clasificación

Métrica	Precisión	Recall	F1-score	IC 95%	Soporte (n)
Accuracy	—	—	0.6749	0.6637, 0.6858	7066
Macro average	0.6871	0.6460	0.6564	0.6410, 0.6705	7066
Weighted average	0.6841	0.6749	0.6708	0.6595, 0.6820	7066

4.3 Métricas generales de desempeño

La Tabla IV sintetiza los indicadores clave de desempeño del modelo sobre el conjunto de validación, destacando una exactitud total del **67,49%**. Si bien LHC-Net [18] presenta una mayor exactitud debido a su mayor complejidad, estos valores avalan la eficacia del enfoque propuesto en tareas multiclase de reconocimiento emocional.

TABLA 4
Resultado del rendimiento de la arquitectura

Modelo	Accuracy %	Precisión %	Recall %	F1-Score %
Pecoraro et al., LHC-Net [18]	0.7442	—	—	—
Hasani & Mahoor, Enhanced Deep 3D CNN [13]	0.7300	—	—	—
Tang, Linear SVM [17]	0.7116	—	—	—
Pascual et al., Light-FER [14]	0.6900	—	—	—
CNN propia (este trabajo)	0.6749	0.6958	0.644	0.654
oodfellow et Goodfellow et al., CNN [11]	0.6550	—	—	—

4.4 Visualización de resultados

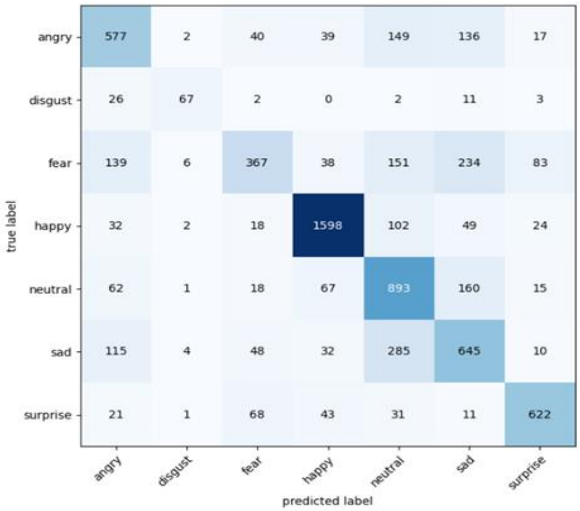

Fig. 2. Matriz de confusión de cada emoción sobre el conjunto de validación.

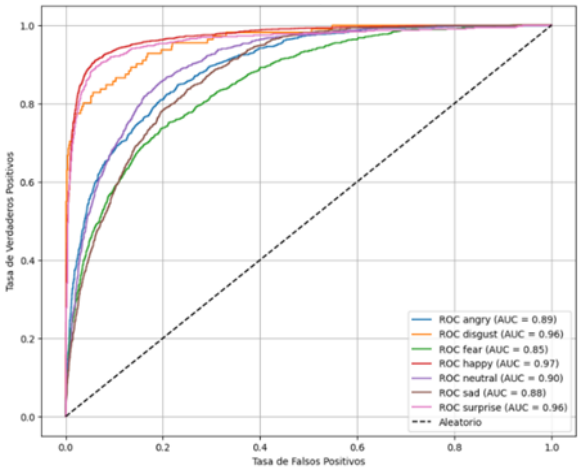

Fig. 3. Curvas ROC por clase sobre el conjunto de validación.

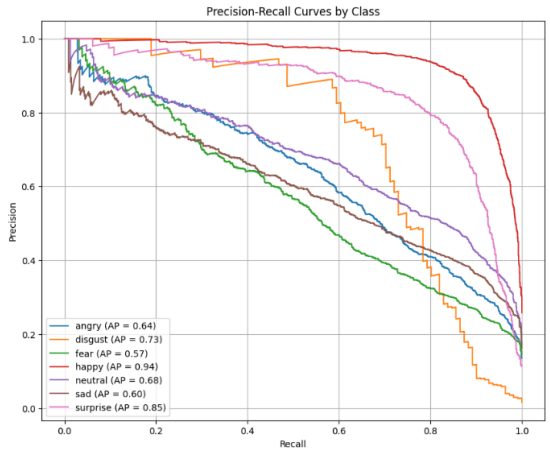

Fig. 4. *Curvas Precisión–Recall por clase sobre el conjunto de validación.*

5 DISCUSIONES

En este estudio se presentó una arquitectura liviana basada en redes neuronales convolucionales (CNN) para el reconocimiento automático de emociones faciales, entrenada y validada sobre el conjunto de datos FER2013. El modelo fue diseñado con un enfoque progresivo en la extracción de características jerárquicas, integrando técnicas como capas de desenfoque (BlurLayer), normalización por lotes y Dropout de 0.25 en los bloques convolucionales y 0.50 en la capa densa, lo cual favoreció la generalización y estabilidad del entrenamiento.

