Universidad de Valladolid

ESCUELA DE INGENIERÍAS INDUSTRIALES

UNIVERSIDAD DE VALLADOLID

ESCUELA DE INGENIERIAS INDUSTRIALES

Grado en Ingeniería en Electrónica Industrial y Automática

Determinación del estado emocional a través del análisis de video facial utilizando técnicas de Deep Learning

Autor:

Falcone, Andres Nolberto

Tutores:

De la Fuente López, Eusebio

Cisnal De la Rica, Ana

Departamento de Ingeniería de Sistemas y Automática

Valladolid, julio de 2025.

RESUMEN

El presente Trabajo de Fin de Grado se centra en el diseño, implementación y evaluación de un sistema automático para el reconocimiento del estado emocional a partir de vídeo facial dinámico mediante técnicas de Deep Learning. El objetivo principal ha sido explorar arquitecturas que integren tanto el análisis espacial como la codificación temporal de las expresiones faciales humanas, evaluando su rendimiento en entornos controlados y no controlados.

Para ello, se han desarrollado y comparado dos enfoques complementarios: una arquitectura basada en Vision Transformer (ViT-B/32), integrada en el marco multimodal CLIP y entrenada con la base de datos DFEW; y un sistema CNN+LSTM adaptado para su ejecución en tiempo real con entrada desde webcam. La arquitectura ViT-B/32 ha sido evaluada de manera formal mediante test directo y validación cruzada con la base MAFW, alcanzando métricas destacadas en emociones como felicidad o tristeza. Por su parte, el modelo CNN+LSTM ha demostrado una operatividad estable en escenarios en vivo, aunque sin evaluación cuantitativa formal.

Los resultados obtenidos demuestran la viabilidad de aplicar redes neuronales profundas al reconocimiento afectivo en vídeo, resaltando tanto las capacidades del sistema como las limitaciones asociadas a la detección de emociones ambiguas o poco representadas. Este trabajo sienta las bases para futuras aplicaciones en el ámbito de la salud, la interacción hombre-máquina y la inteligencia artificial afectiva.

PALABRAS CLAVE:

Reconocimiento emocional, Vídeo facial, Aprendizaje profundo, Vision Transformer, CNN-LSTM.

ABSTRACT

This Final Degree Project focuses on the design, implementation, and evaluation of an automatic system for emotion recognition from dynamic facial video using Deep Learning techniques. The main objective was to explore architectures capable of integrating both spatial feature extraction and temporal encoding of human facial expressions, and to assess their performance in both controlled and real-world scenarios.

To achieve this, two complementary approaches were developed and compared: a model based on the Vision Transformer (ViT-B/32) architecture integrated into the CLIP multimodal framework and trained with the DFEW dataset, and a CNN+LSTM system adapted for real-time inference via webcam input. The ViT-B/32 model was rigorously evaluated through direct testing and cross-validation on the MAFW dataset, achieving strong performance in well-defined emotions such as happiness or sadness. The CNN+LSTM model showed stable qualitative performance in real-time scenarios, although lacking formal quantitative evaluation due to dataset annotation constraints.

The results confirm the feasibility of applying deep neural networks to affective computing in video, highlighting both the strengths of the proposed system and the challenges associated with recognizing ambiguous or underrepresented emotions. This work lays a solid foundation for future applications in fields such as healthcare, human-machine interaction, and affective artificial intelligence.

KEYWORDS:

Emotion recognition, Facial video, Deep learning, Vision Transformer, CNN-LST

Índice General

Índice De Figuras

Índice De Tablas

CAPITULO 1

INTRODUCCIÓN

Este capítulo presenta el contexto general del trabajo, introduciendo la importancia del reconocimiento automático de emociones humanas a través del análisis de video facial.

Se expone la motivación del estudio, sus objetivos, el ámbito de aplicación, la metodología utilizada para el desarrollo y evaluación de los modelos, y finalmente se detalla la estructura global del documento.

1.1. Motivación

El reconocimiento automático del estado emocional humano a través de señales visuales ha emergido como una de las áreas más activas y prometedoras dentro del campo del aprendizaje profundo y la inteligencia artificial. Las expresiones faciales, al ser manifestaciones espontáneas y universales del estado afectivo, ofrecen una vía directa y no invasiva para interpretar la dimensión emocional de una persona. Esta capacidad resulta de gran interés en una amplia variedad de aplicaciones, que van desde interfaces hombre-máquina más intuitivas y empáticas hasta sistemas de apoyo en salud mental, educación emocional, atención al cliente o vigilancia del comportamiento.

En ámbitos como la medicina rehabilitadora, la psicología clínica o la robótica asistencial, poder interpretar el estado emocional del usuario de forma automática y fiable permite desarrollar soluciones tecnológicas más sensibles al contexto humano. Esta información puede servir, por ejemplo, para adaptar dinámicamente una terapia, detectar signos tempranos de malestar emocional, mejorar la comunicación no verbal o incluso prevenir riesgos en situaciones críticas. Asimismo, en el entorno educativo, la detección afectiva puede facilitar un aprendizaje personalizado, adaptado al estado emocional del estudiante.

A pesar de los notables avances obtenidos en los últimos años, el reconocimiento emocional continúa enfrentando importantes retos. Entre ellos destacan la variabilidad interindividual en la expresión de las emociones, la ambigüedad inherente a ciertas categorías afectivas, las diferencias culturales, el desbalance de clases en los conjuntos de datos y las condiciones no controladas propias de entornos reales. Estos factores dificultan tanto la precisión como la generalización de los sistemas, especialmente en escenarios dinámicos donde las emociones evolucionan en el tiempo y pueden ser sutiles o transitorias.

En este contexto, el presente trabajo se enmarca en una línea de investigación en continua evolución, con el propósito de avanzar en la creación de sistemas capaces de interpretar el estado emocional humano a partir del análisis facial en vídeo. Esta labor requiere una aproximación interdisciplinar y constante, que combine conocimientos técnicos con una comprensión profunda del comportamiento humano, y que permita desarrollar herramientas afectivas robustas, éticas y funcionales, destinadas a mejorar la calidad de vida en múltiples contextos sociales, educativos, médicos y tecnológicos.

1.2. Objetivos

El presente Trabajo Fin de Grado tiene como objetivo principal desarrollar y evaluar un sistema computacional capaz de detectar emociones humanas a partir de vídeo facial dinámico, utilizando arquitecturas avanzadas de aprendizaje profundo. Este sistema se enmarca en el campo del reconocimiento emocional automático, una disciplina de creciente interés en aplicaciones médicas, sociales, educativas y de interacción humano-máquina.

Para alcanzar este objetivo general, se han planteado los siguientes objetivos específicos:

Selección y adaptación de un modelo pre-entrenado potente, basado en una arquitectura de aprendizaje profundo, con el propósito de reutilizar su capacidad representacional en tareas de reconocimiento facial emocional.

Entrenamiento supervisado del modelo sobre una base de datos específica ampliamente utilizada en tareas de expresión emocional en vídeo, con el fin de optimizar la red para la clasificación multiclase de emociones básicas.

Evaluación cruzada sobre un dominio distinto, aplicando el modelo entrenado sobre otra base de datos, lo que permite examinar su capacidad de generalización a escenarios más complejos, con emociones espontáneas y diversidad cultural y contextual.

Cálculo y análisis de métricas cuantitativas de rendimiento, tanto individuales por clase (precisión, recall, F1-score) como globales (WAR, UAR, precisión macro y F1-score macro), con el objetivo de ofrecer una evaluación rigurosa y comparativa del modelo.

Implementación de un sistema de prueba funcional en condiciones reales, mediante la aplicación del modelo entrenado sobre vídeo en tiempo real capturado por webcam o mediante archivos grabados, evaluando su respuesta práctica y estabilidad.

Comparación de diferentes arquitecturas para la misma tarea, buscando aplicar también un segundo enfoque basado en otra arquitectura, con el objetivo de explorar soluciones híbridas que combinen extracción espacial y codificación temporal.

En conjunto, estos objetivos persiguen validar la viabilidad de arquitecturas avanzadas de Deep Learning para el reconocimiento automático de emociones faciales en vídeo dinámico, así como avanzar en el diseño de sistemas robustos, adaptativos y contextualizados. Más allá del simple desempeño técnico, el trabajo busca sentar las bases para el desarrollo de soluciones inteligentes afectivas capaces de interpretar el estado emocional humano de manera precisa, continua y en condiciones cercanas a escenarios reales. Este enfoque resulta especialmente relevante en entornos como la robótica médica asistencial, la monitorización psicológica, la educación emocional o la interacción humano-máquina.

1.3. Contexto del trabajo

El presente Trabajo Fin de Grado se inscribe dentro del campo emergente de la inteligencia artificial afectiva (Affective Computing), una disciplina interdisciplinar que combina visión por computador, aprendizaje profundo y neurociencia para dotar a los sistemas computacionales de la capacidad de interpretar y responder a las emociones humanas. En particular, este trabajo se centra en la detección automática de emociones a partir del análisis dinámico de vídeo facial, utilizando para ello arquitecturas de Deep Learning capaces de capturar tanto características espaciales como temporales en secuencias de imágenes.

La necesidad de sistemas capaces de identificar el estado emocional humano de forma precisa y en tiempo real ha crecido en paralelo con el desarrollo de tecnologías aplicadas a la salud digital, la interacción humano-robot, la atención al cliente y la educación emocional. En el ámbito médico, por ejemplo, la posibilidad de monitorizar las emociones de un paciente permite personalizar intervenciones psicológicas o evaluar el progreso de terapias en rehabilitación cognitiva. En entornos educativos, puede facilitar la adaptación del contenido en función del nivel de estrés o motivación del alumno, mientras que en la robótica social y asistencial contribuye a mejorar la empatía y la adaptabilidad de los sistemas autónomos.

El trabajo se desarrolla en un momento en el que los enfoques tradicionales basados en características manuales o expresiones faciales estáticas están siendo superados por métodos que integran el análisis de secuencias temporales. En este sentido, las redes convolucionales 2D combinadas con mecanismos de memoria como las LSTM, así como las nuevas arquitecturas basadas en Transformers (especialmente Vision Transformers), representan el estado del arte en el análisis afectivo dinámico. Estas técnicas permiten extraer información emocional de forma más robusta y contextual, incluso en condiciones de variabilidad expresiva, ruido o iluminación no controlada.

Por todo ello, este trabajo pretende situarse en la intersección entre la investigación teórica y la aplicación práctica de modelos de Deep Learning para el reconocimiento emocional facial, proponiendo soluciones viables que puedan ser escalables a sistemas reales. En particular, se explora su aplicabilidad en contextos médicos y de robótica asistencial, donde la comprensión del estado emocional del

usuario representa un valor añadido esencial para el desarrollo de tecnologías más humanas, éticas y efectivas.

1.4. Metodología

El desarrollo del sistema se ha estructurado en dos enfoques principales. El primero, basado en Vision Transformers (ViT-B/32) dentro del marco CLIP, ha sido entrenado sobre la base DFEW y validado mediante pruebas internas y validación cruzada en el dominio externo MAFW. Este enfoque ha permitido una evaluación cuantitativa precisa mediante métricas.

El segundo enfoque corresponde a una arquitectura CNN+LSTM basada en el trabajo de Elena Ryumina, adaptada e implementada para evaluar su rendimiento en tiempo real mediante webcam. Dado que no se dispone de una base de datos anotada cuadro a cuadro adaptada a este modelo, su evaluación se ha centrado en resultados cualitativos en vídeo. Ambos modelos han sido diseñados y testeados en un entorno modular que permite su futura implementación en plataformas más amplias y han demostrado ser capaces de operar en tiempo real sobre vídeo facial.

1.5. Estructura del documento

El presente Trabajo Fin de Grado se organiza en seis capítulos que permiten una comprensión progresiva del problema abordado, la metodología aplicada y los resultados obtenidos.

Capítulo 1: Introducción. Se exponen la motivación del trabajo, los objetivos perseguidos, el contexto en el que se enmarca la investigación, la metodología seguida y la estructura general del documento.

Capítulo 2: Marco teórico. Se detallan los fundamentos conceptuales del reconocimiento automático de emociones, los principios del aprendizaje profundo aplicados a visión por computador, las arquitecturas seleccionadas (ViT-B/32 y

CNN+LSTM), así como una revisión del estado del arte centrado en el análisis emocional dinámico.

Capítulo 3: Desarrollo del sistema. Se describen las bases de datos utilizadas, las decisiones técnicas adoptadas en el diseño del sistema, el proceso de implementación y entrenamiento de los modelos, así como las configuraciones específicas de cada enfoque.

Capítulo 4: Evaluación y análisis de resultados. Se presentan los resultados cuantitativos obtenidos mediante el modelo Vision Transformer, así como una evaluación cualitativa del sistema CNN+LSTM aplicado a escenarios en tiempo real mediante webcam o vídeo. Se incluyen análisis detallados mediante métricas y matrices de confusión.

Capítulo 5: Comparativa entre modelos. Se realiza un análisis comparativo entre ambos enfoques propuestos, identificando sus principales fortalezas, limitaciones y posibles aplicaciones.

Capítulo 6: Conclusiones y líneas futuras. Se reflexiona sobre el cumplimiento de los objetivos planteados, se sintetizan las conclusiones más relevantes del trabajo y se proponen líneas de mejora y ampliación, especialmente orientadas a la integración del sistema en entornos reales como la robótica médica asistencial o la monitorización emocional.

CAPITULO 2

ESTADO DEL ARTE

En este segundo capítulo de la memoria del Trabajo de Fin de Grado se realizará un breve repaso de diversos artículos y estudios previos relacionados con el proyecto, centrándose especialmente en investigaciones vinculadas al análisis de sentimientos mediante diferentes técnicas. Dicho análisis abarca distintos tipos de datos como texto, imágenes y videos.

Asimismo, se llevará a cabo una revisión y comprensión general de conceptos fundamentales como la Inteligencia Artificial, el Aprendizaje Automático (Machine Learning, ML por sus siglas en inglés) y el Aprendizaje Profundo (Deep Learning, DL por la misma razón). También se incluirá un análisis actualizado sobre el estado actual de la detección de emociones, explorando tanto el nivel de madurez de esta tecnología como los desarrollos más recientes en este campo.

Finalmente, se propondrá una reflexión crítica sobre las posibles ventajas y limitaciones que presentan estas tecnologías en la práctica.

2.1. El concepto de expresión facial

Las emociones faciales se refieren a la expresión visible que una persona muestra en su rostro como reflejo externo de su estado emocional interno, un mecanismo de comunicación no verbal crucial para la interacción social y la comprensión del estado emocional de los demás. Se ha identificado que ciertas expresiones faciales son universales y reflejan emociones básicas como alegría, tristeza, miedo, ira, sorpresa y repulsión. De esta manera, los individuos expresan como se sienten, facilitándose las interacciones sociales y la respuesta a su entorno.

2.2. La influencia de las emociones en la ingeniería de software

Históricamente, el interés científico en las expresiones faciales se remonta a los trabajos seminales de Charles Darwin en el siglo XIX, quien destacó su universalidad y su importancia adaptativa [1]. A partir de entonces, diversos investigadores han tratado de descifrar la complejidad inherente a las emociones humanas, dando lugar a enfoques estructurados. Muchos de ellos, aunque fueran rigurosos y ampliamente utilizado, requerían la intervención manual y la experiencia de codificadores entrenados, lo que limitaba su practicidad en aplicaciones a gran escala y en tiempo real.

La ingeniería de software (IS) ha centrado tradicionalmente su atención en cuestiones técnicas y en el desarrollo de metodologías. Con el auge de la computación moderna, los métodos clásicos basados en la extracción manual de características faciales dieron paso a técnicas más sofisticadas de inteligencia artificial. n los últimos años ha surgido un interés creciente por la influencia de los sentimientos y emociones, especialmente en entornos colaborativos y en metodologías ágiles, donde el componente humano desempeña un papel fundamental. Se ha comprobado que las emociones afectan directamente la productividad, la creatividad y la calidad del software desarrollado, lo que ha llevado a la necesidad de reconocerlas y gestionarlas de manera adecuada [2]. En particular, el desarrollo y la evolución del aprendizaje automático y posteriormente del aprendizaje profundo marcaron un punto de inflexión crucial en la automatización del reconocimiento de emociones faciales. Las redes neuronales convolucionales (CNNs, por sus siglas en inglés) emergieron como la piedra angular de estas nuevas metodologías, permitiendo no solo la automatización del proceso de extracción de características, sino también una mejora sustancial en la precisión y robustez del reconocimiento emocional.

2.3. Historia del estudio de las expresiones faciales

El interés científico por las expresiones faciales se remonta a más de un siglo atrás, impulsado por el reconocimiento de que los rostros constituyen un medio privilegiado para la comunicación de estados emocionales. Uno de los primeros tratamientos sistemáticos de este fenómeno fue realizado por Charles Darwin, quien en 1872 publicó *The Expression of the Emotions in Man and Animals [1].* En esta obra,

Darwin defendía que muchas expresiones emocionales humanas son herencia evolutiva y tienen correlatos observables en otras especies animales. Según su perspectiva, las expresiones faciales surgieron como adaptaciones biológicas que mejoraron las probabilidades de supervivencia, al permitir la transmisión rápida de emociones como el miedo o la agresividad, favoreciendo así la cooperación y la cohesión social [1].

Esta hipótesis evolucionista de Darwin contrastaba radicalmente con las ideas dominantes de su tiempo, que atribuían los comportamientos expresivos a normas sociales aprendidas. Aunque su propuesta sentó las bases para el estudio moderno de las expresiones emocionales, inicialmente tuvo un impacto limitado debido a la ausencia de métodos estandarizados para analizar de manera objetiva la expresión facial.

Fue en el siglo XX cuando la investigación en este campo adquirió nuevo impulso, en parte gracias a los avances metodológicos y al interés por la psicología intercultural. En este contexto destacó la figura de Paul Ekman, quien en los años 1960 condujo estudios transculturales que mostraron que ciertas expresiones faciales —asociadas a emociones básicas como alegría, tristeza, ira, sorpresa, miedo y repulsión— son reconocidas universalmente, incluso en comunidades aisladas de la influencia cultural occidental. Estos hallazgos reforzaron empíricamente la tesis de Darwin sobre la universalidad de las emociones expresadas facialmente [3].

Para facilitar el análisis sistemático de los movimientos faciales, Ekman y Wallace V. Friesen desarrollaron en 1978 el *Facial Action Coding System (FACS).* Este sistema codifica las expresiones faciales a partir de combinaciones de movimientos musculares básicos, conocidos como *Action Units* (AUs, por sus siglas en ingles), permitiendo describir cualquier expresión observable de forma precisa y estandarizada. A través de la combinación de distintas AUs, el FACS proporciona una herramienta robusta para la investigación psicológica, la neurociencia afectiva, las ciencias forenses y más recientemente, para aplicaciones de reconocimiento automático de emociones [4].

Además de Ekman, otros autores como Carroll Izard promovieron enfoques basados en modelos de emociones discretas, mientras que perspectivas posteriores introdujeron modelos dimensionales, como el de valencia-activación, que conceptualizan las emociones en un espacio continuo. Estos enfoques ampliaron el

estudio de las expresiones más allá de las emociones básicas, permitiendo abordar estados afectivos complejos como el orgullo, la vergüenza o el desprecio.

La llegada de las tecnologías digitales y el desarrollo de la Inteligencia Artificial provocaron una transformación en el estudio de las expresiones faciales. A partir de los años 2000, el uso de técnicas de ML permitió automatizar la detección y clasificación de emociones a partir de grandes volúmenes de datos faciales, superando las limitaciones humanas en cuanto a velocidad y objetividad de análisis. Posteriormente, la integración de modelos de DL, especialmente mediante CNNs, posibilitó el aprendizaje directo de patrones complejos desde imágenes faciales crudas, mejorando significativamente la precisión incluso bajo condiciones variables de iluminación, pose y demografía.

