ISBN: 978-628-96613-3-0. ISSN: 2414-6390. Digital Object Identifier: https://dx.doi.org/10.18687/LACCEI2026.1.1.1217

# Intelligent Multimodal System for the Early Detection of Children's Needs and Emotions through Facial Expression Analysis

Hamer Garcia[1]; Standy Cerna[1]; Carlos Mendoza[1]; Jaime Llanos[1]; Laura Bazán[1]
[1]Universidad Privada del Norte, Perú, *N00373411@upn.pe, N00345033@upn.pe, eduardo.santos@upn.pe; jaime.llanos@upn.pe, laura.bazan@upn.pe*

***Abstract*– At present, verbal communication between infants and their caregivers constitutes a clinical and everyday challenge, since many first-time parents and daycare staff often make ambiguous interpretations of the signals used by babies to express their needs. The objective of this research was to develop an intelligent system capable of detecting and classifying infants' emotions in real time through the analysis of their gestures and facial expressions, as well as acoustic signals (crying), implemented and evaluated at the childcare center "El Nido." The study was applied in nature and, according to its approach, quantitative, with a pre-experimental design. The population consisted of 100 infants; however, a non-probabilistic convenience sampling method was used, selecting a sample of 52 infants. Finally, the results showed that the system achieved an 86% accuracy rate in facial landmark detection. Regarding comprehensive emotional classification, the system obtained an overall accuracy of 33.33% and an F1-Score of 32.54%. These metrics, while lower than laboratory models, demonstrate the feasibility of implementing lightweight architectures for supporting the interpretation of infant needs on low-cost hardware.***

***Keywords*— Verbal communication, infant needs, computer vision, deep learning, multimodal intelligent system.***

# Sistema Inteligente Multimodal para la Detección Temprana de Necesidades y Emociones Infantiles mediante Análisis de Expresiones Faciales

Hamer Garcia[1]; Standy Cerna[1]; Carlos Mendoza[1]; Jaime Llanos[1]; Laura Bazán[1]
[1]Universidad Privada del Norte, Perú, *N00373411@upn.pe, N00345033@upn.pe, eduardo.santos@upn.pe; jaime.llanos@upn.pe, laura.bazan@upn.pe*

***Resumen–*** ***Actualmente, la comunicación verbal entre bebés y sus cuidadores constituye un desafío clínico y cotidiano, ya que muchos padres primerizos y personal de guarderías suelen interpretar de forma ambigua las señales que utilizan los bebés para expresar sus necesidades. El objetivo de esta investigación fue desarrollar un sistema inteligente capaz de detectar y clasificar las emociones de los bebés en tiempo real mediante el análisis de sus gestos y expresiones faciales, así como de señales acústicas (llanto). Este sistema se implementó y evaluó en la guardería "El Nido". El estudio fue de naturaleza aplicada y, según su enfoque, cuantitativo, con un diseño preexperimental. La población estuvo compuesta por 100 bebés; sin embargo, se utilizó un método de muestreo por conveniencia no probabilístico, seleccionando una muestra de 52 bebés. Finalmente, los resultados mostraron que el sistema alcanzó una precisión del 86% en la detección de hitos faciales. Respecto a la clasificación emocional integral, se obtuvo una exactitud global del 33.33% y un F1-Score del 32.54%. Estas métricas, aunque inferiores a modelos de laboratorio, demuestran la viabilidad de implementar arquitecturas ligeras para el soporte en la interpretación de necesidades infantiles en hardware de bajo costo.***

***Palabras clave: Comunicación verbal, necesidades infantiles, visión artificial, aprendizaje profundo, sistema inteligente multimodal.***

## I. Introducción

En la actualidad la comunicación no verbal entre lactantes y sus cuidadores constituye un desafío clínico y cotidiano: el llanto, los gestos y las micro expresiones faciales son las principales señales que los bebés usan para expresar hambre, malestar, sueño o necesidad de consuelo, pero su interpretación es a menudo ambigua, especialmente para padres primerizos y personal de guarderías con alta carga laboral. Esta ambigüedad provoca retrasos en la atención apropiada y aumenta la probabilidad de respuestas inadecuadas o tardías por parte del cuidador, ya que el desarrollo de la comunicación en la primera infancia se basa en signos no verbales integrados que preceden al lenguaje y permiten expresar necesidades y emociones [1]. En el contexto de la primera infancia, el llanto y otros comportamientos no verbales constituyen formas tempranas de comunicación que señalan necesidades básicas y responden a la sensibilidad del cuidador [2]. El uso de gestos y otras señales no verbales ha sido documentado como parte del desarrollo comunicativo preverbal de los bebés y como puente hacia interacciones posteriores más complejas [3]. Estas interacciones tempranas de apego, mediadas por señales afectivas y no verbales, son clave para la formación de relaciones seguras y el bienestar emocional del niño [4].