Los resultados obtenidos reflejan un desempeño competitivo. La matriz de confusión (Fig. 2) evidencia una alta capacidad del modelo para reconocer emociones como felicidad (*happy*) y sorpresa (*surprise*), con tasas de acierto destacadas. Sin embargo, también se identificó mayor confusión en emociones como *miedo (fear)* y *tristeza (sad)*, lo cual es consistente con observaciones previas en la literatura que señalan la ambigüedad inherente de dichas expresiones faciales y el reducido número de muestras de estas clases en el conjunto de datos [15], [17], [20], [3], [9].

Asimismo, las curvas ROC (Fig. 3) y de Precisión–Recall (Fig. 4) corroboran una elevada capacidad discriminativa global, con áreas bajo la curva (AUC) superiores a 0,85 en la mayoría de las clases. Esta tendencia también ha sido reportada en trabajos como los de Mollahosseini et al. [15] y Tang [17], donde se destaca la utilidad de estas métricas para evaluar el rendimiento de modelos multiclase, especialmente bajo condiciones de desbalance. Estudios recientes como los de Wang et al. [21] y Xu et al. [22] también confirman que estas curvas son esenciales cuando se trabaja con clases minoritarias o expresiones ambiguas.

Desde el punto de vista arquitectónico, la estructura propuesta guarda coherencia con diseños eficientes orientados a dispositivos embebidos o plataformas educativas, alineándose con enfoques como los de Goodfellow et al. [11], Hasani y Mahoor [13] y Pecoraro et al. [18], quienes también enfatizan la simplicidad estructural sin comprometer el rendimiento. El uso de una arquitectura liviana permitió alcanzar una exactitud del 67,49% y un F1-score macro de 65,64%, cifras que, si bien no superan a modelos más sofisticados como Vision Transformers (ViTs) o arquitecturas híbridas con módulos de atención [19], [20], ofrecen una solución balanceada entre rendimiento, coste computacional y facilidad de implementación [4], [8].

Dado que el conjunto de datos FER2013 es una base de datos pública ampliamente utilizada, resulta esencial contrastar el desempeño del modelo propuesto con trabajos representativos del estado del arte. En la literatura, diversos enfoques basados en redes neuronales convolucionales evaluados sobre FER2013 reportan valores de exactitud que oscilan aproximadamente entre el 65 % y el 72 %, dependiendo de la complejidad de la arquitectura y de las estrategias de entrenamiento empleadas [10], [12], [14]. En este contexto, la exactitud alcanzada por el modelo propuesto (67,49 %) y su F1-score macro (65,64 %) se sitúan dentro del rango reportado por estudios previos, confirmando su competitividad, especialmente considerando su naturaleza liviana y su menor complejidad computacional.

Finalmente, la disponibilidad del código fuente y documentación en un repositorio público [23] refuerza el compromiso con la reproducibilidad científica. Esto no solo permite su adaptación en entornos reales, como aulas universitarias o aplicaciones móviles de bienestar emocional, sino que también habilita futuras extensiones en contextos más complejos o multimodales [9].

La inclusión de la BlurLayer permitió reducir la sensibilidad del modelo a variaciones locales de alta frecuencia, actuando como una forma de regularización espacial sin incrementar el número de parámetros entrenables. Este enfoque se alinea con estudios previos que incorporan filtros de suavizado o anti-aliasing para mejorar la estabilidad y generalización de redes convolucionales [24].

5.1 Limitaciónes del estudios

El presente estudio presenta algunas limitaciones que deben considerarse al interpretar los resultados. En primer lugar, el modelo fue entrenado y evaluado únicamente sobre el conjunto de datos FER2013, por lo que su capacidad de generalización hacia otras bases de datos o escenarios reales requiere ser validada. Asimismo, no se realizó un estudio de ablación que permitiera cuantificar el aporte individual de componentes como la BlurLayer, el esquema de pooling o los niveles de Dropout en el desempeño final del modelo. Finalmente, aunque se desarrolló una implementación para detección de emociones en tiempo real, no se efectuó una evaluación sistemática en diferentes dispositivos o condiciones de operación, por lo que aspectos como la latencia, el consumo de

recursos y la robustez frente a variaciones ambientales deberán analizarse en investigaciones futuras.

6 CONCLUSIÓN

Este estudio presentó una arquitectura optimizada basada en redes neuronales convolucionales para el reconocimiento automático de emociones faciales, validada sobre el conjunto de datos FER2013. El modelo demostró una capacidad destacable para identificar emociones como Feliz (*Happy)* y sorprendido (*Surprise)*, mostrando un comportamiento robusto en clases bien representadas y una respuesta aceptable frente a categorías más ambiguas o desbalanceadas.