En la actualidad, el estudio de las expresiones faciales constituye un campo multidisciplinario que abarca la psicología, la biología evolutiva, la neurociencia computacional, la robótica social y el diseño de sistemas de interacción humano-máquina. Dentro de estas líneas de investigación, el análisis de micro-expresiones —movimientos faciales breves e involuntarios que revelan emociones auténticas— ha adquirido particular relevancia en ámbitos como la seguridad, la evaluación de la veracidad y el entrenamiento en inteligencia emocional [5].

Así, la trayectoria histórica del estudio de las expresiones faciales refleja un campo en constante expansión, donde las ideas fundacionales de Darwin siguen siendo validadas y ampliadas a través de metodologías contemporáneas basadas en la inteligencia computacional y el análisis intercultural.

2.4. Machine Learning

El ML, o aprendizaje automático, es una disciplina perteneciente al campo de la IA que estudia y desarrolla métodos computacionales capaces de aprender patrones y realizar inferencias automáticas a partir de datos, sin ser programados explícitamente para cada tarea concreta. Su fundamento radica en la construcción de algoritmos que, mediante procesos de entrenamiento, optimizan su desempeño ajustando sus parámetros internos para minimizar errores de predicción o clasificación [6].

2.4.1. Historia y evolución

El origen del ML se remonta a mediados del siglo XX. En 1943, Pitts y McCulloch modelaron las neuronas artificiales como unidades computacionales elementales. Posteriormente, Alan Turing (1950) propuso el célebre "Test de Turing" como criterio de inteligencia de las máquinas. En 1952, Arthur Samuel desarrolló uno de los primeros programas de aprendizaje automático, basado en el juego de las damas, que mejoraba su rendimiento con la experiencia. A finales de los años cincuenta, Rosenblatt introdujo el perceptrón, un precursor de las redes neuronales modernas. Tras varias décadas marcadas por los denominados "inviernos de la IA" debido a limitaciones tecnológicas, el renacimiento del ML se consolidó en la década de 2000, impulsado por el incremento de la capacidad computacional, la disponibilidad masiva de datos y avances teóricos significativos [7].

La evolución del ML ha seguido un proceso de especialización y expansión metodológica. Inicialmente centrado en algoritmos de regresión y clasificación supervisada, el campo incorporó posteriormente técnicas no supervisadas, de aprendizaje semi-supervisado y por refuerzo. Con el advenimiento del Big Data y el perfeccionamiento de las arquitecturas de procesamiento paralelo (especialmente mediante *GPUs*), emergió el Deep Learning como una extensión del ML tradicional, caracterizada por la utilización de redes neuronales profundas capaces de aprender representaciones jerárquicas de los datos [8].

2.4.2. Funcionamiento

Un sistema de ML generalmente se estructura en tres fases principales: adquisición de datos, entrenamiento del modelo y evaluación del rendimiento. Estas fases se adaptan a los distintos paradigmas de aprendizaje, que varían según la disponibilidad de información en los datos y la interacción del sistema con el entorno. Actualmente hay tres tipos de aprendizajes: supervisado, no supervisado y por refuerzo. A continuación, se describen cada uno de ellos.

El **aprendizaje supervisado** se caracteriza por el uso de datos de entrenamiento etiquetados. Cada ejemplo del conjunto de datos contiene una entrada (*features*) y su correspondiente salida deseada (*label o target*), lo que permite al modelo aprender una función de mapeo de entradas a salidas. El objetivo principal es encontrar una hipótesis que generalice adecuadamente para predecir correctamente la etiqueta de nuevas instancias no vistas.

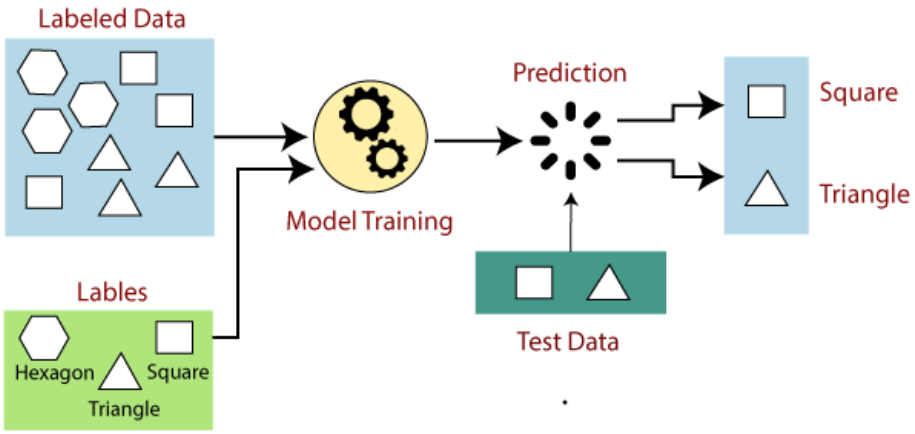

Figura 2.1: Diagrama de un aprendizaje supervisado para clasificación.
Fuente [53]

Durante el entrenamiento, el modelo realiza predicciones sobre los datos de entrada y compara estas predicciones con las etiquetas reales. El error entre la predicción y el valor verdadero se cuantifica mediante una función de pérdida (por ejemplo, error cuadrático medio para regresión o entropía cruzada para clasificación). A través de algoritmos de optimización como el descenso por gradiente, el modelo ajusta sus parámetros internos (por ejemplo, los pesos de una red neuronal) para minimizar la función de pérdida y mejorar su capacidad predictiva (véase la Figura 2.1).

Existen dos grandes tipos de tareas en el aprendizaje supervisado: la clasificación y la regresión. En la clasificación, el objetivo es asignar una etiqueta discreta a cada entrada, como en la identificación de emociones ("feliz", "triste", "neutral"). En la regresión, se busca predecir un valor continuo: el *arousal* o la *valence*.

El rendimiento del modelo entrenado se evalúa mediante métricas específicas como la precisión, la *recall*, la *F1-score* para tareas de clasificación, o el error medio absoluto (MAE, por sus siglas en inglés) y el error cuadrático medio (MSE, por el mismo

motivo) para tareas de regresión. Para asegurar la capacidad de generalización y evitar fenómenos como el sobreajuste (*overfitting*), es práctica común dividir los datos en conjuntos de entrenamiento, validación y prueba.

El aprendizaje supervisado es ampliamente utilizado en aplicaciones como el reconocimiento facial, la detección automática de emociones, el diagnóstico médico asistido por ordenador, el filtrado de spam y la predicción financiera, entre otros, gracias a su efectividad para resolver problemas donde se dispone de grandes cantidades de datos etiquetados de calidad [6].

El **aprendizaje no supervisado** es un paradigma del ML que se caracteriza por trabajar con datos no etiquetados, es decir, conjuntos de datos que no contienen una salida o resultado explícito asociado a las entradas. El objetivo principal es descubrir estructuras ocultas, patrones, agrupaciones o relaciones inherentes en los datos sin supervisión humana directa. A diferencia del aprendizaje supervisado, donde se aprende una función explícita de mapeo entre entradas y salidas, en el aprendizaje no supervisado el modelo infiere representaciones internas que explican la estructura de los datos (véase la Figura 2.2).

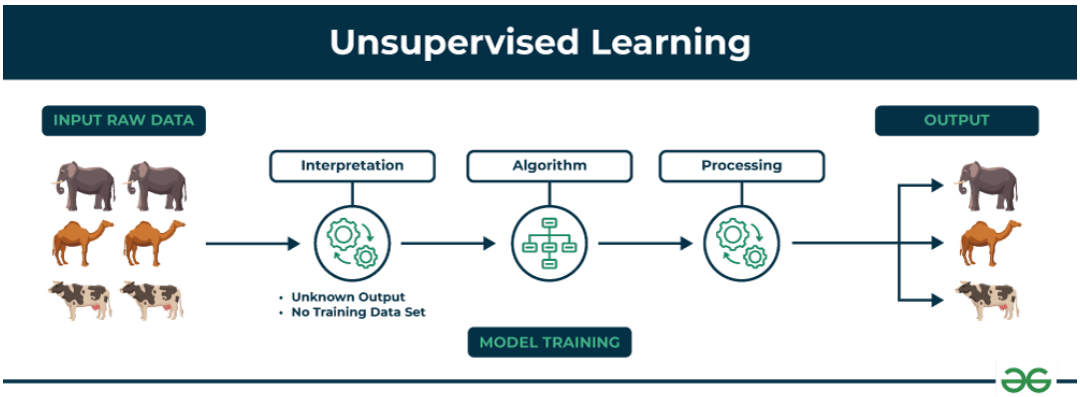

Figura 2.2: Diagrama de aprendizaje no supervisado
Fuente [54]

Las técnicas de aprendizaje no supervisado pueden clasificarse principalmente en dos grandes grupos: *clustering* y reducción de dimensionalidad. En *clustering*, el algoritmo agrupa las muestras en función de su similitud, siendo métodos representativos el algoritmo *k-means*, el *clustering* jerárquico y *DBSCAN* (*Density-Based Spatial Clustering of Applications with Noise*). Estos métodos buscan particionar

el espacio de datos en subconjuntos coherentes (*clusters*) sin necesidad de conocer de antemano las etiquetas. En reducción de dimensionalidad, el objetivo es encontrar una representación de los datos en un espacio de menor dimensión que preserve la mayor cantidad de información posible. Técnicas como el Análisis de Componentes Principales (PCA, por sus siglas en inglés) o los *Autoencoders* son ampliamente utilizadas para este fin.

Este aprendizaje es fundamental en escenarios donde el etiquetado manual de los datos sería costoso o inviable. Permite realizar análisis exploratorios, detectar anomalías, segmentar mercados, comprimir información y preprocesar datos para sistemas supervisados posteriores. Su éxito depende en gran medida de la calidad de las representaciones que aprende y de la capacidad del algoritmo para capturar relaciones semánticas relevantes dentro de los datos [6].

El aprendizaje por refuerzo es un paradigma del ML en el cual un agente aprende a interactuar con un entorno tomando acciones que maximizan una recompensa acumulativa a lo largo del tiempo. A diferencia del aprendizaje supervisado, donde se proporciona explícitamente el resultado correcto para cada entrada, en el aprendizaje por refuerzo el agente debe explorar las consecuencias de sus acciones y aprender del feedback, que recibe en forma de recompensas o penalizaciones.

Figura 2.3: Diagrama de aprendizaje por refuerzo
Fuente [55]

Formalmente, el entorno suele modelarse como un Proceso de Decisión de Markov (MDP, por sus siglas en inglés), definido por un conjunto de estados, un conjunto de acciones, una función de transición (que describe la probabilidad de moverse entre estados dados una acción), y una función de recompensa. El agente observa el estado

actual, selecciona una acción basándose en una política (estrategia de decisión) y, como resultado, transita a un nuevo estado recibiendo una recompensa (véase la Figura 2.3). El objetivo del agente es aprender una política óptima que maximice la suma esperada de recompensas futuras, conocidas como retorno.

El proceso de aprendizaje puede abordarse mediante diversos enfoques, como métodos de valor (por ejemplo, *Q-learning*), métodos de política (por ejemplo, *Policy Gradient*) o combinaciones de ambos (actor-critic). Además, algoritmos modernos como *Deep Q-Networks* (DQN, por sus siglas en inglés) y *Proximal Policy Optimization* (PPO, por el mismo motivo) han permitido aplicar el aprendizaje por refuerzo en entornos de alta dimensión y complejidad mediante el uso de redes neuronales profundas.

El aprendizaje por refuerzo ha demostrado ser particularmente efectivo en tareas donde la secuencia de decisiones afecta el resultado final, como el control de robots, los sistemas de recomendación dinámicos, la optimización de redes de comunicaciones, la automatización de estrategias financieras y el entrenamiento de agentes inteligentes en videojuegos y simulaciones complejas [9].

2.4.3. Aplicación en la detección de emociones

La detección automática de emociones representa una de las aplicaciones más dinámicas del ML actual, especialmente en el ámbito del análisis de expresiones faciales. Mediante el procesamiento de imágenes o vídeos, los algoritmos son capaces de identificar emociones básicas como alegría, tristeza, enfado, sorpresa o miedo. El flujo habitual de trabajo comprende la detección del rostro, la extracción de características relevantes (por ejemplo, posiciones relativas de los ojos, la boca y las cejas) y la clasificación de la emoción mediante modelos entrenados en grandes bases de datos. Inicialmente, se empleaban técnicas clásicas de ML como *Support Vector Machines* (SVM, por sus siglas en inglés) o *Random Forests*, utilizando vectores de características faciales manualmente definidos. Sin embargo, en los últimos años se ha observado una progresiva integración del ML con métodos de DL, lo que ha permitido mejorar significativamente la precisión y la robustez de los sistemas de detección emocional, gracias a su capacidad de aprender automáticamente representaciones más abstractas de los datos. Esta sinergia ha impulsado aplicaciones prácticas en sectores

como la salud mental, la interacción hombre-máquina, el marketing afectivo y la educación personalizada, estableciendo la detección emocional como un área estratégica de desarrollo tecnológico[10].

2.5. Deep Learning

El DL, también llamado aprendizaje profundo es una subdisciplina del aprendizaje automático que se caracteriza por el uso de redes neuronales artificiales con múltiples capas ocultas. Estas arquitecturas permiten transformar progresivamente los datos de entrada en representaciones internas de mayor nivel de abstracción, facilitando la detección automática de patrones complejos sin requerir la intervención humana en la selección manual de características [6]

A través de este enfoque jerárquico de procesamiento, el modelo es capaz de identificar relaciones no triviales en los datos, mejorando sustancialmente el rendimiento en tareas que tradicionalmente presentaban altos niveles de complejidad, como la traducción automática, el análisis del lenguaje natural o, en el caso de este trabajo, la búsqueda de la detección de emociones a través de reconocimiento facial.

2.5.1. Historia y evolución

Aunque la popularización del aprendizaje profundo es reciente, su desarrollo conceptual comenzó hace más de siete décadas. Ha experimentado una evolución significativa desde sus inicios en la década de 1940 hasta convertirse en una herramienta esencial en la inteligencia artificial moderna.

Sus orígenes se remontan a los años 40, cuando McCulloch y Pitts propusieron el primer modelo computacional de neurona artificial, una unidad lógica capaz de realizar operaciones booleanas básicas [11]. Este trabajo sentó las bases del conexionismo y de la idea de que sistemas compuestos por neuronas artificiales podían simular procesos cognitivos.

Poco después, Hebb formuló una teoría neuropsicológica sobre el aprendizaje, proponiendo que la repetida activación simultánea de dos neuronas fortalecía su conexión, lo que posteriormente sería conocido como "regla de Hebb" y serviría de inspiración para mecanismos de aprendizaje en redes artificiales [12].

En la década de 1950, Rosenblatt desarrolló el perceptrón, una red neuronal capaz de aprender a clasificar entradas en categorías mediante un algoritmo de ajuste de pesos. Este avance generó un considerable entusiasmo en la comunidad científica, al sugerir que las máquinas podrían aprender sin necesidad de programación explícita [13]. Sin embargo, ese entusiasmo se vio mermado tras la crítica de Minsky y Papert quienes demostraron que el perceptrón no era capaz de resolver problemas no lineales, como la función XOR, y que las redes de una sola capa eran fundamentalmente limitadas [14]. Esto condujo al llamado "invierno de la inteligencia artificial", una etapa de escepticismo y falta de financiación hacia las redes neuronales.

Durante los años 80, se produjo un resurgimiento del interés gracias al algoritmo de retropropagación del error (*backpropagation*), popularizado por Rumelhart, Hinton y Williams [15]. Este método permitió ajustar los pesos en redes de múltiples capas mediante el cálculo del gradiente descendente del error a lo largo de la red, resolviendo así el problema del aprendizaje en capas ocultas. Este avance teórico fue esencial para el desarrollo de redes profundas, aunque en esa época aún se enfrentaban a limitaciones tecnológicas y de datos.

Fue recién en la década de 2010 cuando el aprendizaje profundo comenzó a alcanzar su potencial, gracias a tres factores clave: la disponibilidad de grandes conjuntos de datos etiquetados, como ImageNet; la mejora del hardware computacional, en especial el uso de unidades de procesamiento gráfico (GPU); y la innovación en arquitecturas de redes neuronales profundas [16]. En particular, el modelo AlexNet, presentado por estos autores, marcó un punto de inflexión al ganar con amplia ventaja la competición ImageNet en 2012. AlexNet empleaba varias capas convolucionales, funciones de activación *ReLU*, *dropout* y entrenamiento en GPU, logrando una mejora sin precedentes en la precisión de clasificación.

Desde entonces, el aprendizaje profundo ha sido aplicado con éxito en numerosos campos, incluyendo visión por computador, procesamiento de lenguaje natural, sistemas de recomendación, medicina y conducción autónoma [16], [15]. En

particular, la evolución de arquitecturas como las CNN, redes neuronales recurrentes (RNN, por sus siglas en ingles), y más recientemente los transformadores, ha permitido abordar tareas cada vez más complejas con alta precisión.

Actualmente, la investigación en aprendizaje profundo se orienta hacia la mejora de la eficiencia computacional, la interpretabilidad de los modelos, y su robustez frente a perturbaciones y sesgos. Además, se exploran modelos generalistas que pueden adaptarse a múltiples dominios con entrenamiento mínimo, lo que refuerza el papel central del DL en el desarrollo de la inteligencia artificial[16].

2.5.2. Funcionamiento

La "profundidad" en el aprendizaje profundo se refiere al número de capas ocultas en una red neuronal.

Las redes neuronales artificiales (ANN, por sus siglas en inglés) son modelos computacionales inspirados en la estructura y funcionamiento del cerebro humano. Estas redes están compuestas por capas de nodos interconectados, denominados neuronas artificiales, que procesan información y aprenden de los datos mediante ajustes en los pesos de las conexiones. Cada neurona recibe entradas numéricas, aplica una función de activación para introducir no linealidad en el modelo, y transmite una salida hacia las siguientes capas. A través del ajuste iterativo de los pesos de conexión, basado en algoritmos de optimización, las redes neuronales aprenden a minimizar el error entre las predicciones realizadas y los resultados esperados [17].

En el ámbito del ML, las redes neuronales constituyen una de las múltiples técnicas utilizadas para modelar y predecir datos. En este ámbito, las redes superficiales pueden tener una o dos capas ocultas. Las redes profundas contienen múltiples capas (decenas o incluso cientos de capas), lo que les permite modelar relaciones no lineales complejas y capturar estructuras latentes presentes en grandes volúmenes de datos [18].

No obstante, cuando estas redes incluyen un gran número de capas ocultas (decenas o incluso cientos de capas) y parámetros ajustables entre la entrada y la salida, se consideran redes neuronales profundas (DNN, por sus siglas en inglés), y dan lugar al paradigma del DL (véase la Figura 2.4). Esta transición hacia una arquitectura más profunda a través del modelamiento de relaciones no lineales complejas y capturar estructuras latentes presentes en grandes volúmenes de datos al modelo permite abstraer representaciones jerárquicas cada vez más ricas. Lo que se traduce en un rendimiento significativamente superior frente a tareas de clasificación, segmentación, análisis secuencial y predicción en entornos de alta dimensionalidad [8].