La relevancia de abordar esta problemática se sustenta en evidencia epidemiológica y psicosocial: el estrés parental y la soledad en cuidadores han sido asociados con un deterioro del bienestar familiar y la calidad de la atención a la infancia [5]. Asimismo, cerca del 40% de los niños en ciertos contextos muestran vínculos emocionales débiles con sus cuidadores, lo que se relaciona con peores resultados en regulación emocional y desarrollo socioemocional a mediano plazo [6]. Estas cifras subrayan que la falta de detección temprana y adecuada de las necesidades infantiles no es un problema aislado, sino un factor de riesgo con implicaciones poblacionales para la salud mental y el apego.

Frente a esta necesidad, las soluciones tecnológicas han avanzado en varias direcciones complementarias: el reconocimiento automático del llanto mediante técnicas de Deep Learning ha mostrado altas precisiones utilizando características audiofónicas (por ejemplo, coeficientes MFCC y arquitecturas CNN/LSTM) para clasificar motivos de llanto como hambre o dolor [7]. Paralelamente, desarrollos en IoT han permitido crear dispositivos integrados y sistemas de monitoreo remotos desde cunas inteligentes con sensores hasta plataformas de vigilancia ambiental que facilitan la recolección continua de señales fisiológicas y ambientales [8], [9]. En el ámbito de la visión artificial, estudios recientes han explorado el análisis de expresiones faciales y movimiento corporal en lactantes aplicando modelos de aprendizaje profundo (incluyendo enfoques de visión por computadora para extraer rasgos faciales relevantes) con resultados prometedores de 97,41% en detección de estados afectivos [10]. Investigaciones que implementan enfoques multimodales y ensambles de modelos han demostrado que la fusión de audio, video y datos de sensores mejora la robustez y la exactitud de la detección respecto a modalidades aisladas [11]. Además, propuestas locales y aplicadas han comenzado a adaptar estos métodos a contextos reales de atención infantil, proponiendo arquitecturas

de sistemas inteligentes capaces de inferir emociones y necesidades en tiempo casi real [12].

A pesar de los avances reportados en las fuentes, persisten brechas relevantes para su aplicación práctica en entornos reales de cuidado infantil. Numerosas soluciones existentes han sido examinadas solamente en entornos controlados o se enfocan solamente en un tipo de análisis, por ejemplo, el audio del llanto o el reconocimiento facial de manera independiente. Esto restringe su habilidad para comprender con mayor profundidad la condición emocional del bebé. Asimismo, varios enfoques priorizan la precisión del modelo sin considerar su implementación en tiempo real mediante dispositivos accesibles, como cámaras integradas en sistemas locales, sin la validación en escenarios reales de guardería.

En este contexto, el presente trabajo tuvo como objetivo general desarrollar un sistema inteligente capaz de detectar y clasificar en tiempo real las emociones de bebés mediante el análisis de sus gestos y expresiones faciales, así como de señales acústicas (llanto), implementado y evaluado en el centro de cuidado infantil "El Nido", ubicado en el distrito de Jesús, Cajamarca, con el propósito de apoyar a las cuidadoras en la interpretación temprana y precisa de las señales emocionales infantiles.

Para alcanzar este objetivo, se plantearon los siguientes objetivos específicos: (i) Analizar la situación actual del centro de cuidado infantil "El Nido" en relación con las dificultades que enfrentan las cuidadoras para interpretar las emociones de los bebés a partir de sus expresiones faciales y vocalizaciones, (ii) Evaluar distintos algoritmos de inteligencia artificial aplicados al reconocimiento de emociones faciales y acústicas, considerando criterios de precisión, latencia y viabilidad de implementación en tiempo real, (iii) Desarrollar una solución basada en técnicas de visión por computadora y modelos de Deep Learning para la clasificación automática de gestos, expresiones faciales y patrones de llanto, y (iv) Evaluar el impacto del sistema propuesto en la reducción de la dificultad de interpretación emocional mediante pruebas realizadas en el entorno real del centro de cuidado infantil.

## II. METODOLOGÍA

La investigación realizada fue de tipo aplicada, debido a que recurre a los conocimientos científicos ya alcanzados en la investigación, para aplicarlos en la solución de un problema específico [13]. De acuerdo con el enfoque, la investigación fue cuantitativa con diseño pre-experimental, evaluando el sistema en cada etapa del proceso de desarrollo. La población estuvo conformada por los 100 lactantes del centro de cuidado infantil "El Nido", ubicado en el distrito de Jesús, perteneciente al departamento de Cajamarca. Se realizó un muestreo no probabilístico por conveniencia, seleccionando una muestra de 52 lactantes, debido a la accesibilidad brindada por el centro y la disponibilidad de la institución.