Las visualizaciones empleadas, incluyendo la matriz de confusión y las curvas ROC y Precision–Recall, permitieron evaluar de forma detallada el rendimiento por clase, evidenciando un equilibrio razonable entre precisión y recall. Estos hallazgos confirman la viabilidad del enfoque propuesto para su implementación en entornos reales, especialmente en contextos educativos o de bienestar emocional, donde se requiere un modelo liviano, reproducible y eficiente computacionalmente.

Como líneas futuras, se propone explorar mecanismos de mejora del desempeño en clases minoritarias mediante técnicas de balanceo de datos o generación sintética [20], [14]. Asimismo, la incorporación de información multimodal —como señales de voz, análisis de texto o expresión corporal— podría enriquecer la interpretación emocional en escenarios complejos [5], [10]. Además, se podrían investigar enfoques auto-supervisados o híbridos basados en atención liviana para mejorar la capacidad del modelo en condiciones no controladas [22], [14]. Cabe precisar que el presente estudio se limitó al conjunto de datos FER2013; por ello, una línea prioritaria de trabajo futuro es la integración de múltiples conjuntos de datos —como RAF-DB, AffectNet o CK+— mediante esquemas de entrenamiento inter-dataset o aprendizaje transferido, con el fin de mejorar la generalización del modelo en escenarios reales y no controlados. Asimismo, un estudio de ablación cuantitativo que evalúe el aporte individual de cada componente arquitectónico (BlurLayer, Dropout, normalización por lotes) se identifica como línea prioritaria para validar empíricamente sus contribuciones.

Finalmente, como demostración de la aplicabilidad del modelo propuesto, se desarrolló una interfaz funcional que permite la detección en tiempo real de emociones faciales utilizando visión por computadora. Esta implementación procesa las imágenes capturadas desde una cámara web, aplica el modelo entrenado y muestra en pantalla la emoción dominante reconocida junto con su porcentaje de probabilidad. Durante las pruebas alcanzó una velocidad promedio de 8 FPS, lo que lo hace viable para aplicaciones educativas, interactivas y de bienestar emocional.

El código fuente completo, junto con instrucciones para su ejecución y documentación técnica, se encuentra disponible en el siguiente repositorio de GitHub: https://github.com/PLINIORZAVALA/emotion-detector.git

Además, las Figuras 5 a 11 muestran ejemplos de la detección en tiempo real, incluyendo la emoción reconocida, su porcentaje de confianza y la velocidad de procesamiento (4–8 FPS):

Fig. 5. Detección de la emoción: enojo (*angry*)

Fig. 6. Detección de la emoción: felicidad (*happy*)

Fig. 7. Detección de la emoción: asco (*disgust*)

Fig. 8. Detección de la emición: tristeza (*sad*)

Fig. 9. Detección de la emoción: miedo (*fear*)

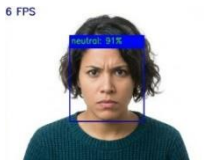

Fig. 10. Detección de la emoción: neutral (*neutral*)

Fig. 11. Detección de la emoción: sorpresa (*surprise*)

REFERENCIAS

[1] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," IEEE Trans. Pattern Anal. Mach. Intell., vol. 31, no. 1, pp. 39–58, 2009, doi: 10.1109/TPAMI.2008.52.

[2] S. Koelstra et al., "DEAP: A database for emotion analysis; Using physiological signals," IEEE Trans. Affect. Comput., vol. 3, no. 1, pp. 18–31, Jan. 2012, doi: 10.1109/T-AFFC.2011.15.

[3] S. Li and W. Deng, "Deep Facial Expression Recognition: A Survey," IEEE Trans. Affect. Comput., vol. 13, no. 3, pp. 1195–1215, 2022, doi: 10.1109/TAFFC.2020.2981446.

[4] C. A. Corneanu, M. O. Simón, J. F. Cohn, and S. E. Guerrero, "Survey on RGB, 3D, Thermal, and Multimodal Approaches for Facial Expression Recognition: History, Trends, and Affect-related Applications," IEEE Trans. Pattern Anal. Mach. Intell., vol. 38, no. 8, p. 1548, Aug. 2016, doi: 10.1109/TPAMI.2016.2515606.

[5] B. du Boulay, "Artificial Intelligence in Education and Ethics," Handb. Open, Distance Digit. Educ., pp. 93–108, Jan. 2023, doi: 10.1007/978-981-19-2080-6_6.

[6] B. Chul and K. Id, "A Brief Review of Facial Emotion Recognition Based on Visual Information," Sensors 2018, Vol. 18, Page 401, vol. 18, no. 2, p. 401, Jan. 2018, doi: 10.3390/S18020401.