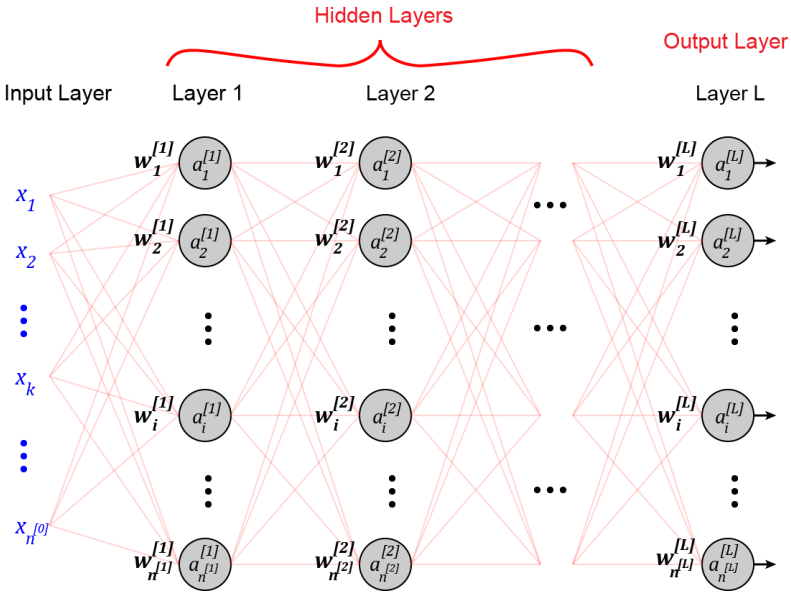

Figura 2.4: Esquema de una DNN y sus capas ocultas.
Fuente [56]

En las DNN, los pesos sinápticos constituyen el elemento clave que permite a la red aprender y generalizar a partir de los datos. Cada conexión entre dos neuronas, tanto en las capas ocultas como en la de salida, tiene asignado un valor numérico denominado peso. Este valor regula la influencia de una neurona sobre otra, determinando qué información es más relevante y debe ser reforzada, y cuál es menos significativa [6].

Durante el proceso de inferencia, cuando una muestra de entrada es introducida en la red, cada neurona multiplica sus entradas por los respectivos pesos asociados. Luego, se suma ese resultado ponderado y se le añade un valor de sesgo (*bias*), para

finalmente aplicar una función de activación no lineal, que permite a la red modelar relaciones complejas entre variables [19]. Esta operación se repite en cada capa, de modo que la información es progresivamente transformada hasta generar una salida final. Es aquí donde entra en juego el termino de entrenamiento de una red neuronal.

El entrenamiento de una DNN implica ajustar iterativamente los pesos sinápticos que conectan las neuronas, con el objetivo de minimizar una función de pérdida que cuantifica la discrepancia entre las predicciones del modelo y los valores reales. Este proceso se realiza mediante algoritmos de optimización como el descenso del gradiente, y se ve facilitado por técnicas como la retropropagación del error, que permite distribuir el error de manera eficiente a través de las capas de la red [15]. Se computan los gradientes de la función de pérdida respecto a cada peso, lo que permite ajustar dichos valores en la dirección donde el error se vea más reducido.

Este ajuste está regulado por un parámetro conocido como tasa de aprendizaje (*learning rate*), que define el tamaño del paso en cada actualización. Un valor demasiado alto puede provocar inestabilidad, mientras que uno muy bajo puede hacer que el entrenamiento sea lento o se estanque [6]. Gracias a este mecanismo de ajuste continuo, los pesos se convierten en los responsables directos del conocimiento adquirido por la red, ya que determinan qué patrones, características y relaciones ha aprendido el modelo a lo largo del entrenamiento.

En redes neuronales profundas, que pueden tener decenas o incluso cientos de capas, los pesos se organizan en matrices de gran tamaño y se optimizan con ayuda de bibliotecas computacionales especializadas que aprovechan el paralelismo de las GPU. A medida que los datos fluyen a través de la red, los pesos permiten extraer características cada vez más abstractas y representativas, lo que explica el alto rendimiento de estos modelos en tareas como clasificación de imágenes, reconocimiento de voz o análisis secuencial [8].

Dentro del paradigma del Deep Learning, existen diversas tipologías de redes neuronales y arquitecturas, cada una adaptada a diferentes tipos de datos y problemas específicos. Entre las más destacadas se encuentran las CNN, RNN y la arquitectura mediante *Transformers*. Además, las redes generativas adversarias (GAN, por sus siglas en inglés) han ganado popularidad por su capacidad para generar datos sintéticos

realistas en la actualidad. A continuación, se explicarán más en detalle, las más conocidas y fundamentalmente las empleadas para la realización de este proyecto.

2.5.3. Redes Neuronales Convolucionales (CNN):

Son una arquitectura especializada dentro del aprendizaje profundo, especialmente eficaz para el procesamiento de datos con estructura espacial o temporal, como imágenes, video o audio. A diferencia de las redes neuronales tradicionales de tipo *feed-forward* totalmente conectadas, las CNN explotan la localización espacial mediante capas de convolución,

El objetivo es extraer y transformar características relevantes de los datos de entrada. La primera de estas capas es la capa convolucional, en la cual se aplican filtros o *kernels* sobre la imagen u otra matriz de entrada. Cada filtro es una pequeña matriz de pesos entrenables que se desplaza (convoluciona) sobre la entrada, generando un mapa de activación que resalta ciertas características locales, como líneas, esquinas o texturas. Estos filtros aprenden automáticamente qué patrones son importantes para la tarea durante el proceso de entrenamiento. De esta forma se reduce drásticamente el número de parámetros necesarios. Lo que conlleva a que cada neurona de salida conecta solo con una pequeña vecindad de la entrada) y a la capacidad de detectar una característica sin importar su posición exacta. Estas propiedades permiten que la CNN identifique automáticamente las características relevantes en los datos sin supervisión manual, extrayendo jerárquicamente patrones simples en capas iniciales y patrones más complejos en capas profundas [20].

Originalmente inspiradas en la organización del córtex visual biológico, donde neuronas responden a regiones locales del campo visual, las CNN han demostrado un rendimiento sobresaliente en visón por computador, incluidas tareas como clasificación de imágenes, reconocimiento facial y procesamiento de voz.

Una CNN típica, además de estar compuesta por varias de estas capas de convolución mencionadas, posteriormente, incorporan unas capas de agrupamiento (*pooling*), cuya función es reducir la dimensión espacial de los mapas de activación, preservando la información más relevante y finalmente unas capas totalmente conectadas que actúan como clasificadores (véase la Figura 2.5). En cada convolución,

el filtro realiza un producto punto con una región de la imagen produciendo un mapa de características. Tras varias convoluciones, se insertan capas de *pooling* que reducen la resolución espacial de las características. La técnica de *pooling* más conocida es el *Max Pooling*.

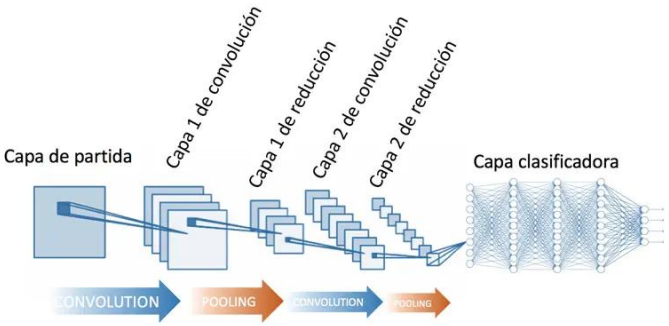

Figura 2.5: Esquema de una CNN y sus distintas capas.
Fuente [57]

El *Max Pooling* selecciona el valor máximo dentro de una ventana móvil, lo que aporta invarianza a traslaciones pequeñas en la entrada y reduce el coste computacional (véase la Figura 2.6). Las capas convolucionales y de agrupamiento pueden apilarse varias veces, generando representaciones cada vez más abstractas del contenido original [8].

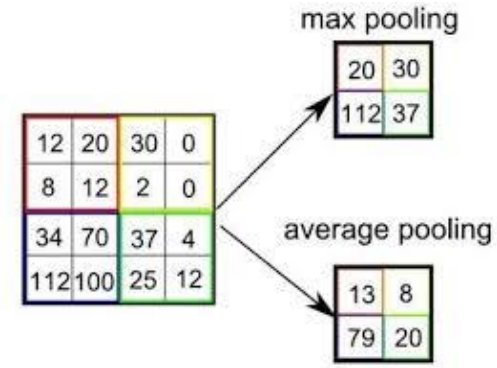

Figura 2.6: Diferencias entre *Max Pooling* y *Average Pooling*
Fuente [58]

Una vez extraídas las características relevantes, los mapas de activación resultantes se aplanan y se pasan a través de capas densamente conectadas (*fully connected layers),* que funcionan como clasificadores. Estas capas analizan las

características aprendidas para tomar una decisión final, como asignar una etiqueta de clase a una imagen. El proceso completo se entrena de extremo a extremo mediante retropropagación y optimización basada en el descenso del gradiente, ajustando los pesos de los filtros y las conexiones internas para minimizar el error entre la predicción y la salida esperada.

Gracias a esta estructura jerárquica y al aprendizaje automático de representaciones, las CNN han demostrado una eficacia sobresaliente en tareas de visión artificial, como el reconocimiento facial, la clasificación de objetos, la segmentación de imágenes médicas y el análisis de escenas en tiempo real.

Las redes neuronales convolucionales (CNN) han transformado profundamente el campo de la visión artificial. Una de sus principales ventajas radica en su capacidad para extraer características automáticamente de forma jerárquica, evitando así el diseño manual de descriptores. Además, gracias a su estructura local y al compartimiento de pesos, logran una reducción significativa del número de parámetros entrenables, lo que facilita la generalización del modelo y reduce el riesgo de sobreajuste. También permiten el entrenamiento de modelos profundos a gran escala de manera eficiente en comparación con redes densamente conectadas [20].

Otro aspecto fundamental es el uso de operaciones de *pooling*, que aportan cierta invariancia a traslaciones y pequeñas distorsiones en las imágenes, permitiendo conservar la información más relevante y reducir la sensibilidad a detalles irrelevantes [20]. Como resultado, las CNN han alcanzado un rendimiento sobresaliente en tareas de visión por computador, superando incluso el rendimiento humano en ciertos escenarios de clasificación de imágenes a gran escala.

Sin embargo, también presentan limitaciones. El entrenamiento de CNN profundas con millones de parámetros exige grandes volúmenes de datos y altos recursos computacionales, lo que puede dificultar su implementación en entornos con recursos limitados. Además, seleccionar adecuadamente la arquitectura, los hiper-parámetros y los métodos de regularización no es trivial, dado que la profundidad de la red introduce una complejidad considerable. Por otro lado, debido a que las CNN capturan principalmente patrones locales, necesitan redes muy profundas o técnicas adicionales (como el *global pooling*) para modelar relaciones espaciales de largo alcance. También se les atribuye el problema de ser poco interpretables, al actuar como

una “caja negra” cuyo funcionamiento interno no siempre puede explicarse fácilmente, a pesar de los avances en visualización de filtros o activaciones [21].

Las arquitecturas de redes neuronales convolucionales (CNN) han sido fundamentales en el avance del reconocimiento facial automático y en tareas complejas como la detección de emociones. Modelos como AlexNet, VGG y ResNet han demostrado un rendimiento notable al procesar imágenes mediante el aprendizaje jerárquico de características visuales [20]. Su evolución ha dado lugar a redes más eficientes como DenseNet y EfficientNet, capaces de extraer representaciones más profundas con menos parámetros.

En este contexto, el presente proyecto emplea la base de datos AffectNet, una de las colecciones más completas y utilizadas para el reconocimiento de emociones en imágenes faciales. AffectNet contiene más de un millón de imágenes anotadas con expresiones emocionales, valencia y activación, recopiladas en condiciones reales desde internet [22]. Su variedad y volumen permiten entrenar modelos CNN robustos para la clasificación de emociones, lo cual es fundamental para el desarrollo de este sistema de análisis del estado emocional a partir de video facial.

2.5.4. Redes Neuronales Recurrentes (RNN):

Las RNN son un tipo de arquitectura dentro del aprendizaje profundo diseñada específicamente para trabajar con datos secuenciales o temporales, como series temporales, texto, audio o video (véase la Figura 2.7). A diferencia de las redes neuronales tradicionales, estas tienen la capacidad de mantener información de pasos anteriores en la secuencia, gracias a conexiones recurrentes que permiten retroalimentar su propia salida hacia la entrada en el siguiente paso temporal .[6]

En una RNN, cada neurona en una capa no solo recibe datos de la entrada actual, sino también un estado oculto que contiene información del paso anterior. Esto le permite a la red tener memoria a corto plazo, ya que puede influir sus predicciones actuales con base en lo que ha visto previamente. Matemáticamente, el estado oculto

se actualiza en cada paso temporal combinando la entrada del momento con el estado anterior mediante funciones lineales y no lineales, como la activación tanh o ReLU [19].

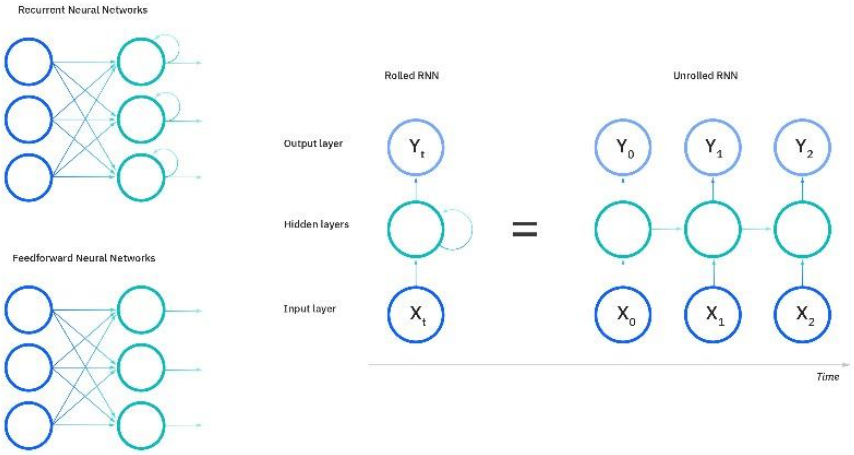

Figura 2.7: Redes neuronales prealimentadas vs RNN
Fuente [59]

Conceptualmente, una RNN puede desplegarse en el tiempo formando una secuencia de copias de la red, lo que equivale a una profundidad dinámica igual a la longitud de la secuencia. Gracias a esta capacidad, las RNN y sus variantes han sido ampliamente aplicadas en tareas como modelado de lenguaje, traducción automática, análisis de sentimientos, reconocimiento de voz, y en general en cualquier contexto donde el orden de los datos es relevante para la predicción [23].

Sin embargo, las RNN estándar presentan problemas cuando se trata de aprender dependencias a largo plazo, debido al fenómeno conocido como desvanecimiento o explosión del gradiente, que dificulta el ajuste adecuado de los pesos durante el entrenamiento [24]. Para abordar esta limitación, se introdujeron arquitecturas avanzadas que incluyen mecanismos de compuertas (*gating*) que regulan el flujo de información en la red: las LSTM y las GRU.

Las **LSTM** utilizan tres compuertas: entrada, olvido y salida. La compuerta de entrada decide cuánta información nueva se almacena en la memoria; la compuerta de olvido controla qué parte de la memoria anterior se descarta; y la compuerta de salida determina qué parte de la memoria se utiliza para generar la salida actual [25]. Gracias

a este diseño, las LSTM permiten que los gradientes se propaguen a lo largo del tiempo sin desvanecerse rápidamente, lo cual mejora la capacidad de aprendizaje en secuencias largas. Estas redes han demostrado un alto rendimiento en tareas como la traducción automática, el reconocimiento de voz o la generación de texto.

Las **GRU**, A diferencia de las LSTM, fusionan el estado oculto con la memoria y utilizan solo dos compuertas: una de actualización, que decide cuánta información nueva se incorpora, y otra de reinicio, que regula cuánto del estado anterior se conserva.

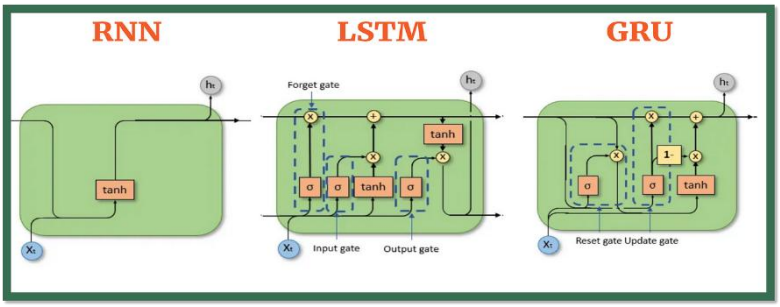

Figura 2.8: RNN vs LSTM vs GRU
Fuente [60]

Este diseño más simple permite un entrenamiento más rápido y eficiente, con un rendimiento comparable al de las LSTM en muchos casos [26].

En general, se ha observado que las GRU pueden igualar o incluso superar a las LSTM cuando las dependencias temporales no son muy largas, mientras que las LSTM ofrecen una leve ventaja cuando se requiere almacenar información durante más pasos. En cualquier caso, tanto LSTM como GRU representan mejoras sustanciales sobre las RNN simples, ya que mitigan el problema del desvanecimiento del gradiente mediante mecanismos internos que regulan el flujo de información (véase la Figura 2.8).

Una ventaja clave de las RNN es que pueden adaptarse a secuencias de cualquier longitud y procesarlas paso a paso, lo que las hace idóneas para aplicaciones como el procesamiento del lenguaje natural, el análisis de series temporales, el reconocimiento de voz y tareas en tiempo real como sistemas de streaming y robótica. No obstante, además del gradiente desvaneciente, las RNN no permiten un procesamiento paralelo eficiente, ya que cada paso depende del anterior. Esto ralentiza su entrenamiento e inferencia frente a arquitecturas paralelizables como los

Transformers [17].También presentan baja interpretabilidad, dificultando comprender qué patrones temporales específicos ha aprendido la red, lo que limita su uso en entornos donde se requiere explicabilidad [27].

Para superar algunas de estas limitaciones y mejorar la capacidad temporal de los modelos, se ha adoptado un enfoque híbrido que combina CNN con RNN. Las primeras son eficaces extrayendo características espaciales de imágenes o secuencias de vídeo, mientras que las segundas modelan su evolución temporal. Esta combinación ha sido eficaz en tareas como el análisis de video, reconocimiento de acciones o detección de emociones [20].

Aunque los *Transformers* han ganado terreno por su eficiencia paralela y capacidad para modelar relaciones de largo alcance sin recurrencia, las RNN —especialmente en combinación con CNN— siguen siendo una herramienta poderosa y válida dentro del aprendizaje profundo, especialmente en dominios donde el tiempo y la secuencia son elementos críticos.

2.5.5. Transformers:

Los *Transformers* son una arquitectura de aprendizaje profundo que revolucionó el tratamiento de secuencias al eliminar por completo el uso de recurrencias y convoluciones. Su base está en el mecanismo de auto-atención, que permite modelar relaciones entre elementos de entrada sin necesidad de procesarlos de forma secuencial [28]. Esto mejora tanto la eficiencia computacional como la capacidad para capturar dependencias a largo plazo, superando limitaciones comunes en redes recurrentes como el desvanecimiento del gradiente.

El modelo original propuesto por Vaswani et al. se compone de dos bloques principales: un codificador (*encoder*) y un decodificador (*decoder*). El *encoder* transforma la entrada completa en un conjunto de representaciones latentes mediante capas de atención y normalización. El *decoder*, a su vez, genera la salida paso a paso, utilizando atención sobre los elementos anteriores generados y también sobre la representación del *encoder* [28]. Esta estructura es especialmente eficaz en tareas de

traducción, pero también ha sido adaptada a clasificación, resumen automático, y procesamiento de vídeo (véase la Figura 2.9).