Para la recolección de datos, se utilizaron técnicas de observación tecnológica y métricas de rendimiento de algoritmos. Como instrumento de validación de las emociones detectadas, se empleó una ficha de cotejo contrastada con el juicio de las cuidadoras.

El sistema integró herramientas de visión computacional y Deep learning, tales como OpenCV, Deepface y MTCNN, cuya confiablidad en la detección de hitos faciales fue validada mediante pruebas de precisión (Accuracy), obteniendo un valor de 86%. El análisis e interpretación de los datos se realizó mediante el procesamiento de señales en tiempo real y análisis estadístico para comparar la tasa de aciertos del sistema frente a la interpretación humana manual.

Finalmente, para medir la eficiencia del modelo en dispositivos de borde (edge computing), se consideró la unidad de medida de Tiempo de Inferencia (ms) y el consumo de recursos computacionales, optimizando la arquitectura para operar bajo la siguiente relación de rendimiento (1).

$$Acc = \frac{VP+VN}{P+N} \quad (1)$$

Donde:

- VP: Verdaderos Positivos (Emociones correctamente identificadas).
- VN: Verdaderos Negativos.
- P / N: Total de casos positivos y negativos respectivamente.

Esta relación permite asegurar que el sistema no solo sea preciso en la detección, sino también viable para su ejecución en hardware de bajo costo dentro del entorno de la guardería.

Respecto a los factores éticos, se obtuvo el consentimiento informado de los padres o tutores legales de los menores, asegurando la confidencialidad de los datos biométricos y el uso exclusivo de las imágenes para fines de investigación académica, cumpliendo con la Ley de Protección de Datos Personales. Asimismo, el procesamiento de datos se realizó de forma local para garantizar la privacidad.

## III. RESULTADOS

### *A. Resultados del objetivo específico 1: Analizar la situación actual del centro de cuidado infantil "El Nido" respecto a las dificultades que enfrentan las cuidadoras para interpretar las emociones de los bebés a partir de sus expresiones faciales.*

La situación que presentó el centro "El Nido" se evaluó en base a las dificultades que enfrentan las cuidadoras para interpretar las emociones de los bebés, considerando las dimensiones de: a) Capacidad de identificación emocional (con sus indicadores de nivel de acierto en la interpretación y tiempo de respuesta ante el llanto) y b) Percepción de seguridad en el cuidado (con su indicador de nivel de confianza reportado). Estos resultados representaron el diagnóstico inicial (pre-test), obtenidos mediante una guía de observación y una encuesta estructurada, los cuales se muestran en la Tabla I.

TABLA I
LA SITUACIÓN QUE PRESENTA EL CENTRO "EL NIDO" EN LA DETENCIÓN DE NECESIDADES Y EMOCIONES INFANTILES ANTES (PRE-TEST) DEL USO DEL SISTEMA

| DIMENSIÓN | INDICADOR | RESULTADO |
|---|---|---|
| Capacidad de identificación emocional | Nivel de acierto en interpretación (1-5) | 2.1 puntos de los 5 posibles |
| | Tiempo de respuesta ante el llanto (min) | 4.5 minutos en promedio |
| Percepción de seguridad | Nivel de confianza en la interpretación (1-5) | 1.8 puntos de los 5 posibles |
| | Frecuencia de duda entre hambre/sueño (%) | 65% de las cuidadoras presentan duda frecuente |

Cabe destacar que la eficiencia de las cuidadoras para atender a las necesidades y emociones infantiles fue medida a través de la observación y una encuesta estructurada que permitía evidenciar si se cumplía con la necesidad infantil.

*B. Resultados del objetivo específico 2: Evaluar distintos algoritmos de inteligencia artificial aplicados al reconocimiento de emociones faciales y acústicas.*

Con el objetivo de identificar el algoritmo más adecuado para la detección de estados y expresiones infantiles en condiciones controladas, se compararon varios enfoques para el análisis facial y acústico, considerando como criterios principales: precisión (accuracy), latencia de inferencia (ms) y viabilidad de ejecución en tiempo real sobre hardware modesto.

Para la evaluación facial se consideraron métodos de detección/primer procesamiento (Haar Cascade mediante OpenCV y MTCNN) y clasificadores basados en redes CNN de diverso tamaño (ResNet preexistente en bibliotecas como DeepFace y MobileNetV2 implementado con TensorFlow/Keras). Para la clasificación acústica del llanto se evaluaron arquitecturas basadas en CNN y modelos preentrenados de audio como YAMNet (TensorFlow), utilizando MFCC y espectrogramas como características de entrada cuando fue necesario (TABLA II).