[7] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression," 2010 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. - Work. CVPRW 2010, pp. 94–101, 2010, doi: 10.1109/CVPRW.2010.5543262.

[8] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks," IEEE Signal Process. Lett., vol. 23, no. 10, pp. 1499–1503, Oct. 2016, doi: 10.1109/LSP.2016.2603342.

[9] R. Ranjan, C. D. Castillo, and R. Chellappa, "L2-constrained Softmax Loss for Discriminative Face Verification," Jun. 2017, Accessed: Jul. 23, 2025. [Online]. Available: http://arxiv.org/abs/1703.09507

[10] M. Arul Vinayakam Rajasimman, R. K. Manoharan, N. Subramani, M. Aridoss, and M. G. Galety, "Robust Facial Expression Recognition Using an Evolutionary Algorithm with a Deep Learning Model," Appl. Sci. 2023, Vol. 13, Page 468, vol. 13, no. 1, p. 468, Dec. 2022, doi: 10.3390/APP13010468.

[11] I. J. Goodfellow et al., "Challenges in representation learning: A report on three machine learning contests," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 8228 LNCS, no. PART 3, pp. 117–124, 2013, doi: 10.1007/978-3-642-42051-1_16.

[12] L. Zahara, P. Musa, E. Prasetyo Wibowo, I. Karim, and S. Bahri Musa, "The Facial Emotion Recognition (FER-2013) Dataset for Prediction System of Micro-Expressions Face Using the Convolutional Neural Network (CNN) Algorithm based Raspberry Pi," 2020 5th Int. Conf. Informatics Comput. ICIC 2020, Nov. 2020, doi: 10.1109/ICIC50835.2020.9288560.

[13] B. Hasani and M. H. Mahoor, "Facial Expression Recognition Using Enhanced Deep 3D Convolutional Neural Networks," IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work., vol. 2017-July, pp. 2278–2288, Aug. 2017, doi: 10.1109/CVPRW.2017.282.

[14] A. M. Pascual et al., "Light-FER: A Lightweight Facial Emotion Recognition System on Edge Devices," Sensors 2022, Vol. 22, Page 9524, vol. 22, no. 23, p. 9524, Dec. 2022, doi: 10.3390/S22239524.

[15] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild," IEEE Trans. Affect. Comput., vol. 10, no. 1, pp. 18–31, Jan. 2019, doi: 10.1109/TAFFC.2017.2740923.

[16] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," 3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc., Sep. 2014, Accessed: Jul. 23, 2025. [Online]. Available: https://arxiv.org/pdf/1409.1556

[17] Y. Tang, "Deep Learning using Linear Support Vector Machines," Jun. 2013, Accessed: Jul. 23, 2025. [Online]. Available: https://arxiv.org/pdf/1306.0239

[18] R. Pecoraro, V. Basile, and V. Bono, "Local Multi-Head Channel Self-Attention for Facial Expression Recognition," Inf., vol. 13, no. 9, Nov. 2021, doi: 10.3390/info13090419.

[19] H. Li, M. Sui, F. Zhao, Z. Zha, and F. Wu, "MVT: Mask Vision Transformer for Facial Expression Recognition in the wild," Jun. 2021, Accessed: Jul. 23, 2025. [Online]. Available: http://arxiv.org/abs/2106.04520

[20] W. Niu, K. Zhang, D. Li, and W. Luo, "Four-player GroupGAN for weak expression recognition via latent expression magnification," Knowledge-Based Syst., vol. 251, p. 109304, Sep. 2022, doi: 10.1016/J.KNOSYS.2022.109304.

[21] C. Wang, J. Zeng, S. Shan, and X. Chen, "Multi-Task Learning of Emotion Recognition and Facial Action Unit Detection with Adaptively Weights Sharing Network," Proc. - Int. Conf. Image Process. ICIP, vol. 2019-September, pp. 56–60, Sep. 2019, doi: 10.1109/ICIP.2019.8802914.

[22] S. Bhattacharya, H. Li, J. Xia, and W. Xu, "SimPPG: Self-supervised photoplethysmography-based heart-rate estimation via similarity-enhanced instance discrimination," Smart Heal., vol. 28, p. 100396, Jun. 2023, doi: 10.1016/J.SMHL.2023.100396.

[23] D. Revelo Luna, *Face_Emotion*. GitHub repository. [Online]. Available: https://github.com/DavidReveloLuna/Face_Emotion. Accessed: Jul. 23, 2025.

[24] R. Zhang, "Making Convolutional Networks Shift-Invariant Again," in Proc. 36th Int. Conf. Mach. Learn. (ICML), vol. 97, pp. 7324–7334, 2019.