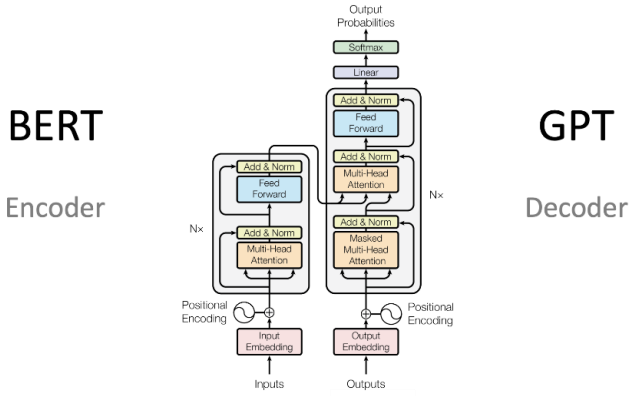

Figura 2.9: Diagrama de un *Transformer: encoder y decoder*.
Fuente [61]

Una ventaja distintiva de los *Transformers* es su capacidad de paralelización, ya que cada elemento de la secuencia se procesa en paralelo mediante el mecanismo de *self-attention*. Esto permite un entrenamiento mucho más rápido que en redes RNN, donde cada paso depende del anterior. Además, el uso de múltiples cabezas de atención (multi-head attention) permite aprender diferentes relaciones semánticas en paralelo.

En visión por computador, esta arquitectura ha dado lugar a modelos como los *Vision Transformers* (ViT), que dividen la imagen en parches y los tratan como tokens secuenciales, permitiendo aplicar el mismo modelo de atención originalmente diseñado para texto [29]. Con suficiente cantidad de datos, estos modelos han alcanzado o superado el rendimiento de las CNN en clasificación de imágenes.

Sin embargo, los *Transformers* también presentan desventajas. El mecanismo de auto-atención tiene una complejidad cuadrática respecto a la longitud de la secuencia, lo que aumenta significativamente el coste computacional. Para secuencias largas (como vídeos extensos o textos de muchos párrafos), esto puede volverse prohibitivo. Se están desarrollando variantes como *sparse attention* o arquitecturas más eficientes como *Performer* para reducir esta carga [30]. Por otro lado, no incorporan inductivos espaciales como las CNN (por ejemplo, la localización o la jerarquía visual), por lo que requieren mayor volumen de datos para

aprender regularidades que una CNN capta de forma más directa. Aun así, su flexibilidad y capacidad de generalización los han posicionado como una de las arquitecturas más versátiles del aprendizaje profundo actual.

La adopción de los *Transformers* ha impulsado una gran variedad de modelos especializados. En NLP (*Natural Language Processing*), destacan BERT (modelo codificador bidireccional pre-entrenado) y GPT (modelo autorregresivo generativo), que han establecido nuevos estándares en tareas lingüísticas. En visión, ViT (*Vision Transformer*) es el modelo más conocido, mientras que otros como DETR (*Detection Transformer*) y *Segmenter* aplican atención para detección y segmentación de objetos. También han surgido modelos multimodales como CLIP (*Contrastive Language–Image Pretraining*) o Florence, capaces de procesar simultáneamente texto e imágenes.

En el contexto de este trabajo, ViT modelar eficazmente la relación entre diferentes regiones del rostro a lo largo del tiempo, lo cual resulta útil para detectar emociones faciales en secuencias de vídeo.

2.5.6. Aplicación en la detección de emociones

El Deep Learning ha revolucionado el campo de la detección de emociones faciales, ofreciendo soluciones precisas tanto para el análisis de imágenes estáticas como de secuencias de vídeo. Las arquitecturas más empleadas en la literatura son las redes neuronales convolucionales (CNN), los modelos híbridos CNN+LSTM, las redes convolucionales 3D (3D CNN) y los *Transformers*.

Las CNN han sido esenciales para el reconocimiento de emociones en imágenes fijas. Estas redes aprenden de forma jerárquica: las primeras capas extraen patrones de bajo nivel (bordes, texturas), mientras que las más profundas capturan características faciales complejas asociadas con emociones específicas [20]. Gracias al uso de filtros locales y la compartición de pesos, las CNN son eficientes en cómputo y robustas ante variaciones de iluminación, escala o posición del rostro [21]. Han demostrado un rendimiento superior respecto a los enfoques tradicionales basados en descriptores manuales.

Para la detección de emociones en vídeo, se han explorado arquitecturas que capturan la dimensión temporal. Una de las combinaciones más exitosas es la de CNN con redes neuronales recurrentes como LSTM, donde la CNN extrae descriptores visuales de cada fotograma y la LSTM modela la evolución de estos vectores a lo largo del tiempo [31]. Esta estrategia permite reconocer emociones que se desarrollan de manera progresiva o fugaz, y es especialmente útil en aplicaciones como monitorización emocional o interfaces interactivas [32].

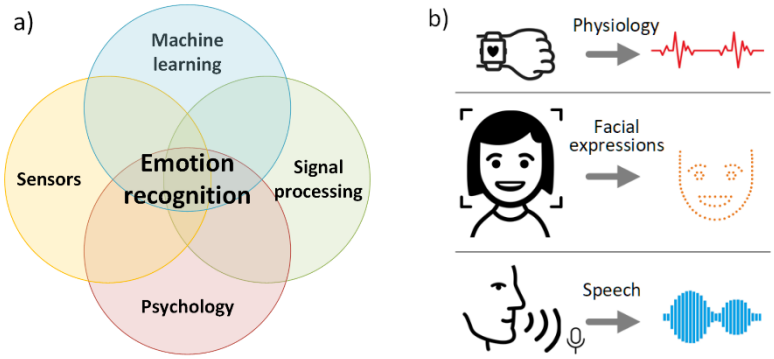

Figura 2.10: Componentes en el reconocimiento automático de emociones. Fuente [62]

Las 3D CNN también han demostrado eficacia en el análisis de vídeo, al aplicar filtros que consideran simultáneamente las dimensiones espaciales y temporales. Esto permite detectar directamente patrones de movimiento facial (como la progresión de una sonrisa), sin necesidad de una red recurrente adicional [33], [34]. Si bien estas redes requieren mayor capacidad computacional, su capacidad de integración temprana de la información temporal las hace muy potentes en contextos dinámicos.

En los últimos años, los *Transformers* han emergido como una alternativa destacada. Adaptados del procesamiento de lenguaje natural al dominio visual mediante *Vision Transformers (ViT)*, estos modelos dividen la imagen en parches y aplican auto-atención para determinar las regiones más informativas del rostro[28], [35]. A diferencia de las CNN, los Transformers pueden capturar relaciones espaciales de largo alcance, lo cual mejora la interpretación de expresiones complejas (véase la figura 2.11). En el ámbito del vídeo, se han desarrollado versiones que integran la dimensión temporal

dentro del mecanismo de atención, permitiendo analizar secuencias sin una estructura estrictamente secuencial [36].

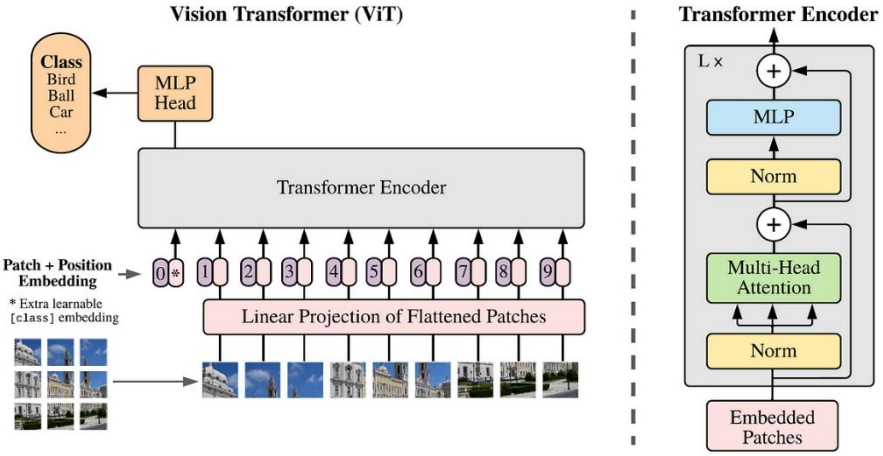

Figura 2.11: Diagrama de un Vision Transformer (ViT).
Fuente [63]

En síntesis, el DL ofrece un conjunto diverso y potente de arquitecturas para la detección emocional, adaptables tanto a imágenes estáticas como a secuencias de vídeo. La selección del enfoque depende de los requisitos de la aplicación, el tipo de datos disponibles y los recursos computacionales.

2.6. Métodos de DL para detección del estado emocional de manera dinámica

Cuando se busca detectar emociones en vídeo o en tiempo real, es fundamental capturar la dimensión temporal de las expresiones faciales. A diferencia de las imágenes estáticas, los vídeos contienen información sobre la progresión, la transición y la duración de las emociones, lo que permite una interpretación más precisa del estado emocional. En este contexto, se han desarrollado modelos específicamente diseñados para explotar esa información temporal.

Una estrategia habitual es el uso de arquitecturas híbridas como CNN+LSTM mencionada en el apartado previo. En este esquema, una CNN extrae representaciones

visuales de cada fotograma de una secuencia, y una red LSTM modela la evolución temporal de dichas representaciones. Esta arquitectura permite identificar no solo qué emoción está presente, sino cómo se desarrolla en el tiempo, lo que es especialmente valioso para emociones que emergen gradualmente o de forma ambigua (véase la figura 2.12). En el presente proyecto, este enfoque ha sido implementado para detectar emociones de pacientes a partir de vídeo facial, con el objetivo de dotar al sistema de sensibilidad ante cambios emocionales sutiles.

Otra arquitectura implementada en este trabajo es el ViT, concretamente el modelo ViT-B/32. Los Transformers aplicados a secuencias visuales pueden extenderse al dominio temporal mediante atención espacio-temporal, permitiendo comparar diferentes instantes de la secuencia sin la necesidad de seguir un orden fijo. Esto permite detectar relaciones globales y evaluar qué momentos del vídeo contienen información emocional relevante. Los ViT destacan por su capacidad para adaptarse a cambios de iluminación, poses o expresiones parciales, lo que mejora la generalización en condiciones no controladas.

En el análisis en tiempo real es crucial mantener un equilibrio entre precisión y eficiencia computacional. Las soluciones implementadas deben procesar fotogramas de forma continua, sin latencias perceptibles. Tanto CNN+LSTM como ViT pueden adaptarse a entornos semi en tiempo real, y con las optimizaciones adecuadas (por ejemplo, reducción de resolución, modelos ligeros, uso de aceleradores como GPU o TPU), es posible alcanzar tasas de procesamiento adecuadas para aplicaciones interactivas.

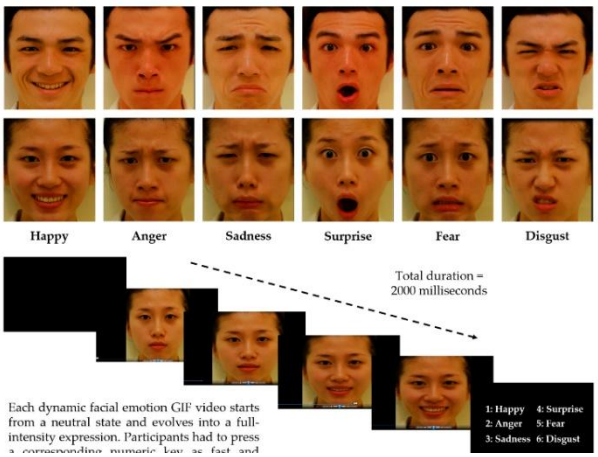

Figura 2.12: Secuencias de la evolución de seis emociones.
Fuente [64]

2.7. Ventajas e inconvenientes de la detección emocional

La detección automática de emociones mediante técnicas de DL ha transformado significativamente el análisis del comportamiento humano a través de la observación facial en imágenes, secuencias de vídeo y señales en tiempo real. Gracias al entrenamiento con grandes volúmenes de datos visuales, estos sistemas son capaces de identificar emociones básicas —como alegría, tristeza, miedo o sorpresa— con una precisión considerablemente superior a la de métodos tradicionales basados en reglas manuales o descriptores faciales [38]. Esta tecnología ha sido aplicada con éxito en áreas como la educación, el entretenimiento, el análisis del comportamiento del usuario y la atención al cliente, y está ganando relevancia en ámbitos sensibles como el bienestar emocional y la salud mental [37].

Una de las principales ventajas del DL radica en su capacidad para aprender automáticamente representaciones complejas a partir de los datos, sin necesidad de diseñar manualmente rasgos específicos como el ángulo de las cejas o la curvatura de los labios. Este enfoque no solo simplifica el proceso de extracción de características, sino que mejora la robustez del sistema frente a variaciones en iluminación, pose o calidad de la imagen [38]. Además, el procesamiento emocional en tiempo real se ha vuelto factible gracias a avances en hardware como GPU y TPU, permitiendo la integración de estos sistemas en plataformas interactivas que reaccionan dinámicamente a los estados emocionales del usuario. Esto tiene un impacto notable en contextos como la enseñanza personalizada o las interfaces adaptativas, donde detectar emociones como la frustración o el entusiasmo puede mejorar la experiencia del usuario [39].

No obstante, estas tecnologías presentan desafíos considerables. Uno de los principales es la dependencia de grandes conjuntos de datos etiquetados y diversos. La precisión del modelo está directamente relacionada con la representatividad de estos datos. En muchos casos, los *datasets* existentes contienen sesgos demográficos que afectan el rendimiento del sistema en ciertos grupos, particularmente minorías étnicas o poblaciones subrepresentadas [39]. Este sesgo algorítmico puede reproducir desigualdades sociales, comprometiendo la equidad del sistema e incluso perpetuando discriminaciones involuntarias.

Otro obstáculo importante es la escasa interpretabilidad de las redes neuronales profundas. Aunque su rendimiento en tareas de clasificación es elevado, su lógica interna resulta opaca, lo que dificulta comprender las decisiones tomadas por el modelo. Esta falta de transparencia —conocida como el problema de la "caja negra"— limita su implementación en entornos donde la trazabilidad y la explicabilidad son esenciales, como en aplicaciones clínicas o de diagnóstico psicológico [37].

También existen limitaciones relacionadas con los recursos computacionales. Los modelos de mayor rendimiento requieren un elevado poder de procesamiento, tanto en la fase de entrenamiento como en la de inferencia. Esto dificulta su implementación en sistemas con restricciones de hardware, como dispositivos móviles o robots autónomos. Si bien se han desarrollado técnicas para optimizar y reducir el tamaño de los modelos, aún persiste el desafío de mantener un equilibrio entre eficiencia y precisión [38].

Además de los retos técnicos, la adopción de sistemas de reconocimiento emocional plantea serias cuestiones éticas. La detección de emociones mediante webcams, por ejemplo, puede percibirse como una forma de vigilancia emocional, especialmente si se realiza sin el consentimiento explícito del usuario. Dado que las emociones forman parte de la intimidad del individuo, su análisis automatizado exige garantías sólidas de privacidad y confidencialidad, así como mecanismos de consentimiento informado. Estas preocupaciones son especialmente críticas en contextos como el educativo o el clínico, donde la relación de poder entre el sistema y el usuario puede influir en la percepción de libertad y participación [39].

Por otra parte, estudios recientes han identificado errores sistemáticos en la clasificación de emociones en función de la etnia o el color de piel. Por ejemplo, se ha observado que algunos modelos tienden a etiquetar las expresiones faciales de personas negras como más negativas que las de personas blancas, reproduciendo así sesgos raciales presentes en los datos de entrenamiento [39]. Este tipo de sesgos representa un riesgo ético significativo, ya que puede impactar negativamente a grupos vulnerables y comprometer la justicia y la equidad del sistema.

En contextos clínicos, como el del presente proyecto, donde se analizan las emociones de pacientes a través de vídeo facial, estas consideraciones éticas cobran aún mayor relevancia. La información emocional debe manejarse bajo estrictos criterios

de confidencialidad y con pleno respeto por los principios del consentimiento informado, la transparencia del sistema y la no maleficencia. Además, es esencial evitar la interpretación automatizada de estados emocionales como base única para decisiones clínicas, ya que las emociones humanas son complejas y pueden estar influidas por factores contextuales que escapan al análisis facial [40]

CAPITULO 3

METODOLOGÍA

En este tercer capítulo de la memoria se detalla de forma sistemática la metodología seguida para llevar a cabo el desarrollo e implementación del sistema de reconocimiento emocional a partir de vídeo facial. Constituye la base técnica y experimental sobre la que se asienta el análisis y los resultados presentados en los capítulos posteriores.

En primer lugar, se describen las bases de datos utilizadas, fundamentales para el entrenamiento y validación de los modelos. A continuación, se justifica la selección de los modelos de aprendizaje profundo empleados en el presente trabajo, así como sus características arquitectónicas. Finalmente, se presentan las métricas de evaluación seleccionadas para medir el desempeño de los modelos, tanto en términos de precisión como de robustez.

3.1. Bases de datos/*datasets*

A continuación, se enseñan las bases de datos empleadas para la realización del trabajo.

3.1.1. DFEW

La base de datos *Dynamic Facial Expressions in-the-Wild* (DFEW, por sus siglas en inglés) es una base de datos publicada en 2020, enfocada en el reconocimiento de expresiones faciales dinámicas (en video) captadas en entornos realistas o no controlados. Fue presentada por Xingxun Jiang y colaboradores en la conferencia ACM Multimedia 2020 como un nuevo conjunto de datos de gran tamaño para impulsar la investigación en reconocimiento de expresiones faciales en “el mundo real” [41].

El dataset DFEW consta exactamente con 16.372 clips de video de expresiones faciales dinámicas que han sido recopilados desde alrededor de 1.500 películas comerciales de diversas partes del mundo. Cada clip captura una expresión facial en condiciones de entornos reales, con variaciones significativas en iluminación, oclusiones y pose de la persona. Esto la convierte en una de las mayores bases de datos de expresiones faciales dinámicas hasta la fecha.

Figura 3.1: Secuencia de imágenes de un clip del dataset DFEW.
Fuente [Figura de elaboración propia a partir de la carpeta del dataset]

Si bien no se especifica un número exacto de sujetos únicos, es reconocida por tener una alta diversidad demográfica en cuanto a etnia, edad y género de las personas en los videos aportando diversidad en características faciales y contextos culturales.

Los clips varían en longitud. La mayoría son cortos (entre 2 y 5 segundos, y aproximadamente un 33% dura menos de 2 segundos), aunque también hay clips de menos de 2 segundos de duración o de más de 5 segundos cuando la expresión es prolongada (véase la Figura 3.1). Esta diversidad en duraciones permite cubrir desde micro-expresiones breves hasta reacciones emocionales más largas.

Cada clip está categorizado en 7 clases de emociones básicas (las siete emociones discretas clásicas de Ekman): Alegría (*Happy*), Tristeza (*Sad*), Ira (*Angry*), Sorpresa (*Surprise*), Asco/Repulsión (*Disgust*), Miedo (*Fear*) y Neutral.

En la tabla siguiente (véase la Tabla 3.1), se resume el número de clips con etiqueta definida en cada emoción.

Emoción	N.º de clips	Porcentaje
Neutral	2.709	22,5%
Felicidad	2.488	20,6%
Ira	2.229	18,5%
Tristeza	2.008	16,7%
Sorpresa	1.498	12,4%
Miedo	981	8,1%
Repulsión	146	1,2%

Tabla 3.1: Distribución de clips con etiqueta emocional única por categoría. Fuente [Tabla de elaboración propia]

La anotación de emociones en la base de datos DFEW se realizó mediante un riguroso proceso de etiquetado humano, con el objetivo de obtener tanto etiquetas individuales como distribuciones emocionales fiables. Para ello, se contrataron 12 anotadores expertos a través de una plataforma de crowdsourcing (JD Crowdsourcing, China), quienes recibieron formación específica en reconocimiento de expresiones emocionales. Su tarea consistía en identificar la emoción predominante en cada clip de vídeo, eligiendo entre las siete emociones básicas de Ekman: alegría, tristeza, ira, sorpresa, miedo, repulsión o neutral.