Los experimentos se llevaron a cabo sobre un conjunto de datos compuesto por 52 vídeos provistos por la institución (sin permiso para pruebas in situ) y complementado con muestras de bases públicas para entrenamiento y validación cruzada. Debido al tamaño limitado del set local, los resultados deben interpretarse como evidencia preliminar de comportamiento y no como evaluación en despliegue real.

TABLA II
COMPARACIÓN DE ALGORITMOS PARA EL RECONOCIMIENTO DE EMOCIONES FACIALES Y ACÚSTICAS (RESULTADOS PRELIMINARES)

| MODALIDAD | ALGORITMO / BIBLIOTECA | VIABILIDAD EN TIEMPO REAL |
|---|---|---|
| Facial | Haar Cascade (OpenCV) | Alta (detección rápida, baja complejidad) |
| Facial | MTCNN | Media (mejor robustez ante oclusión) |
| Facial | MobileNetV2 (TensorFlow/Keras) | Alta–Media (buena relación precisión/latencia) |
| Facial | ResNet (DeepFace) | Media–Baja (más preciso, mayor latencia) |
| Acústica | CNN simple (entrenada con MFCC) | Alta (eficiente en edge) |

Los resultados mostraron que los modelos de mayor capacidad (ResNet, CNN-LSTM) tienden a alcanzar precisiones levemente superiores, pero con una penalización importante en latencia. En función del balance entre precisión y respuesta temporal y considerando el tamaño y la naturaleza del conjunto de vídeo disponible se seleccionó la arquitectura final basada en:

OpenCV para las etapas de captura y preprocesamiento (detección rápida de rostros con Haar/algoritmos DNN según necesidad),

MobileNetV2 (TensorFlow/Keras) como clasificador facial por su buena relación precisión/latencia y facilidad de optimización para edge, y

YAMNet (TensorFlow) para clasificación acústica del llanto, aprovechando su preentrenamiento en múltiples clases de audio y su capacidad de fine-tuning.

*C. Resultados del objetivo específico 3: Desarrollo de la solución basada en visión por computadora y deep learning para la clasificación automática de emociones*

El desarrollo de la solución propuesta siguió un *pipeline* multimodal que integra información visual y acústica, diseñado para operar en dispositivos de borde con capacidades computacionales limitadas. El proceso se estructuró en tres componentes principales: el entrenamiento y ajuste fino de los modelos, el diseño del flujo de procesamiento del sistema y la estrategia de fusión multimodal junto con su despliegue.

Para el componente visual se adoptó un enfoque de transferencia de aprendizaje. Como punto de partida se utilizó el conjunto de datos FER2013, el cual contiene imágenes faciales etiquetadas en las clases *anger*, *disgust*, *fear*, *happy*, *sad*, *surprise* y *neutral*. Sobre esta base se entrenó inicialmente un clasificador convolucional basado en la arquitectura MobileNetV2, implementado en TensorFlow/Keras, aprovechando su eficiencia computacional y su capacidad de generalización en escenarios con recursos limitados. Posteriormente, se realizó un proceso de *fine-tuning* empleando exclusivamente imágenes de bebés, conformadas por un

conjunto local extraído de cincuenta y dos videos proporcionados y por muestras públicas filtradas por edad, con el objetivo de adaptar las representaciones aprendidas a las proporciones faciales y rasgos característicos de los lactantes. El flujo general del proceso de entrenamiento y ajuste fino del modelo visual se muestra en la Fig. 1.

.

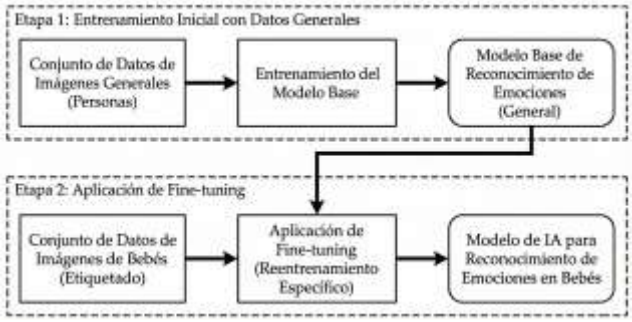


*Fig. 1 Diagrama de flujo del proceso de Entrenamiento de IA*

En paralelo, para el componente acústico se empleó YAMNet como modelo base, debido a su preentrenamiento en dominios amplios de audio. Este modelo fue ajustado utilizando fragmentos de llanto extraídos de los videos disponibles y datos públicos complementarios. El preprocesamiento del audio consistió en el remuestreo a 16 kHz, la segmentación en ventanas temporales de uno a dos segundos y la extracción de coeficientes MFCC, complementados con espectrogramas cuando fue requerido, junto con una normalización por canal. Dependiendo de la disponibilidad de datos, se ajustaron las capas superiores de YAMNet o se entrenó una red convolucional ligera directamente sobre las representaciones MFCC.