Cada uno de los 16.372 clips fue evaluado por un subconjunto de 10 anotadores de forma independiente. Esto generó una distribución de votos por clip, es decir, cuántos evaluadores seleccionaron cada emoción. Esta estrategia permitió captar no solo la emoción dominante, sino también posibles ambigüedades o mezcla de emociones percibidas. Para garantizar una base “limpia” con etiquetas claras, los autores definieron un criterio de mayoría: si una emoción era seleccionada por un número suficiente de anotadores (por ejemplo, al menos 6 de 10), esa emoción se asignaba como etiqueta única. Aplicando este criterio, se obtuvieron etiquetas únicas para 12.059 clips (alrededor del 74% del total). Los clips restantes se mantuvieron con su distribución de votos para reflejar la ambigüedad percibida.

Finalmente, la calidad del etiquetado fue evaluada mediante el coeficiente Kappa de Fleiss, una medida estadística de acuerdo entre anotadores múltiples. Los resultados mostraron un nivel de acuerdo sustancial (κ entre 0,61 y 0,80), siendo k el coeficiente mencionado. Esto confirma la fiabilidad de las anotaciones para su uso en investigación y entrenamiento de modelos de aprendizaje automático.

DFEW es una base de datos de acceso público para fines de investigación no comerciales. Los creadores la han puesto a disposición a través de un sitio web oficial (repositorio GitHub Pages) donde se pueden descargar los archivos, sujetos a ciertos términos y procedimientos que se han realizado en su totalidad. Cabe destacar que esta base de datos es ampliamente utilizada como referencia en estudios de reconocimiento de expresiones faciales dinámicas.

Se ha empleado para evaluar modelos como CNN 3D, CNN+LSTM y arquitecturas con atención o Transformers. También se combina frecuentemente con otras bases *in-the-wild* en tareas de aprendizaje transferido y evaluación cruzada. Su realismo y diversidad lo hacen útil tanto en contextos académicos como en aplicaciones prácticas como sistemas de interacción hombre-máquina o análisis emocional automatizado.

3.1.2. MAFW

La base de datos *Multimodal Affective Faces in the Wild* (MAFW, por sus siglas en inglés) constituye un recurso de referencia en el reconocimiento automático de expresiones faciales dinámicas en vídeo bajo condiciones naturales e incontroladas. Fue presentada por Liu et al. en 2022 como una propuesta innovadora que supera las limitaciones de bases estáticas o artificiales, al incorporar expresiones emocionales espontáneas en contextos realistas y multimodales [42]. A diferencia de conjuntos de datos posados o recogidos en laboratorio, MAFW introduce complejidades propias del mundo real, como la variabilidad en pose, iluminación, oclusiones parciales y riqueza cultural.

MAFW está compuesto por 10.045 clips de vídeo extraídos de múltiples fuentes audiovisuales: películas, series de televisión y vídeos breves procedentes de una amplia gama geográfica (China, Japón, Corea, Europa, América, India, entre otras). Los

contenidos abarcan géneros diversos como drama, comedia, suspense, ciencia ficción, programas de variedades o entrevistas, lo que incrementa la diversidad expresiva, étnica y contextual de las muestras. En términos de duración, un 11 % de los clips dura menos de 2 segundos, el 68 % entre 2 y 5 segundos, y el 21 % restante supera los 5 segundos, ofreciendo así cobertura tanto de micro-expresiones como de reacciones emocionales sostenidas.

Uno de los principales aportes de MAFW es su enfoque multimodal: cada clip incluye no solo la secuencia de imágenes en vídeo, sino también una pista de audio sincronizada y descripciones textuales de la escena emocional, disponibles en chino e inglés [42]. Este diseño permite el análisis integrado de señales visuales, auditivas y semánticas, ampliando las posibilidades para sistemas de aprendizaje profundo multimodal.

El sistema de anotación de MAFW es exhaustivo. Cada clip fue evaluado por 11 expertos en codificación facial, entrenados específicamente en micro-expresiones y unidades de acción facial (AU) [42]. Las emociones anotadas abarcan 11 categorías discretas: ira, asco, miedo, felicidad, tristeza, sorpresa, desprecio, ansiedad, impotencia, decepción y neutralidad, extendiendo así el repertorio más allá de las seis emociones básicas de Ekman. En este trabajo, no obstante, se empleará un subconjunto reducido con las siete emociones básicas mediante un filtrado previo.

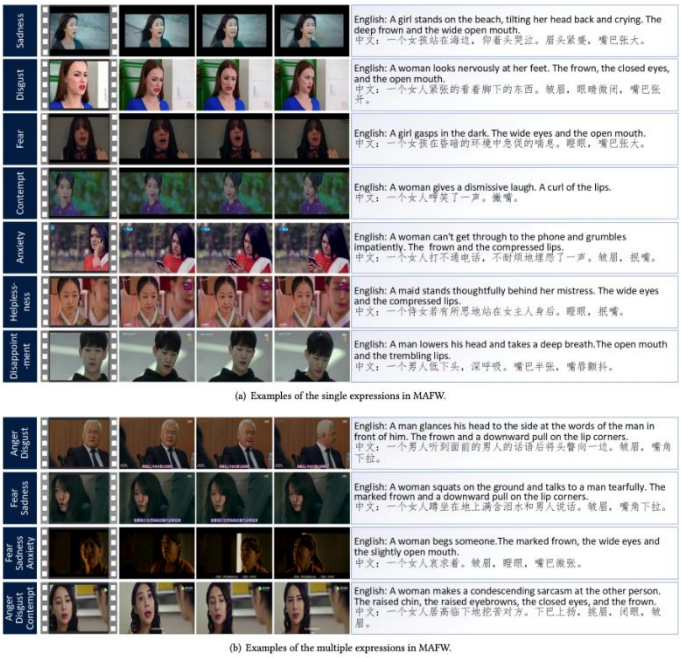

Figura 3.2. Ejemplo de expresiones únicas y múltiples en MAFW
Fuente [65]

El dataset ofrece tres tipos de anotaciones por clip: etiqueta de expresión única (una sola emoción dominante), etiqueta de expresión compuesta (combinaciones de múltiples emociones) y descripción textual contextualizada de la situación emocional (véase la figura 3.2).

En total, se identificaron 9.172 clips con emociones únicas y 4.058 clips con etiquetas múltiples, categorizados en 32 clases compuestas (por ejemplo, “ira + asco”, “miedo + tristeza”, etc.). Todos los clips con expresiones múltiples conservan una etiqueta dominante que permite su reutilización como expresión simple cuando el análisis así lo requiera.

La distribución emocional es notoriamente desbalanceada, reflejando la ocurrencia natural de emociones en medios audiovisuales: “tristeza” representa un 16 % de las muestras con etiqueta única, “ira” un 15,2 %, “felicidad” un 13,6 % y “neutralidad” un 12,4 %. En contraposición, emociones como “desprecio”, “impotencia” o “decepción” aparecen en menos del 3 % de los casos. Este desbalance dota al conjunto de un valor adicional al forzar a los modelos a enfrentarse a clases minoritarias o ambiguas [42].

Desde una perspectiva metodológica, MAFW ofrece protocolos oficiales de validación, incluyendo tareas de clasificación simple, clasificación compuesta y validación cruzada entre etiquetas dominantes y distribuciones emocionales. Esta versatilidad lo ha convertido en un *benchmark* emergente para la evaluación de arquitecturas como Transformers temporales, modelos con atención, enfoques multimodales, y métodos de aprendizaje supervisado robustos frente a etiquetado ruidoso o subjetivo [42].

En definitiva, MAFW es una de las bases de datos más completas y exigentes del ámbito de la computación afectiva contemporánea. Su riqueza estructural, cobertura emocional amplia, etiquetado experto y enfoque multimodal la posicionan como un recurso clave para validar sistemas que aspiren a funcionar en entornos realistas, donde las emociones no son discretas ni fácilmente separables, sino complejas, ambiguas y culturalmente moduladas.

3.2. Selección de los modelos

A continuación, se enseñan los modelos seleccionados para la elaboración del trabajo así como sus arquitecturas.

3.2.1. CLIP (ViT/B-32).

El modelo *Vision Transformer* ViT-B/32 representa una de las aproximaciones más innovadoras en el campo de la visión por computador, al trasladar directamente los principios de los *Transformers* —originalmente desarrollados para el procesamiento de lenguaje natural— al análisis de imágenes. A diferencia de las redes convolucionales tradicionales, que emplean filtros locales y estructuras jerárquicas para capturar características espaciales, ViT-B/32 divide la imagen en parches regulares de 32×32 píxeles. Estos parches se aplanan y se proyectan a un espacio de representación mediante capas lineales, siendo posteriormente procesados por capas de auto-atención multi-cabeza, lo que permite capturar dependencias globales entre diferentes regiones de la imagen desde las primeras etapas del procesamiento [29].

Esta arquitectura evita los *inductive biases* característicos de las CNN —como la invariancia traslacional— y, aunque ello conlleva una mayor dependencia de grandes volúmenes de datos para el entrenamiento, permite modelar interacciones espaciales complejas, lo cual resulta esencial en tareas como el reconocimiento emocional facial, donde las señales relevantes pueden estar dispersas por todo el rostro. A nivel interno, ViT-B/32 utiliza bloques repetidos de atención, normalización y redes *feedforward*, integrados mediante conexiones residuales, lo que contribuye a una representación robusta y flexible de la información visual [29].

En el marco de este trabajo, ViT-B/32 se ha integrado dentro del modelo multimodal **CLIP** (Contrastive Language–Image Pretraining), que entrena simultáneamente un codificador visual y uno textual en un espacio semántico compartido (véase la Figura 3.3). Durante la inferencia, se emplean *embeddings* textuales predefinidos —como "*anger*", "*happiness*" o "*sadness*"— que se comparan con los *embeddings* generados por ViT-B/32 a partir de secuencias de vídeo facial. El sistema analiza clips de vídeo de 8 fotogramas, extraídos dinámicamente desde *webcam* o archivos, y estima la emoción dominante mediante una inferencia conjunta de similitud entre texto e imagen.

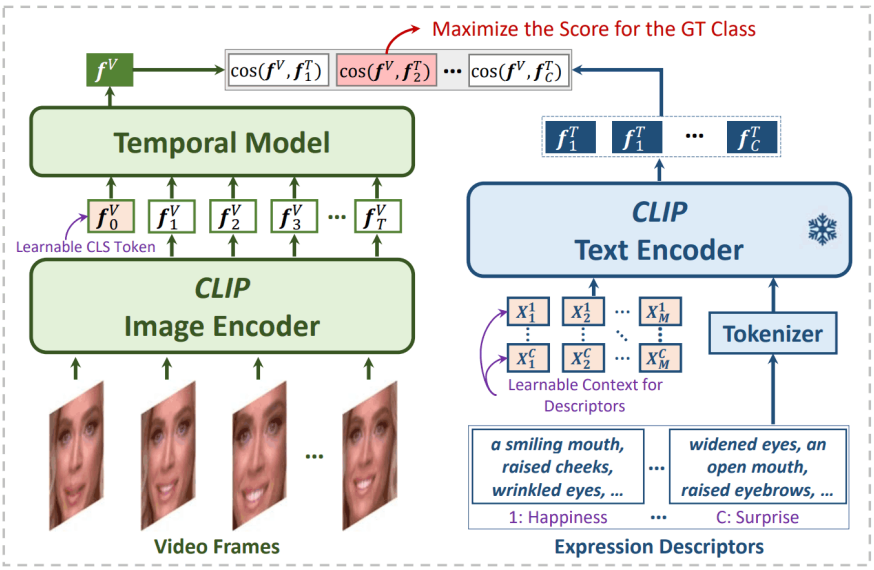

Figura 3.3. Esquema del modelo CLIP.
Fuente [66]

La base de datos empleada para el entrenamiento y evaluación ha sido exclusivamente DFEW, una colección de vídeos etiquetados con expresiones emocionales capturados en condiciones realistas, no controladas. Esta elección permite validar la robustez del modelo frente a variaciones en iluminación, pose, movimiento y calidad del vídeo, factores que suelen dificultar la detección emocional en aplicaciones prácticas[41].

El presente trabajo ha adaptado e implementado una versión del sistema propuesto por Zhao y Patras [43], en el que se emplea ViT-B/32 dentro del marco CLIP para el reconocimiento dinámico de expresiones faciales. El sistema desarrollado

permite analizar en tiempo real la evolución emocional del paciente a partir de vídeo facial, integrando detección automática de rostro, preprocesamiento, segmentación temporal y clasificación mediante atención visual-lingüística [43]. Gracias a la capacidad de ViT para capturar micro-expresiones y variaciones sutiles del rostro, el sistema logra una detección emocional precisa y eficiente, adecuada para tareas de monitorización continua.

3.2.2. CNN+LSTM

El modelo híbrido CNN+LSTM ha emergido como una de las arquitecturas más efectivas para la detección de emociones faciales en vídeo, gracias a su capacidad de combinar la representación espacial de las CNNs con el modelado secuencial de las RNNs tipo LSTM. Esta integración es especialmente útil en tareas donde las emociones no solo están presentes en imágenes estáticas, sino también en la evolución dinámica de los gestos faciales a lo largo del tiempo.

En esta arquitectura, cada fotograma del vídeo es procesado por una red convolucional profunda, en este caso una ResNet50, que actúa como extractor de características. Las CNNs son muy eficaces para identificar patrones espaciales complejos, como movimientos de los labios, elevación de cejas o fruncimiento del entrecejo, que corresponden a emociones específicas. ResNet50, en particular, introduce conexiones residuales que permiten un entrenamiento más eficiente y profundo sin problemas de desvanecimiento del gradiente, lo cual resulta clave para capturar sutilezas faciales [44].

El resultado de esta etapa es un vector de características por fotograma, el cual se inserta en una secuencia temporal utilizando una ventana deslizante. Esta secuencia se alimenta a un módulo LSTM, que se encarga de modelar las relaciones a largo plazo entre los distintos fotogramas. A través de mecanismos de memoria interna y compuertas de entrada, olvido y salida, las LSTM son capaces de retener información relevante de estados emocionales pasados, facilitando así la detección de emociones transitorias, mixtas o de baja intensidad que podrían ser ignoradas en una imagen individual [25].

La salida de las LSTM, que encapsula tanto la información espacial como su evolución temporal, se pasa a una capa completamente conectada seguida de una función *softmax* que calcula la probabilidad de pertenencia a cada clase emocional. El sistema permite, así, una clasificación robusta en condiciones reales como cambios de iluminación, movimiento del sujeto, o variación de ángulo de la cámara. Además, la arquitectura se adapta bien a tareas en tiempo real, lo que la convierte en una opción práctica para aplicaciones clínicas, robótica social o interfaces afectivas.

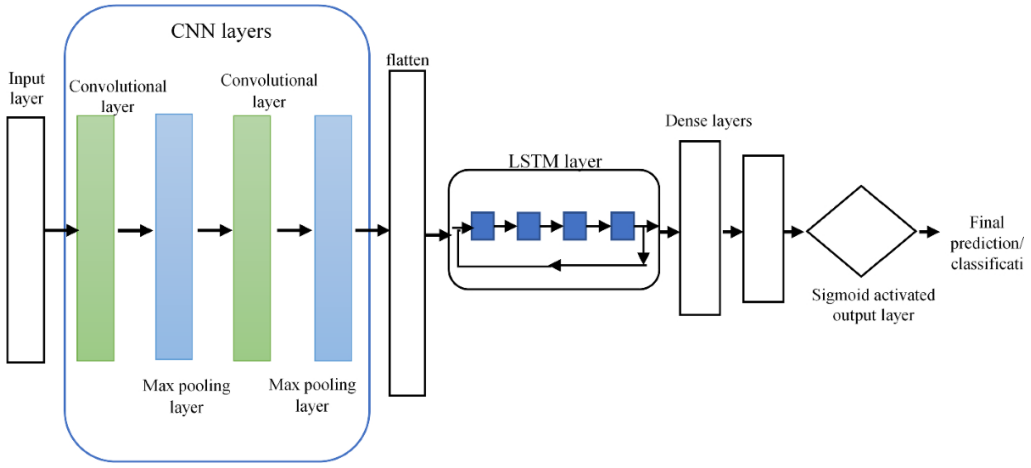

Figura 3.4: Esquema del modelo híbrido CNN+LSTM.
Fuente [67]

Además de los modelos CNN+LSTM (véase la Figura 3.4), otros enfoques utilizados en el análisis emocional dinámico incluyen redes convolucionales tridimensionales (3D CNN), como C3D o I3D, que operan directamente sobre bloques espacio-temporales del vídeo. También destacan los previamente mencionados *Vision Transformers* y modelos híbridos como *TimeSformer*, que modelan explícitamente la atención en las dimensiones espacial y temporal, aunque requieren una gran capacidad computacional y grandes volúmenes de datos para alcanzar su rendimiento óptimo [45].

En el presente trabajo, se he desarrollado una implementación propia de un modelo CNN+LSTM basado en vídeo facial en tiempo real. La arquitectura utiliza ResNet50 como extractor convolucional de características y un doble LSTM para modelado temporal. Esta implementación se inspira en el proyecto de Elena Ryumina, *Static and Dynamic Facial Emotion Recognition Using the Emo-AffectNet Model*, en el que se entrenó una red CNN+LSTM sobre el conjunto de datos AffectNet [46]. A partir

de su código y diseño, he adaptado y optimizado el sistema para el análisis del estado emocional del paciente, integrando detección facial con *MediaPipe* y operando sobre secuencias de vídeo reales.

3.3. Métricas

Para evaluar el rendimiento del sistema de detección emocional en vídeo facial, se han empleado diversas métricas propias de la clasificación supervisada multiclase, así como otras específicas para tareas dinámicas en tiempo real. Estas métricas permiten valorar tanto la eficacia global del sistema como su comportamiento frente a clases desequilibradas, aspecto crucial en problemas afectivos donde ciertas emociones son inherentemente minoritarias.

3.3.1. Accuracy (métrica individual por clase).

La *accuracy* es una métrica global, que también puede calcularse a nivel de clase individual, ampliamente utilizada en clasificación supervisada. Representa la proporción de predicciones correctas respecto al total de instancias evaluadas y se define como:

$$Accuracy = \frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i}$$

Donde TP (true positives) son los verdaderos positivos, TN los verdaderos negativos, FP los falsos positivos y FN los falsos negativos [47].

En este caso, se considera la proporción de veces que el modelo acierta prediciendo correctamente o descartando una clase i, respecto al total de casos que involucraban dicha clase. Esta interpretación permite observar de forma detallada si el sistema tiene dificultades particulares para reconocer o distinguir ciertas emociones como “asco” o “sorpresa”. No obstante, al incluir tanto verdaderos negativos como positivos en su cálculo, esta métrica puede verse inflada en clases minoritarias con muchos negativos, por lo que es importante interpretarla junto a otras métricas sensibles al desbalance como el *recall* o la precisión macro.

3.3.2. Precisión (métrica individual por clase)

La precisión (o *positive predictive value*) es una métrica individual por clase que mide la proporción de verdaderos positivos sobre el total de instancias predichas como pertenecientes a esa clase

$$Precision = \frac{TP}{TP + FP}$$

Esta métrica es especialmente útil en tareas como la clasificación emocional, donde se desea evitar falsos positivos en clases como “miedo” o “tristeza”, debido a su sensibilidad social o clínica [48].