El flujo general de funcionamiento del sistema se ilustra en la Fig. 2, donde se presenta el *pipeline* completo desde la captura de imágenes RGB y audio hasta la obtención del resultado final. En la rama visual, las imágenes capturadas por la cámara son sometidas a un proceso de detección y preprocesamiento facial antes de ser clasificadas por el modelo MobileNetV2 ajustado, produciendo una salida probabilística asociada a la modalidad visual. De manera análoga, en la rama acústica, el audio capturado por el micrófono es segmentado, preprocesado y analizado por el clasificador basado en YAMNet, generando una medida de confianza asociada a la modalidad acústica.

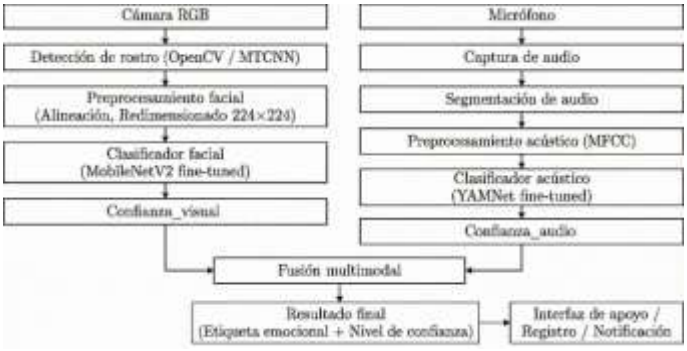


*Fig. 2 Diagrama de bloques del pipeline multimodal del sistema.*

Las salidas de ambas modalidades son combinadas mediante una estrategia de fusión multimodal basada en confianza ponderada. En este enfoque, las probabilidades normalizadas de cada clasificador se combinan utilizando pesos asignados a la modalidad visual y acústica, seleccionándose como resultado final la etiqueta emocional que maximiza la probabilidad conjunta. Como alternativas exploratorias, se evaluaron estrategias más avanzadas de fusión, tales como el *stacking* mediante un meta-clasificador y reglas heurísticas que priorizan la modalidad acústica en presencia de llanto sostenido.

Finalmente, como parte de la validación técnica y la documentación del sistema, se incluye una interfaz gráfica de apoyo para cuidadoras, presentada en la Fig. 3. Esta interfaz muestra la emoción detectada junto con su nivel de confianza y funciona como un elemento de apoyo para la toma de decisiones en entornos reales de cuidado infantil.


*Fig. 3 Mockup de la interfaz gráfica del sistema.*

*D. Resultados del objetivo específico 4: Evaluar el impacto del sistema propuesto en la reducción de la dificultad de interpretación emocional*

La evaluación del impacto del sistema propuesto se realizó mediante un análisis cuantitativo de los resultados generados automáticamente a partir del procesamiento de un conjunto de 52 videos infantiles. Dado que no fue posible ejecutar pruebas

en un entorno institucional real, la evaluación se centró en medir la consistencia del sistema, la reducción de ambigüedad y la coherencia entre variables emocionales, como indicadores del impacto potencial en la interpretación de estados infantiles.

En primer lugar, la distribución de emociones visuales dominantes detectadas por el sistema se presenta en la Fig. 4, donde se observa que el modelo fue capaz de identificar diversas emociones faciales, principalmente estados de sorpresa y neutralidad, así como reconocer explícitamente situaciones en las que no fue posible una detección facial confiable. Este comportamiento evidencia que el sistema evita forzar clasificaciones erróneas, lo cual es relevante para reducir interpretaciones incorrectas en contextos reales.

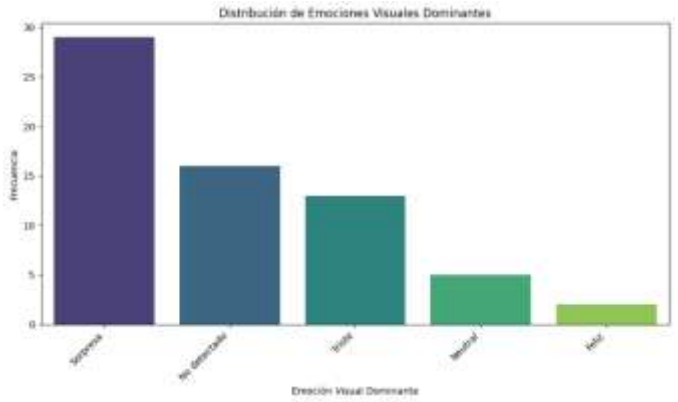


*Fig. 4 Distribución de emociones visuales dominantes detectadas por el sistema.*

Adicionalmente, la relación entre la emoción visual dominante y el nivel promedio de estrés estimado se muestra en la Fig. 5. Los resultados indican que emociones asociadas a estados negativos, como tristeza, presentan niveles de estrés superiores en comparación con estados neutrales o positivos. Esta coherencia entre las variables visuales y temporales refuerza la validez interna del sistema y demuestra su capacidad para integrar la información emocional de manera consistente.

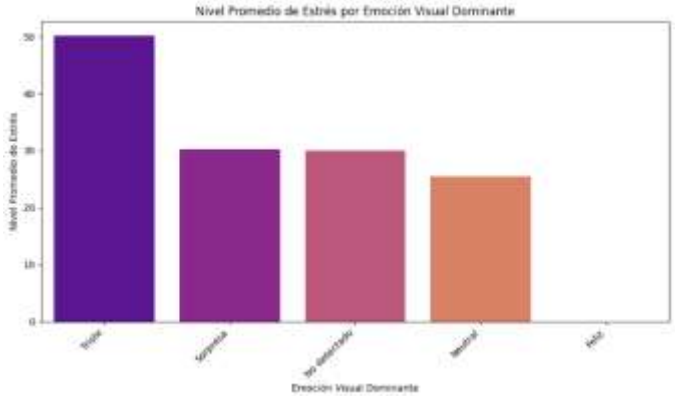


*Fig. 5 Nivel promedio de estrés en función de la emoción visual dominante.*

Finalmente, la síntesis automática realizada por el sistema se refleja en la distribución de conclusiones presentada en la Fig. 6, donde se aprecia que el sistema genera estados interpretativos comprensibles, tales como "Normal", "Feliz/Contento" e "Inquieto/Molesto", integrando múltiples variables sin depender exclusivamente de una única señal. Esta capacidad de síntesis contribuye a reducir la carga cognitiva asociada a la interpretación manual de señales no verbales.

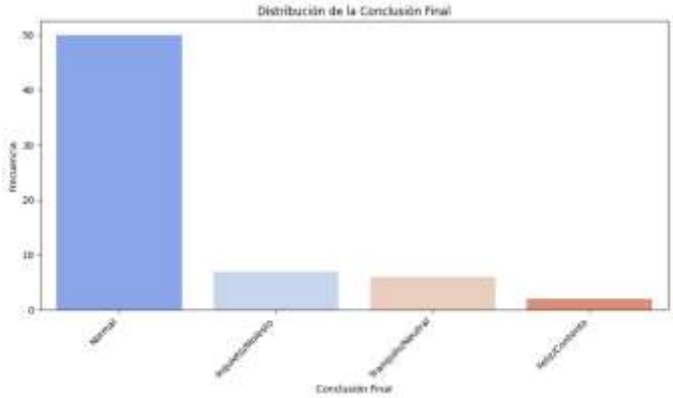


*Fig. 6 Distribución de la conclusión generada por el sistema multimodal.*

En conjunto, los resultados apoyados por las Fig. 4–6 muestran que el sistema propuesto posee el potencial de reducir la dificultad de interpretación emocional mediante una evaluación automática coherente, rápida y explicable. No obstante, estos hallazgos corresponden a una evaluación preliminar basada en datos controlados, por lo que se plantea como trabajo futuro la validación del sistema en entornos reales de cuidado infantil con participación directa de los cuidadores.

## IV. DISCUSIÓN

La presente investigación propone un sistema experimental diseñado para ser ejecutado en dispositivos de consumo masivo (computadoras de escritorio y dispositivos móviles), integrando análisis de audio y video. Los resultados obtenidos, la detección técnica de rasgos faciales mostró una alta fiabilidad del 86%, la clasificación final de emociones y necesidades alcanzó una exactitud del 33.33% y un F1-Score del 32.54%, contrastan con las métricas del estado del arte, evidenciando el desafío técnico que implica adaptar modelos de Inteligencia Artificial para operar bajo las restricciones de recursos propios de una aplicación de usuario final.

Al contrastar con soluciones unimodales de audio, Liang et al. [7] reportaron una exactitud superior (60-64%) clasificando causas del llanto mediante CNNs, sin embargo, su enfoque se centra puramente en la maximización de la clasificación acústica en entornos de laboratorio. A diferencia de ello, el presente sistema asume la complejidad computacional de procesar dos flujos de datos simultáneos (audio y video) en una

arquitectura pensada para ser desplegable en una app o software de escritorio, lo que obligó a sacrificar profundidad en el modelo acústico en favor de la integración multimodal.

La brecha es más notoria frente a arquitecturas de visión profunda de alto rendimiento. Cha [10] alcanzó un 97.41% de exactitud y Kristian et al. [11] un F1-Score del 99.1%. Es crucial destacar que estos antecedentes emplean arquitecturas densas (como ResNet y Ensambles de Autoencoders) que requieren una alta capacidad de cómputo, no apta para la ejecución fluida en dispositivos móviles estándar. La presente propuesta, al priorizar la portabilidad y la ejecución en hardware accesible, evidencia que las arquitecturas ligeras (necesarias para este tipo de aplicaciones) sufren una degradación significativa en el rendimiento (cayendo al 33.33%) al enfrentarse a la sutileza de las emociones negativas en lactantes, como se observó en la nula detección de la etiqueta "Enojado".