3.3.3. Recall (Sensibilidad) (métrica individual por clase)

El *recall*, también denominado sensibilidad (o *true positive rate)*, mide la capacidad del modelo para detectar todas las instancias reales de una clase concreta. Se define como:

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

En el reconocimiento emocional, un alto *recall* indica que la mayoría de las emociones reales de una categoría (por ejemplo, tristeza) han sido correctamente detectadas. Esta métrica es crucial cuando el coste de omitir una clase es elevado, como podría ocurrir con emociones asociadas a estados psicológicos delicados. El *recall* forma parte del balance que establece el *F1-score* junto con la precisión [49].

3.3.4. F1-score (métrica individual por clase)

El *F1-score* es la media armónica entre precisión y *recall*. Ofrece una medida equilibrada del rendimiento, particularmente útil en escenarios con clases poco representadas. Se calcula como:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Este indicador proporciona una evaluación más equilibrada cuando existe un desbalance entre las clases. Es especialmente útil en reconocimiento de emociones, donde ciertas emociones (como la repulsión o el desprecio) están poco representadas en los *datasets* [50]. Un *F1-score* alto indica que el modelo tiene tanto una alta cobertura de instancias reales como una baja tasa de falsos positivos.

3.3.5. Weighted Average Recall (WAR) (métrica global)

La WAR (recuerdo promedio ponderado) es una métrica global que evalúa el rendimiento global del sistema ponderando el *recall* de cada clase según su frecuencia real en el conjunto de datos:

$$WAR = \frac{\sum_{i=1}^{N} TP_i}{\sum_{i=1}^{N}(TP_i + FN_i)}$$

La WAR coincide con la *accuracy* en términos matemáticos, pero interpretada como promedio ponderado del rendimiento por clase. Es especialmente útil para medir el comportamiento del sistema en condiciones reales donde algunas emociones dominan en frecuencia, como la neutralidad o la alegría [51].

3.3.6. Unweighted Average Recall (UAR) (métrica global)

La UAR, formalmente, se define como el promedio aritmético del *recall* obtenido en cada clase individual, sin tener en cuenta su frecuencia relativa

$$UAR = \frac{1}{N}\sum_{i=1}^{N}\frac{TP_i}{TP_i + FN_i}$$

Donde: N es el número total de clases, TPi Representa los verdaderos positivos para la clase y FNi representa los falsos negativos de la clase

Esta métrica garantiza una evaluación equitativa entre clases, penalizando al modelo si falla en clases minoritarias como “asco” o “desprecio”, que suelen estar infrarrepresentadas en las bases de datos emocionales. Por ello, es especialmente

adecuada en el campo del reconocimiento facial emocional, y ha sido adoptada como métrica estándar en múltiples benchmarks afectivos [48], [52].

La diferencia entre UAR y WAR radica en la forma de promediar: mientras que la WAR concede a cada clase un peso proporcional a su frecuencia, la UAR calcula el *recall* de cada emoción y luego hace la media sin tener en cuenta cuántos ejemplos hay de cada una. Así, la WAR puede ocultar un pobre desempeño en categorías minoritarias si hay muchas muestras de clases dominantes, pero ofrece una visión alineada con la distribución real de datos; en cambio, la UAR garantiza que cada emoción aporte igual al resultado final, penalizando duramente los fallos en emociones poco representadas y proporcionando una evaluación más equilibrada en contextos de desbalance.

3.3.7. Precisión macro (métrica global)

La macro precisión calcula la media aritmética de la precisión individual por clase:

$$Precision_{macro} = \frac{1}{N}\sum_{i=1}^{N}\frac{TP_i}{TP_i + FP_i}$$

Al no ponderar por la cantidad de muestras, permite evaluar si el modelo mantiene un rendimiento homogéneo entre clases frecuentes y minoritarias, siendo especialmente útil en contextos desbalanceados como los que presentan las bases de datos afectivas [50].

3.3.8. F1-score macro (métrica global)

El *F1-score* macro representa la media aritmética de los *F1-scores* obtenidos para cada clase, evaluando así el rendimiento global del sistema sin depender de la distribución de las muestras:

$$F1_{macro} = \frac{1}{N}\sum_{i=1}^{N} 2 \times \frac{Precision_i \times Recall_i}{Precision_i + Recall_i}$$

Esta métrica se considera una de las más exigentes, ya que penaliza errores en cualquier clase con igual severidad. Es particularmente valiosa para evaluar sistemas en entornos clínicos, educativos o sociales donde todas las emociones deben ser consideradas igualmente relevantes [51].

3.3.9. Matriz de confusión

La matriz de confusión es una herramienta esencial para analizar el rendimiento de un clasificador multi-categoría. Es una tabla que resume las predicciones correctas e incorrectas hechas por el modelo, diferenciadas por clase. Cada fila representa las instancias reales de una clase, mientras que cada columna representa las predicciones realizadas:

$$Matriz\ de\ confusión = \begin{bmatrix} TP_1 & FP_{1,2} & \dots \\ FN_{2,1} & TP_2 & \dots \\ \vdots & \vdots & \ddots \end{bmatrix}$$

Su análisis facilita identificar emociones que el modelo confunde habitualmente entre sí, como “miedo” y “sorpresa”, ayudando a refinar tanto el diseño del modelo como la calidad de los datos [49].

3.3.10. FPS

Los FPS (*frames per second*) miden el número de fotogramas que el sistema puede procesar por segundo en tiempo real. En aplicaciones de vídeo, como la detección emocional continua, el FPS es crítico para determinar si el sistema puede

operar en tiempo real o no. Generalmente se considera que un rendimiento aceptable en reconocimiento de emociones comienza a partir de los 5-15 FPS [52].

Esta métrica depende no solo del modelo, sino también de la eficiencia del preprocesamiento, la lectura del vídeo y el hardware utilizado (GPU, CPU, RAM).

CAPITULO 4

RESULTADOS Y DISCUSIÓN

En este cuarto capítulo del trabajo, se presentan y analizan los resultados obtenidos tras la evaluación de los modelos propuestos para la detección de emociones faciales en vídeo.

En particular, se aborda el rendimiento del sistema basado en el modelo Vision Transformer ViT-B/32, entrenado y testeado un conjunto de datos de DEFW y posteriormente validado sobre la base de datos MAFW. Se calcularán y justificarán las métricas implementadas. Finalmente, se incluye el análisis del comportamiento del sistema en condiciones de ejecución en tiempo real, tanto mediante webcam como a través de archivos de vídeo, midiendo el rendimiento operativo en términos de frames por segundo (FPS).

Por otro lado, se expone también el trabajo realizado con una arquitectura híbrida CNN+LSTM, basada en el proyecto original de Elena Ryumina, centrado en el reconocimiento emocional estático y dinámico con el modelo Emo-AffectNet [46]. Aunque debido a limitaciones técnicas no se ha podido integrar este segundo modelo en un entorno de evaluación cuantitativa sobre datasets anotados, sí se ha verificado su correcto funcionamiento sobre vídeo en tiempo real mediante detección facial con *MediaPipe*, lo que ofrece una base sólida para futuras comparativas.

4.1. Evaluación del modelo (ViT/B-32). Métricas

El archivo CLIP_metricas.py constituye el script principal empleado para la evaluación del modelo Vision Transformer ViT-B/32. Su propósito es calcular métricas de clasificación multiclase a partir de las predicciones generadas por el modelo sobre distintos conjuntos de datos.

Este script implementa una arquitectura modular que integra diversos componentes definidos en scripts auxiliares: video_dataloader.py, encargado de la carga de datos; video_transform.py, responsable de las transformaciones de entrada; y *Generate_Model.py*, que define la arquitectura del modelo.

El flujo de trabajo se basa en cargar un modelo previamente entrenado y aplicarlo de manera secuencial sobre los clips de los conjuntos de entrenamiento (*train*), prueba (*test*) y validación (*validation*). Para cada clip, se genera una predicción que se compara con su etiqueta real, construyendo así matrices de confusión. Estas matrices permiten derivar métricas tanto globales como específicas por clase emocional.

El modelo fue entrenado y evaluado sobre la base DFEW, empleando un subconjunto definido por los autores del dataset: 9.356 clips para entrenamiento y 2.341 para prueba. Esta configuración permite una comparación directa con otros enfoques del estado del arte.

Para evaluar la capacidad de generalización del modelo, se llevó a cabo una validación cruzada externa sobre la base MAFW, que incluye 7.576 clips. Esta base, caracterizada por expresiones espontáneas y condiciones no controladas, fue preprocesada mediante el script generate_mafw_annotation.py, el cual reorganiza las anotaciones en un formato compatible con las siete emociones básicas reconocidas por el modelo.

La evaluación combina análisis cuantitativos y cualitativos. En el plano cuantitativo, se emplean matrices de confusión, a partir de las cuales se calculan métricas estándar como accuracy, precisión, recall y F1-score. Además, se incorporan métricas específicas como WAR y UAR, debido al desbalance típico en los conjuntos de datos emocionales.

Las métricas se calcularon para tres fases: entrenamiento sobre DFEW, prueba sobre DFEW y validación cruzada sobre MAFW. Este enfoque permite analizar el ajuste del modelo al conjunto de entrenamiento y su generalización ante datos nuevos.

4.1.1. Fase 1: Entrenamiento

Durante esta fase se procesaron 9.356 clips.

La matriz de confusión absoluta (véase la tabla 4.1) revela que la clase “felicidad” presenta la mayor cantidad de aciertos con 1.724 predicciones correctas, seguida de la emoción “neutral” con 1.581 clips y de “ira” con 1.088. Las clases con menor número de aciertos absolutos son “repulsión” con 46 aciertos y “miedo” con 287 aciertos, evidenciando una clara dificultad del modelo para distinguir estas emociones, posiblemente por su menor representación o ambigüedad visual.

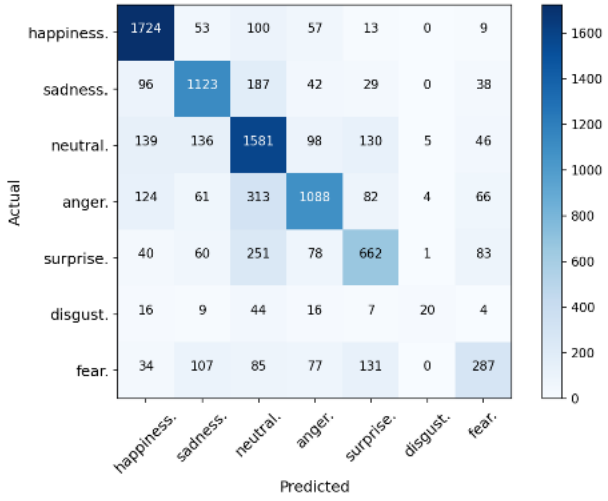

Tabla 4.1. Matriz de confusión absoluta - Resultados (Entrenamiento)
Fuente [Tabla de elaboración propia]

Se registran confusiones frecuentes entre ciertas clases emocionales, lo que pone de manifiesto las dificultades del modelo para distinguir emociones con expresiones faciales similares o solapadas. En particular, se registra una elevada tasa de errores entre “tristeza” y “neutral”, con 187 clasificaciones incorrectas; asimismo, “ira” es confundida con “neutral” en 313 ocasiones. También destaca la confusión entre “miedo” y “sorpresa”, con 131 casos.

Para facilitar una mejor comparación entre clases con distinto número de ejemplos, se elabora la matriz normalizada que expresa los valores en porcentaje respecto al total de muestras por clase verdadera (véase la tabla 4.2).

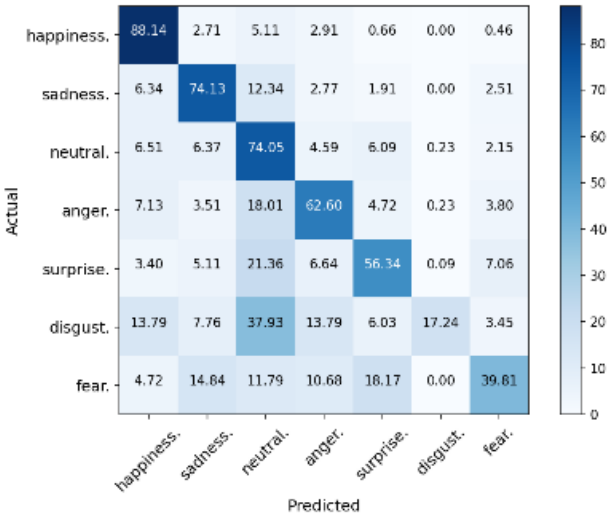

Tabla 4.2. Matriz de confusión normalizada - Resultados (Entrenamiento)
Fuente [Tabla de elaboración propia]

De forma lógica, los resultados porcentuales obtenidos a partir de la matriz de confusión normalizada guardan una correspondencia directa con los valores absolutos previamente observados. Esto se debe a que la matriz normalizada representa la distribución relativa de aciertos y errores por clase, calculada a partir de la matriz absoluta. De esta forma, se permite observar el rendimiento relativo para cada clase.

Pasando los valores a porcentaje, se observa que el modelo acierta correctamente en el 88.14 % de los casos para la clase “felicidad”, lo que demuestra una alta capacidad discriminativa para esta emoción. “Tristeza” y “neutral” alcanzan tasas de acierto del 74.13 % y 74.05 %, respectivamente. Sin embargo, emociones como “repulsión” (17.24 %) y “miedo” (39.81 %) presentan rendimientos considerablemente más bajos, revelando una menor precisión del modelo para identificar estas emociones.

Por otra parte, como se evidenciaba también en la matriz absoluta, se observan errores sistemáticos relacionados con confusiones específicas. Por ejemplo, “repulsión” es confundido con “neutral” en un 37.93 % de los casos, y “miedo” se predice incorrectamente como “tristeza” en un 14.84 % de las ocasiones, o como “sorpresa” en un 18.17 % de los casos.

Las métricas individuales por clase (véase la tabla 4.3) reflejan este comportamiento:

Train DEFW	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Felicidad	92.72	79.34	88.14	83.51
Tristeza	91.26	72.50	74.13	73.30
Neutral	83.60	61.73	74.05	67.33
Ira	89.12	74.73	62.60	68.13
Sorpresa	90.33	62.81	56.34	59.40
Repulsión	98.87	66.67	17.24	27.40
Miedo	92.73	53.85	39.81	45.77

Tabla 4.3. Métricas individuales multiclase - Resultados (Entrenamiento)
Fuente [Tabla de elaboración propia]

Las métricas individuales por clase (véase tabla 7.3) cuantifican el rendimiento del modelo en términos de *accuracy*, *precisión*, *recall* y *F1-score*. “Felicidad” destaca con un F1-score del 83.51 %, seguida por “tristeza” (73.30 %) y “ira” (68.13 %). La clase “neutral” alcanzó un F1-score de 67.33 %, penalizada por una precisión algo baja (61.73 %), pese a su elevado *recall* (74.05 %).

En las emociones con menor rendimiento, como “repulsión” (F1-score: 27.40 %) y “miedo” (45.77 %), los valores reflejan una considerable dificultad del modelo para identificarlas correctamente. En particular, aunque “repulsión” presenta una precisión del 66.67 %, su bajo *recall* (17.24 %) compromete el equilibrio de la predicción, lo cual se traduce en un F1-score muy bajo.

En conjunto, estos resultados reflejan que el modelo logra una buena detección en emociones predominantes o con rasgos faciales más marcados, mientras que encuentra mayores desafíos en aquellas emociones menos representadas o más ambiguas, como también se evidencia en las matrices de confusión.

Para finalizar de detallar los resultados del entrenamiento, se disponen los valores obtenidos para las métricas globales (véase la tabla 7.4).

Métrica	Valor (%)
WAR	69.31
UAR	58.90
Precisión macro	67.37
F1 macro	60.69

Tabla 4.4. Métricas globales por muestra - Resultados (Entrenamiento)
Fuente [Tabla de elaboración propia]

Estos valores reflejan que el modelo alcanza un rendimiento global representado por un WAR del 69.31 %, lo que indica una buena capacidad para clasificar correctamente teniendo en cuenta la distribución de clases. No obstante, el valor de UAR, situado en 58.90 %, revela una disminución de rendimiento al considerar cada clase por igual, lo que sugiere desequilibrios en la predicción entre emociones. La precisión macro alcanzada fue del 67.37 %, mostrando que, en promedio, el modelo identifica correctamente las instancias positivas de cada clase. Por último, el *F1-score* macro se sitúa en 60.69 %, lo cual refleja un rendimiento moderado en la compensación entre precisión y exhaustividad al considerar todas las emociones de forma equitativa. Estos resultados en conjunto reflejan un desempeño sólido en clases dominantes, pero con margen de mejora en aquellas más difíciles o minoritarias.

4.1.2. Fase 2: Test

En la fase de prueba, el modelo fue evaluado sobre un total de 2.341 clips del conjunto DFEW:

Mediante las matrices de confusión, se ha evidenciado un rendimiento globalmente sólido pero desigual entre las distintas clases emocionales. Las matrices de confusión, tanto absoluta como normalizada (véase las tabla 4.5 y 4.6), indican que “felicidad” fue la emoción mejor clasificada, con 457 aciertos (93.46 %), seguida de “neutral” (384 aciertos, 71.91 %) e “ira” (290 aciertos, 66.67 %). Estas emociones, al presentar expresiones faciales más marcadas y mayor representación en los datos, fueron reconocidas con elevada fiabilidad.

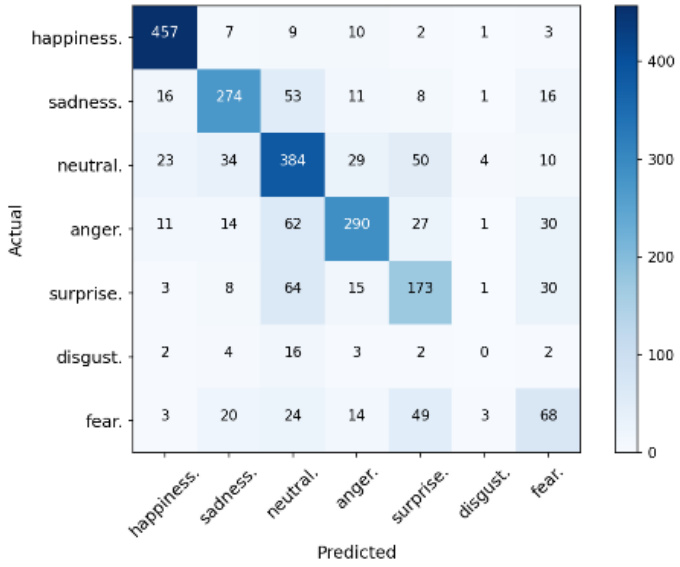

Tabla 4.6. Matriz de confusión absoluta - Resultados (Prueba)
Fuente [Tabla de elaboración propia]

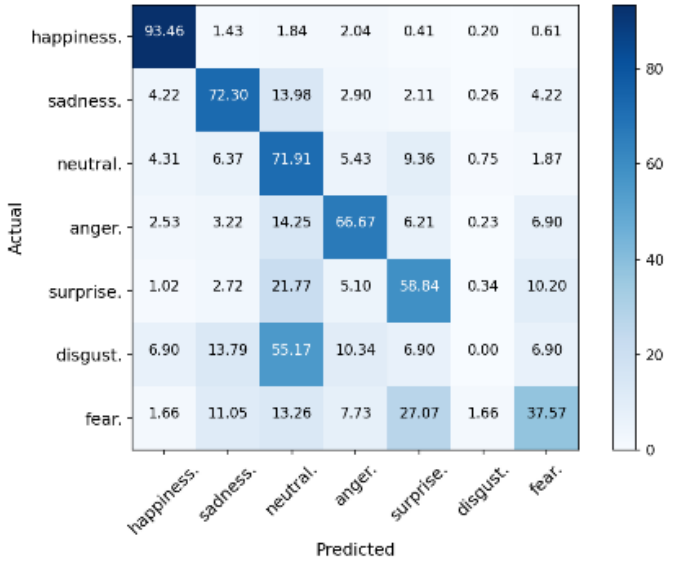

Tabla 4.5. Matriz de confusión normalizada - Resultados (Prueba)
Fuente [Tabla de elaboración propia]

En cambio, el reconocimiento de emociones más sutiles o con menor representación fue notablemente más débil. “Sorpresa” registró 173 aciertos (58.84 %) y “miedo” 68 (37.57 %), reflejando la dificultad del modelo para discriminar clases ambiguas o poco frecuentes. Especialmente destacable es el caso de “repulsión”, para

la cual no se obtuvo ningún acierto, evidenciando la completa incapacidad del modelo para reconocer esta emoción en el conjunto de prueba.