En comparación con soluciones de hardware dedicado, como el sistema IoT de Chandnani et al. [8], que reporta precisiones superiores al 98% utilizando sensores físicos (temperatura, gas), el presente trabajo aborda un problema de naturaleza distinta. Mientras ellos dependen de sensores externos costosos o específicos para obtener datos precisos, nuestra solución busca democratizar el monitoreo utilizando los periféricos estándar (cámara y micrófono) de dispositivos cotidianos. El bajo rendimiento obtenido confirma que depender exclusivamente de software para interpretar señales biológicas complejas sin sensores biométricos dedicados sigue siendo un reto abierto.

Finalmente, respecto al trabajo de Ríos Arévalo [12] (79.63% de exactitud en niños de inicial), la comparación resalta la dificultad inherente al sujeto de estudio. A diferencia de los niños en edad preescolar, cuyas expresiones faciales son socialmente aprendidas y más claras, los lactantes analizados en nuestro sistema presentan micro-expresiones ambiguas. Aunque el presente modelo logró un 55.6% de F1-Score en la detección de "Felicidad", el resto de las emociones se vieron afectadas por la variabilidad propia de los bebés y las limitaciones de un modelo optimizado para no saturar la memoria o procesador de un dispositivo común.

En conclusión, si bien los antecedentes demuestran que es posible alcanzar exactitudes superiores al 90% utilizando recursos ilimitados, el presente estudio delimita la línea base de desempeño para sistemas "ligeros" o accesibles. El 33.33% de exactitud indica que, para llevar estas tecnologías del laboratorio a una app funcional en el teléfono de un padre, se requiere aún optimizar drásticamente la relación costo-beneficio de las redes neuronales ligeras o integrar mecanismos de calibración personalizada para cada bebé.

## V. Conclusiones

Se analizó la situación actual del centro de cuidado infantil “El Nido” en relación con las dificultades que enfrentan las cuidadoras para interpretar las emociones de los bebés a partir de sus expresiones faciales y vocalizaciones, En este diagnóstico inicial, las cuidadoras alcanzaron apenas un promedio de 2.1 puntos (de 5 posibles) en su capacidad de identificación emocional, lo que refleja un bajo nivel de acierto en la interpretación. Asimismo, obtuvieron un 1.8 (de 5 puntos) en su percepción de seguridad, indicando un nivel de confianza muy reducido al momento de atender al infante. Por otro lado, el 65% de las cuidadoras manifestó tener dudas frecuentes sobre la interpretación correcta de las necesidades.

Se evaluaron distintos algoritmos de inteligencia artificial aplicados al reconocimiento de emociones faciales y acústicas, considerando criterios de precisión, latencia y viabilidad de implementación en tiempo real, optando por MobileNetV2. Posteriormente se desarrolló una solución basada en técnicas de visión por computadora y modelos de Deep Learning para la clasificación automática de gestos, expresiones faciales y patrones de llanto.

Finalmente, se evaluó el impacto del sistema propuesto en la interpretación emocional. Si bien la detección técnica de rasgos faciales mostró una alta fiabilidad del 86%, la clasificación final de emociones y necesidades alcanzó una exactitud del 33.33%. A pesar de esta cifra, el sistema demostró ser capaz de generar estados interpretativos coherentes (como 'Normal' o 'Inquieto') integrando audio y video en tiempo real. Esto representa un avance significativo respecto al diagnóstico inicial, donde las cuidadoras presentaban dudas frecuentes en el 65% de los casos, validando el potencial de la herramienta como soporte en entornos de cuidado infantil con recursos tecnológicos modestos.

## VI. Recomendaciones

Ampliar el tamaño de la muestra e incorporar lactantes de diferentes contextos socioculturales y rangos etarios, con el fin de mejorar la generalización de los resultados. Esto permitiría fortalecer la validez externa del sistema y reducir el sesgo asociado al muestreo no probabilístico por conveniencia utilizado en la presente investigación.

Se sugiere implementar el sistema en contextos reales durante periodos más extensos, integrándolo en la rutina diaria de centros de cuidado infantil y entornos clínicos. Esto permitirá evaluar su estabilidad, robustez y rendimiento ante variaciones ambientales reales como ruido, iluminación, movimiento y múltiples estímulos simultáneos.

Asimismo, incorporar sensores biométricos no invasivos (como frecuencia cardíaca, temperatura corporal o patrones respiratorios) para enriquecer el análisis multimodal. Esta integración permitiría mejorar la fiabilidad del sistema y reducir la ambigüedad en la detección de emociones complejas.