El análisis de errores revela patrones sistemáticos, como la tendencia del modelo a clasificar de forma errónea instancias de “sorpresa”, “ira” y “tristeza” como “neutral”, lo cual sugiere una sobre-clasificación hacia esta categoría intermedia. Asimismo, se detectaron errores recurrentes entre “miedo” y “sorpresa” (27.07 %) y entre “repulsión” y “neutral” (13.79 %), que refuerzan la hipótesis de solapamiento expresivo entre estas emociones.

A continuación, se disponen los resultados de las métricas individuales por clase (véase la tabla 4.7).

Test DEFW	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Felicidad	96.16	88.74	93.46	91.04
Tristeza	91.80	75.90	72.30	74.05
Neutral	83.85	62.75	71.91	67.02
Ira	90.30	77.96	66.67	71.87
Sorpresa	88.94	55.63	58.84	57.19
Repulsión	98.29	0.00	0.00	0.00
Miedo	91.29	42.77	37.57	40.00

Tabla 4.7. Métricas individuales multiclase - Resultados (Prueba)
Fuente [Tabla de elaboración propia]

La emoción “felicidad” obtuvo un rendimiento sobresaliente, con una precisión del 88.74 %, recall del 93.46 % y un F1-score de 91.04 %. “Tristeza” (74.05 %) e “ira” (71.87 %) también registraron buenos resultados. En el caso de “neutral”, se observó un recall elevado (71.91 %) pero una precisión más moderada (62.75 %), lo que sugiere que otras clases fueron clasificadas como “neutral” en situaciones ambiguas.

Por el contrario, las emociones “sorpresa” (F1-score: 57.19 %) y “miedo” (40.00 %) mostraron limitaciones significativas, mientras que “repulsión” no fue reconocida en absoluto teniendo un rendimiento nulo. Estos resultados destacan la

necesidad de mejorar la capacidad del sistema para identificar emociones menos frecuentes o con patrones faciales menos distintivos.

Para finalizar la fase de test, se observan las métricas globales obtenidas por el modelo (véase la tabla 4.8), que reflejan un rendimiento general aceptable. El WAR alcanzó un 70,31 %, indicando un buen rendimiento general ponderado por frecuencia, mientras que el UAR se situó en un 57,25 %, lo que evidencia una caída al considerar todas las clases por igual.

Métrica	Valor (%)
WAR	70.31
UAR	57.25
Precisión macro	57.68
F1 macro	57.31

Tabla 4.8. Métricas globales por muestra - Resultados (Prueba)
Fuente [Tabla de elaboración propia]

El modelo tiende a funcionar mejor con clases mayoritarias y a fallar en clases menos representadas o más difíciles de discriminar. Este patrón se confirma con la precisión macro del 57,68% y F1 macro del 57,31%, que se ven afectadas por las pobres predicciones en *repulsión y miedo*.

4.1.3. Fase 3: Validación

Durante la fase de validación cruzada utilizando la base de datos MAFW (7.576 clips), el modelo ViT-B/32 previamente entrenado con DFEW ha sido evaluado sobre un conjunto de muestras externas con características más espontáneas, naturales y variadas. Esta validación ha sido esencial para comprobar la capacidad de generalización del sistema fuera del dominio original de entrenamiento y está relacionada directamente con el objetivo del trabajo. Estos, reflejados tanto en las matrices de confusión como en las métricas individuales y globales, indican un

rendimiento consistente y técnicamente óptimo, teniendo en cuenta la complejidad del nuevo dominio.

En la matriz de confusión absoluta (véase tabla 4.9), se aprecia que la clase “felicidad” ha sido reconocida correctamente en 1.103 casos, seguida por “tristeza” con 1.113 y “neutral” con 823 aciertos. Se garantiza, por lo tanto, que él tiene una buena capacidad para identificar emociones con expresiones faciales bien definidas. La clase “ira” también presenta un resultado elevado, con 807 aciertos. Por el contrario, se observa una disminución en el rendimiento para emociones más ambiguas como “sorpresa” (651 aciertos), “miedo” (189) y especialmente “repulsión”, que solo ha sido correctamente clasificada en 34 ocasiones. Estas cifras, aunque ligeramente inferiores a las obtenidas en la fase de prueba con DFEW, resultan esperables debido al carácter más desafiante y no controlado de la base MAFW.

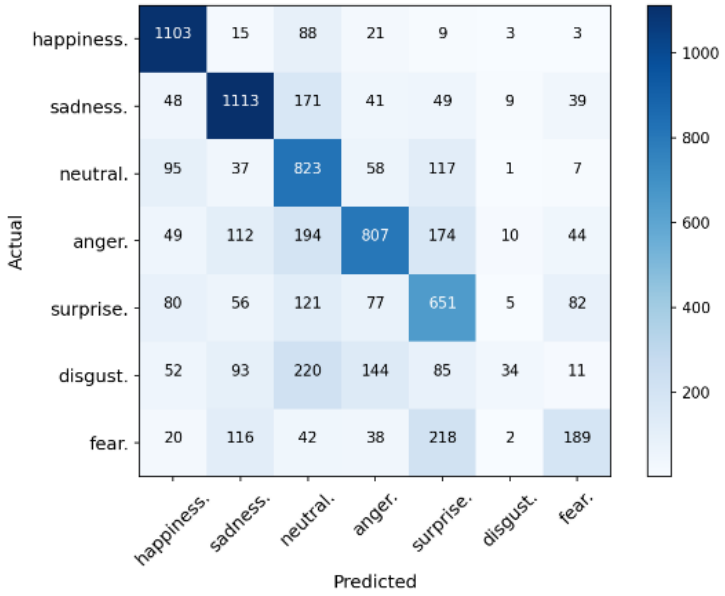

Tabla 4.9. Matriz de confusión absoluta - Resultados (Validación)
Fuente [Tabla de elaboración propia]

La matriz normalizada (véase la tabla 4.10) confirma esta tendencia. “Felicidad” mantiene un elevado nivel de acierto (88.81 %), seguida de “tristeza” (75.71 %), “neutral” (72.32 %) y “ira” (58.06 %). No obstante, la tasa de reconocimiento de “asco” cae al 5.32 % y “miedo” se sitúa en un 30.24 %, destacando nuevamente las dificultades del modelo para discriminar estas emociones. Estas cifras son coherentes con los errores sistemáticos ya observados durante el entrenamiento y test, donde se constató una

tendencia a confundir estas clases con otras de mayor representación o con rasgos faciales superpuestos.

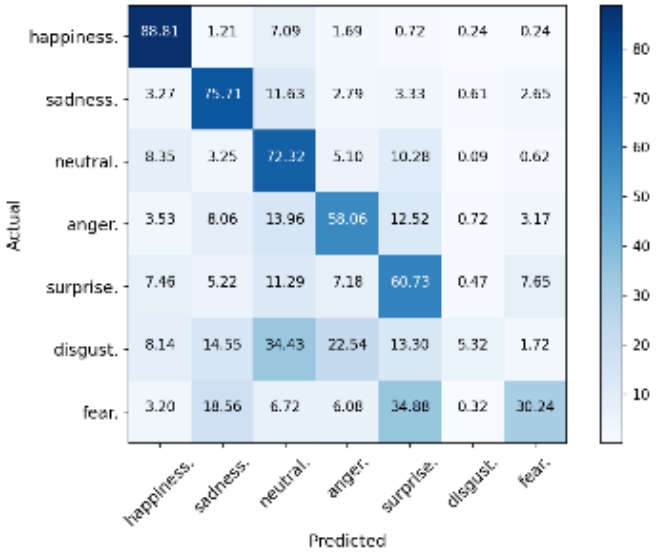

Tabla 4.10. Matriz de confusión normalizada - Resultados (Validación)
Fuente [Tabla de elaboración propia]

Las métricas individuales por clase, dispuestas arriba, respaldan este análisis (véase la tabla 4.11). El *F1-score* para “felicidad” alcanza el 82.04 %, y se mantienen valores adecuados para “tristeza” (73.90 %) y “ira” (62.66 %), mientras que “neutral” logra un F1 de 58.85 %, penalizado por una menor precisión (49.61 %). En contraste, “sorpresa” (54.82 %), “miedo” (37.80 %) y “asco” (9.67 %) reflejan un rendimiento inferior, con *recall* muy reducido y dificultades para clasificar correctamente instancias de estas categorías.

Validación MAFW	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Felicidad	93.62	76.23	88.81	82.04
Tristeza	89.63	72.18	75.71	73.90
Neutral	84.81	49.61	72.32	58.85
Ira	87.30	68.04	58.06	62.66
Sorpresa	85.84	49.96	60.73	54.82
Repulsión	91.62	53.12	5.32	9.67
Miedo	91.79	50.40	30.24	37.80

Tabla 4.11. Métricas individuales multiclase - Resultados (Validación)
Fuente [Tabla de elaboración propia]

Finalmente, las métricas globales obtenidas en esta validación (véase la tabla 4.11) indican que el modelo mantiene un comportamiento robusto, con un WAR del 62.30 % que confirma una capacidad de clasificación efectiva ponderada por clases. El UAR desciende a 55.88 %, mostrando un rendimiento más bajo en las clases minoritarias, aunque sin representar una caída crítica. La precisión macro (59.93 %) y el F1 macro (54.25 %) confirman que, pese a los retos que plantea la base MAFW, el modelo generaliza con eficacia en condiciones no vistas durante el entrenamiento. Estos resultados refuerzan la validez del sistema como herramienta para el reconocimiento emocional dinámico en entornos más realistas, destacando su solidez en emociones bien definidas y señalando al mismo tiempo áreas claras de mejora en emociones complejas o poco representadas.

Métrica	Valor (%)
WAR	62.30
UAR	55.88
Precisión macro	59.93
F1 macro	54.25

Tabla 4.12. Métricas globales por muestra - Resultados (Validación)
Fuente [Tabla de elaboración propia]

4.2. Evaluación del modelo (ViT/B-32). Análisis de video en tiempo real y vídeos

Este script implementa un sistema integral para el análisis dinámico de emociones faciales a partir de secuencias de vídeo, ya sean capturadas en tiempo real mediante webcam o procedentes de archivos previamente grabados. El sistema se basa en una arquitectura ViT-B/32 integrada con el modelo CLIP, que permite la clasificación multimodal a partir de información textual y visual, complementada con un mecanismo de aprendizaje de *prompts*. Inicialmente, el sistema permite configurar múltiples parámetros mediante línea de comandos: selección del dataset (DFEW, MAFW), hiper-parámetros de entrenamiento, número de contextos semánticos, longitud del clip o modo de entrada (webcam, archivo o ambos). Posteriormente, se cargan las clases emocionales y sus descripciones asociadas, que permiten generar *embeddings*

textuales compatibles con el espacio de representación del modelo CLIP. El modelo ViT-B/32 se carga desde un *checkpoint* entrenado y se envuelve en la clase personalizada *GenerateModel*, que combina visión y lenguaje para realizar inferencia multimodal. Esta arquitectura se activa en modo evaluación, lista para el análisis emocional dinámico.

El flujo de análisis se basa en un preprocesamiento por etapas: detección facial mediante *Haar Cascades*, recorte de la región facial, redimensionado a 224x224 píxeles, conversión a tensor y normalización. Los frames se almacenan en un buffer temporal que actúa como ventana deslizante. Cuando se alcanzan 8 o 16 fotogramas consecutivos (según la configuración), estos se agrupan como un clip y se introducen al modelo para su análisis conjunto. La salida es una distribución de probabilidad sobre las emociones básicas, de la cual se extrae la clase con mayor puntuación y su nivel de confianza. Esta información se proyecta en pantalla en tiempo real, junto a la tasa de procesamiento medida en FPS. El sistema incluye mecanismos de control de tiempo para mantener una frecuencia de muestreo estable (por ejemplo, 10 FPS), ajustando dinámicamente la velocidad según los recursos del sistema.

El software permite operar en tres modos: solo webcam, solo archivo de vídeo o ambos de forma simultánea mediante hilos paralelos. Esta versatilidad permite validar su funcionamiento en múltiples contextos y comparar el rendimiento entre entradas en vivo y offline. En conjunto, se trata de una herramienta robusta y modular, orientada al reconocimiento emocional dinámico con aplicaciones potenciales en robótica afectiva, interacción humano-máquina, análisis psicológico o monitorización en salud.

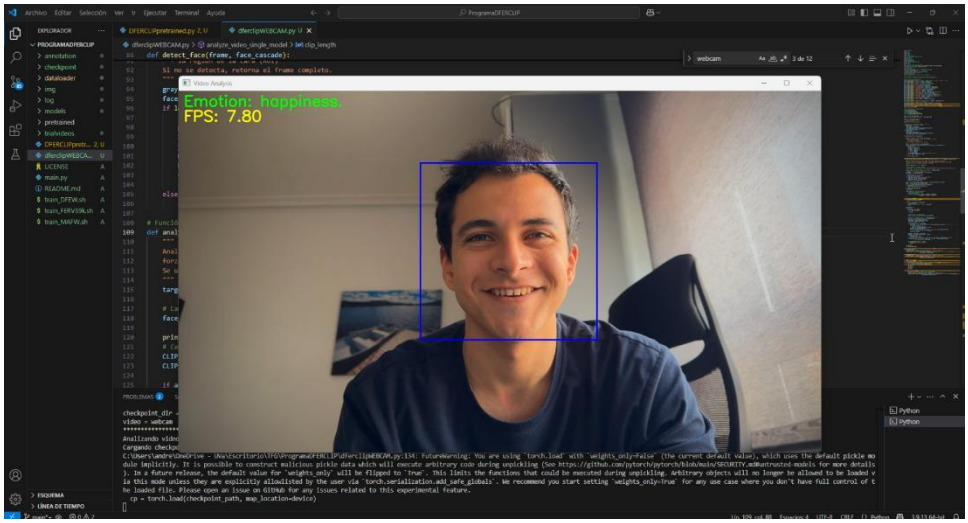

Figura 4.1. Ejemplo de la interfaz de salida del modelo ViT-B/32
Fuente [Figura de elaboración propia]

Este modelo, basado en *Vision Transformers*, analiza de forma conjunta las características espaciales y temporales del clip completo, y genera como salida una distribución de probabilidad sobre las emociones básicas predefinidas (véase la figura 4.1). La emoción con mayor probabilidad se visualiza sobre la imagen, junto con su nivel de confianza (porcentaje) y la tasa de procesado expresada en frames por segundo (FPS).

Para demostrar de forma tangible su operatividad, se ha capturado una secuencia de imágenes en la que se representan las siete emociones básicas simuladas por mí mismo, el propio autor, frente a la cámara. Esta demostración permite validar visualmente el rendimiento en tiempo real del sistema. Los resultados muestran una identificación precisa y rápida de emociones prototípicas como “felicidad”, “ira” o “neutral”, con niveles de confianza superiores al 90 % (véase la figura 4.2) . Por el contrario, emociones menos frecuentes o más ambiguas, como “miedo” o “tristeza”, presentan menor estabilidad en las predicciones, en línea con las limitaciones ya observadas en las fases cuantitativas del estudio.

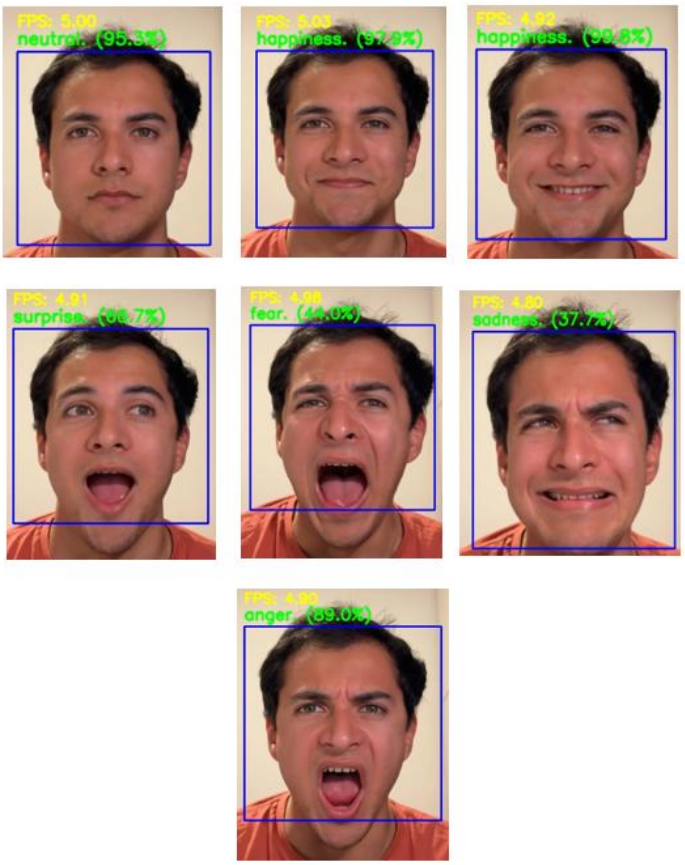

Figura 4.2. Frames a tiempo real detectando emociones en ViT. Webcam
Fuente [Figura de elaboración propia]

Durante las pruebas, el sistema ha mantenido una tasa de procesamiento media de entre 4.8 y 5.1 FPS, adecuada para aplicaciones básicas en tiempo real. Se ha observado una reducción de 2-3 FPS cuando el sistema proyecta texto sobre la imagen en cada iteración, lo cual es atribuible a la sobrecarga de renderizado continuo en la CPU empleada (gama media). Cabe esperar que con hardware más potente (GPU) o mediante optimización del pipeline gráfico, se logren mejoras significativas en este aspecto.

Adicionalmente, se ha evaluado el comportamiento del sistema sobre archivos de vídeo, aplicando el mismo flujo de detección e inferencia. En concreto, se procesó un clip de un paciente de la base DEAP, ampliamente utilizada en investigación neuro-afectiva. El sistema mostró una detección facial estable y generó inferencias emocionales consistentes a lo largo del clip, alcanzando también tasas cercanas a 4.8 FPS. Emociones como “felicidad” y “neutral” fueron reconocidas con alta fiabilidad (véase la figura 4.3), mientras que “miedo” o “repulsión” mostraron predicciones más variables, resultado coherente con las tendencias observadas en el entrenamiento.

Se muestran los resultados de predicción sobre clips correspondientes a un sujeto en condiciones controladas.