Finalmente realizar procesos de validación conjunta con profesionales de la psicología infantil, pediatría y neurodesarrollo, con el fin de contrastar las inferencias del

sistema con criterios clínicos especializados, fortaleciendo su aplicabilidad terapéutica y preventiva.

## AGRADECIMIENTOS

Un reconocimiento especial a la administradora y a todo el personal del centro de cuidado infantil "El Nido" en Cajamarca; su invaluable colaboración fue fundamental durante las etapas de recolección de datos y pruebas experimentales, permitiendo la validación del sistema en un entorno real. Finalmente, agradecemos a los padres de familia por su confianza y por permitir que este desarrollo tecnológico contribuya al fortalecimiento del bienestar y desarrollo emocional de los infantes.

## REFERENCIAS

[1] M. Díaz, "1.6 La comunicación no verbal en el desarrollo integral de los niños," en Comunicación no verbal en el desarrollo integral de los niños, Editorial Pueblo y Educación, 2023, pp. 66–69. Disponible en: https://editorial.redipe.org/index.php/1/catalog/download/206/357/7154 (consultado dic. 2025).

[2] H. Soto Hernández et al., "Evaluación de la comunicación social para infantes de 6 a 18 meses: síntesis de mapeo sistemático," Revista Neuropsicología Latinoamericana, vol. 17, no. 2, pp. 1–14, 2025.

[3] C. Farkas, "Desarrollo de la comunicación gestual intencionada en bebés: Estudio de un caso." https://summapsicologica.cl/index.php/summa/article/view/91

[4] L. Z. Asiain and E. B. Herreras, "Comunicación no verbal en la infancia: estudio comparativo infantes 4 años versus 6 años," *Indivisa Boletín De Estudios E Investigación*, no. 17, pp. 9–43, Jun. 2016, doi: 10.37382/indivisa.vi17.230.

[5] "ECIC 2024: La soledad y el estrés parental impactan el bienestar familiar y el desarrollo infantil en Perú – Copera Infancia," Oct. 18, 2024. https://coperainfanciaperu.com/es/ecic-2024-la-soledad-y-el-estres-parental-impactan-el-bienestar-familiar-y-el-desarrollo-infantil-en-peru/

[6] J. Rhodes, "Nearly 40% of US children lack strong emotional bonds with their parents," *The Chronicle of Evidence-Based Mentoring*, Aug. 08, 2024. https://www.evidencebasedmentoring.org/nearly-40-of-us-children-lack-strong-emotional-bonds-with-parents/

[7] Y.-C. Liang, I. Wijaya, M.-T. Yang, J. R. Cuevas Juarez, and H.-T. Chang, "Deep learning for infant cry recognition," *MDPI*, May 23, 2022. https://www.mdpi.com/1660-4601/19/10/6311 (accessed Dec. 27, 2025).

[8] K. Chandnani *et al.*, "A novel smart baby cradle system utilizing IoT sensors and machine learning for optimized parental care," *Scientific Reports*, vol. 15, no. 1, p. 19080, May 2025, doi: 10.1038/s41598-025-02691-8.

[9] H. Alam *et al.*, "IoT Based Smart Baby Monitoring System with Emotion Recognition Using Machine Learning," *Wireless Communications and Mobile Computing*, vol. 2023, pp. 1–11, Apr. 2023, doi: 10.1155/2023/1175450.

[10] J. Cha, "Deep learning based infant cry analysis utilizing Computer Vision," *ACADEMIA*, Jan. 2022, [Online]. Available: https://www.academia.edu/108011572/Deep_Learning_Based_Infant_Cry_Analysis_Utilizing_Computer_Vision

[11] Y. Kristian, N. Simogiarto, M. T. A. Sampurna, E. Hanindito, and V. Visuddho, "Ensemble of multimodal deep learning autoencoder for infant cry and pain detection," *F1000Research*, vol. 11, p. 359, Jan. 2023, doi: 10.12688/f1000research.73108.2.

[12] J. M. Rios Arevalo, "Sistema inteligente detector de emociones para evaluar el estado de ánimo y mejorar la enseñanza en niños de inicial, Comas," *Repositorio De La Universidad César Vallejo*, Dec. 06, 2024. https://repositorio.ucv.edu.pe/handle/20.500.12692/164431 (accessed Dec. 27, 2025).

[13] Quintero Cordero, Y. J., Molina Prendes, N., Bustillos Peña, M. A., Pastora Alejo, B. (2023). EL ENFOQUE DE COMPETENCIAS APLICADO A LA FORMACIÓN EN INVESTIGACIÓN CIENTÍFICA. *Revista Panamericana De Pedagogía*, *37*, 25-37. https://doi.org/10.21555/rpp.vi37.2852