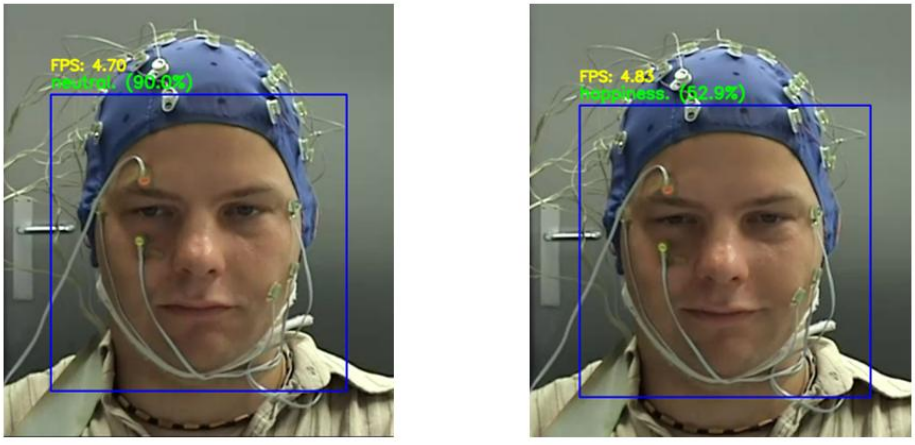

Figura 4.3. Secuencia de frames mediante archivo de vídeo en ViT
Fuente [Figura de elaboración propia]

Aunque no se aprecia una mejora sustancial frente al análisis en vivo, sí se ha registrado una ligera ganancia en estabilidad, posiblemente derivada de la ausencia de

latencia y la naturaleza más controlada del vídeo grabado, lo que reduce el ruido externo (iluminación, ángulos, resolución, etc.).

En conjunto, los resultados funcionales del sistema confirman que el modelo ViT-B/32 no solo ofrece buenos resultados en entornos de validación controlada, sino que también mantiene un comportamiento sólido y fiable en escenarios prácticos, ya sea con webcam o con vídeos offline. Esta dualidad entre robustez teórica y aplicabilidad real consolida al modelo como una herramienta prometedora para aplicaciones sensibles al componente emocional, como la interacción humano-máquina adaptativa, la robótica social o la terapia psicológica asistida. No obstante, se reitera la necesidad de abordar sus limitaciones en la clasificación de emociones menos frecuentes o sutiles. Las dificultades para identificar con precisión emociones como “miedo” o “repulsión”, especialmente en condiciones no ideales, sugieren que es necesario explorar mejoras en la arquitectura, en la estrategia de entrenamiento o en la representatividad de los datos, para incrementar su sensibilidad y cobertura en contextos reales.

4.3. Evaluación del modelo (CNN+LSTM). Análisis de video en tiempo real y vídeos

Además del modelo basado en Vision Transformers, se ha desarrollado e implementado un segundo sistema de análisis emocional dinámico utilizando una arquitectura híbrida CNN+LSTM, adaptada del proyecto original de Elena Ryumina [46]. Esta arquitectura combina la potencia de una red convolucional profunda (ResNet-50) para extraer características espaciales a partir de cada fotograma de vídeo, con la capacidad secuencial de una red neuronal LSTM bidireccional, que permite modelar la evolución temporal de las emociones. El modelo ha sido adaptado e implementado con una serie de mejoras técnicas, incluyendo un punto de entrada más flexible (*run_video(input_source)*), el uso de *argparse* y funciones de registro mediante *datetime*, así como cambios de nomenclatura y mejoras en la presentación visual, como la visualización optimizada de etiquetas y tasas de FPS.

El sistema realiza detección facial a través del módulo *FaceMesh* de MediaPipe, el cual proporciona una estimación robusta y rápida de los *landmarks* faciales. A partir de los puntos detectados, se calcula un recorte automático del rostro, que es

posteriormente preprocesado y transformado para alimentar al modelo ResNet-50. Cada imagen procesada genera un vector de características que se almacena temporalmente en un *buffer* circular. Una vez acumulada una secuencia de longitud fija (por ejemplo, 10 frames), los vectores son concatenados y entregados a la red LSTM, que genera como salida la distribución de probabilidad sobre las siete emociones básicas. La predicción más probable se proyecta sobre la imagen, junto con el valor porcentual de confianza y la tasa de procesado (FPS), similar al enfoque implementado en el modelo basado en Vision Transformers.

Para validar el sistema, se ha llevado a cabo una evaluación en tiempo real a través de una webcam, siendo yo, el propio autor del presente trabajo, el sujeto de prueba.

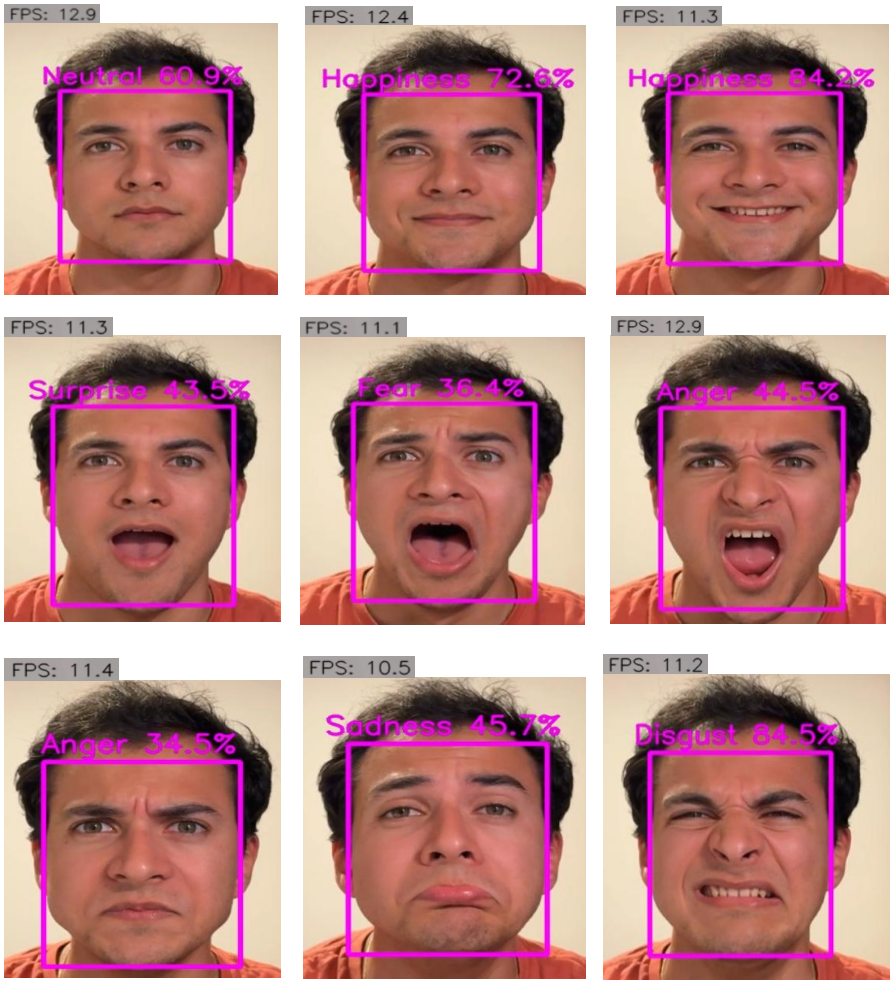

Figura 4.4. Frames a tiempo real detectando emociones en CNN-LSTM. Webcam
Fuente [Figura de elaboración propia]

En esta prueba (véase la figura 4.4), se simularon distintas expresiones faciales representativas de las siete emociones básicas, con el objetivo de comprobar la robustez y fluidez del sistema bajo condiciones reales. El sistema fue capaz de detectar y clasificar de forma coherente las emociones expresadas, mostrando una alta responsividad y continuidad visual. Durante toda la sesión, el sistema mantuvo una tasa media de procesamiento de entre 10 y 13 FPS, superando significativamente el rendimiento observado con el modelo ViT-B/32, que se situaba entre 4.8 y 5.1 FPS.

Como prueba adicional de generalización, de la misma manera que se ha procedido con el modelo anterior, se ha aplicado el sistema sobre un vídeo en tiempo real de un sujeto de la base de datos DEAP (véase la figura 7.16). En esta evaluación se observaron ligeras dificultades en la identificación de micro-expresiones, como una sonrisa sutil que fue etiquetada como “neutral” (30.1 %) y no como “felicidad”, y posteriormente fue reconocida como “felicidad” con una baja confianza del 26.3 %. Este comportamiento evidencia la sensibilidad del modelo CNN+LSTM ante variaciones de baja intensidad emocional y plantea la necesidad de ajustar los umbrales de confianza o refinar el entrenamiento para mejorar la sensibilidad.

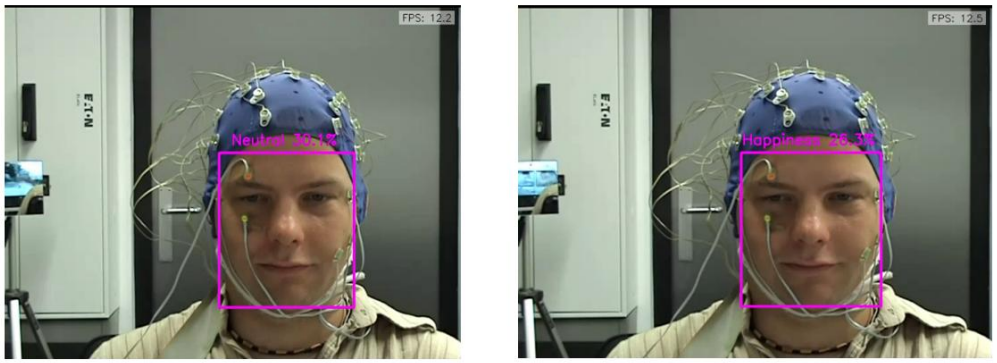

Figura 4.5. Secuencia de frames mediante archivo de vídeo en CNN-LSTM
Fuente [Figura de elaboración propia]

Pese a los buenos resultados funcionales, uno de los principales inconvenientes de esta arquitectura radica en la imposibilidad, en este trabajo, de calcular métricas cuantitativas estándar (precisión, *F1-score*, *recall*, etc.). Esta restricción se debe a la complejidad de la estructura del sistema, que emplea dos modelos pre-entrenados independientes —uno para la extracción espacial (ResNet-50 con AffectNet) y otro para el análisis temporal (LSTM con Aff-Wild2)— lo que dificulta la evaluación conjunta sin

una reimplementación completa del pipeline de validación. En contraste, el modelo ViT-B/32 permite una evaluación integrada al tratarse de una única arquitectura entrenable y trazable.

4.4. Logros, utilidad y limitaciones

El presente trabajo ha implementado y evaluado dos enfoques distintos para el reconocimiento automático de emociones en vídeo facial dinámico: por un lado, un sistema basado en *Vision Transformers* (ViT-B/32) integrado con el modelo CLIP, y por otro, una arquitectura híbrida CNN+LSTM inspirada en el trabajo de Elena Ryumina [46]. Ambos modelos han sido validados objetivamente mediante datos anotados y subjetivamente en condiciones prácticas mediante webcam y análisis en tiempo real o mediante archivo de video. Sin embargo, mientras que el modelo ViT-B/32 permite una evaluación cuantitativa estandarizada al estar entrenado de forma integrada sobre la base DFEW, el modelo CNN+LSTM no permite obtener métricas comparables. Esta limitación se debe a que la arquitectura CNN+LSTM está compuesta por dos modelos pre-entrenados por separado —ResNet-50 con *AffectNet* para la extracción espacial y una LSTM entrenada sobre Aff-Wild2 para la secuencia temporal—, lo que dificulta una trazabilidad y evaluación conjunta sin un rediseño completo del pipeline. Esta diferencia metodológica impone restricciones significativas a la hora de realizar un análisis estadístico riguroso de su rendimiento.

El modelo ViT-B/32 ha sido entrenado de forma integrada sobre la base de datos DFEW y posteriormente sometido a validación cruzada sobre MAFW), un corpus más complejo y naturalista. Esta metodología ha permitido realizar una evaluación objetiva, considerando tanto métricas globales como individuales por clase. En la fase de entrenamiento y prueba en DFEW, el modelo obtuvo resultados consistentes, con un WAR del 69.31 % y un *F1-score* macro del 60.69 % en entrenamiento, y un WAR del 70.31 % y F1 macro del 57.31 % en test. Estos valores reflejan una sólida capacidad de generalización dentro del dominio de entrenamiento (véase la tabla 5.1). Emociones como “felicidad” y “neutral” fueron reconocidas con alta fiabilidad, mientras que categorías como “miedo” y “repulsión” presentaron un menor rendimiento, debido probablemente a su baja representación y alta ambigüedad expresiva.

	Entrenamiento (%)	Prueba (%)	Validación (%)
WAR	69.31	70.31	62.30
UAR	58.90	57.25	55.88
Precisión macro	67.37	57.68	59.93
F1 macro	60.69	57.31	54.25

Tabla 4.133. Comparación métricas globales por fases DFEW-MAFW
Fuente [Tabla de elaboración propia]

En cuanto a la validación cruzada con MAFW, el rendimiento fue técnicamente satisfactorio, alcanzando un WAR del 62.30 %, UAR del 55.88 % y un F1 macro del 54.25 %. Aunque estas cifras son algo inferiores a las obtenidas en DFEW, muestran una generalización aceptable frente a un entorno externo más exigente, caracterizado por expresiones espontáneas, anotaciones complejas y condiciones visuales variables. Las emociones de “felicidad” y “tristeza” mantuvieron un buen nivel de reconocimiento, con *F1-score*s del 82.04 % y 73.90 % respectivamente, lo que confirma la robustez del modelo frente a señales emocionales marcadas incluso en situaciones realistas.

Desde un punto de vista comparativo, el modelo ViT-B/32 se posiciona en un rango competitivo respecto al estado del arte. Modelos tradicionales como VGG-16 o ResNet-18 acoplados a LSTM suelen alcanzar entre un 58 % y 65 % de F1 macro en DFEW, pero sus resultados descienden hasta el 50–55 % en validaciones cruzadas sobre conjuntos como AFEW o Aff-Wild2. Por su parte, arquitecturas más recientes como *TimeSformer o Video Swin Transformer* pueden superar el 65 % en F1 macro, aunque a costa de una mayor demanda computacional y una limitada evaluación externa. En este contexto, ViT-B/32 proporciona un equilibrio eficaz entre rendimiento, coste computacional y aplicabilidad real, destacando además por integrar explícitamente una validación cruzada sobre un dominio externo como MAFW, lo cual aporta un valor añadido y refuerza la fiabilidad del sistema.

En la evaluación práctica en tiempo real, el modelo ViT-B/32 fue probado mediante simulación de expresiones por parte del autor frente a una cámara webcam. Se observó una alta coherencia entre las emociones expresadas y las predicciones del modelo, especialmente en clases como “felicidad”, “neutral” o “ira”, con niveles de

confianza superiores al 90 %. Emociones como “tristeza” o “miedo” presentaron más variabilidad, reflejando la dificultad del modelo para capturar emociones menos evidentes. El sistema logró un rendimiento medio de entre 4.8 y 5.1 frames per second (FPS), aceptable para aplicaciones interactivas básicas aunque limitado por el uso de CPU.

El modelo CNN+LSTM, en cambio, ofreció una experiencia fluida y eficiente en tiempo real, alcanzando tasas de entre 10 y 13 FPS, lo que constituye una ventaja destacada para aplicaciones embebidas o con limitaciones de hardware. En las pruebas realizadas, el sistema respondió correctamente a las emociones representadas por el autor del estudio, y funcionó también sobre vídeo externo (DEAP), si bien mostró ciertas dificultades para captar emociones de baja intensidad, como una sonrisa sutil que fue inicialmente clasificada como “neutral”. Este comportamiento sugiere que, aunque el sistema es sensible y operativo, puede requerir refinamientos adicionales para aumentar su precisión perceptual. No obstante, la carencia de métricas cuantitativas impide realizar una evaluación formal y limita su comparabilidad directa con otros modelos.

En términos comparativos, el modelo ViT-B/32 destaca por su trazabilidad evaluativa, coherencia estadística y enfoque multimodal basado en lenguaje visual, lo que lo hace idóneo En paralelo, se implementó y probó una arquitectura CNN+LSTM, construida a partir de la extracción de características espaciales con una ResNet-50 pre-entrenada en AffectNet y una LSTM temporal entrenada sobre Aff-Wild2. Este modelo no permite obtener métricas cuantitativas comparables, debido a la falta de un entrenamiento conjunto que permita rastrear el flujo completo de procesamiento. No obstante, en pruebas prácticas mediante webcam y archivos de vídeo (por ejemplo, desde la base DEAP), el sistema alcanzó entre 10 y 13 FPS, ofreciendo una mayor fluidez operativa que el modelo ViT. El sistema reconoció correctamente emociones expresadas por el autor, aunque mostró cierta insensibilidad ante emociones de baja intensidad, como sonrisas sutiles mal clasificadas como “neutral”.

Ambos enfoques presentan ventajas y limitaciones complementarias. El modelo ViT-B/32 destaca por su robustez cuantitativa, trazabilidad evaluativa y su capacidad para generalizar en dominios nuevos. Sin embargo, su principal debilidad radica en la ausencia de una modelización explícita de la dimensión temporal, lo que puede comprometer su eficacia en emociones que evolucionan progresivamente. Por otro lado, el modelo CNN+LSTM es más eficiente computacionalmente y mejor adaptado a

entornos en tiempo real, aunque su falta de cohesión en el entrenamiento impide una evaluación objetiva y dificulta la comparación con modelos estandarizados.

Otra limitación compartida por ambos modelos es la dificultad para reconocer emociones minoritarias o ambiguas como “miedo” o “repulsión”. Esta debilidad, visible tanto en los resultados cuantitativos como en la observación directa, resalta la necesidad de seguir perfeccionando los algoritmos, ya sea mediante técnicas de aumento de datos, ponderación por clase o incorporación de mecanismos atencionales. Asimismo, el desequilibrio entre clases presente en DFEW y MAFW afecta la capacidad del modelo para aprender representaciones equitativas, como lo demuestra la brecha constante entre las métricas WAR y UAR.

Considerando todos los elementos analizados, este trabajo ha demostrado que ambos modelos —aunque distintos en naturaleza, arquitectura y operatividad— son viables para tareas de detección emocional en vídeo facial. El ViT-B/32 ofrece rigor evaluativo y capacidad de generalización, mientras que el CNN+LSTM proporciona eficiencia y velocidad en tiempo real. Ambos enfoques sientan las bases para el desarrollo de sistemas más sofisticados, sensibles y adaptables al reconocimiento emocional humano.

CAPITULO 5

CONCLUSIONES Y LÍNEAS FUTURAS

En el sexto y último capítulo del trabajo se presentan las principales conclusiones derivadas del proceso de desarrollo, evaluación y validación de los modelos implementados para el reconocimiento dinámico de emociones faciales mediante aprendizaje profundo. Asimismo, se identifican las líneas futuras más relevantes que podrían permitir una mejora del sistema propuesto o su adaptación a nuevos contextos de aplicación.

Constituye una reflexión crítica y constructiva sobre los resultados obtenidos, el grado de cumplimiento de los objetivos planteados y las oportunidades de evolución tecnológica del sistema diseñado.

5.1. Consecución de objetivos

En términos generales, los objetivos planteados al inicio del presente Trabajo Fin de Grado han sido alcanzados de forma satisfactoria.

El objetivo principal consistía en el desarrollo, implementación y evaluación de un sistema computacional para la determinación automática del estado emocional a través del análisis de vídeo facial dinámico, utilizando técnicas avanzadas de Deep Learning. Para ello, se propusieron dos enfoques arquitectónicos diferenciados: un modelo basado en Vision Transformers (ViT-B/32) dentro del marco multimodal CLIP, y un modelo híbrido CNN+LSTM orientado a aplicaciones en tiempo real.

El trabajo se ha centrado principalmente en la validación formal del modelo ViT-B/32, el cual ha sido entrenado sobre la base DFEW y validado mediante test directo y validación cruzada sobre la base MAFW, evidenciando su capacidad discriminativa y su rendimiento en entornos complejos y realistas. Se han obtenido métricas globales e

individuales que han permitido realizar un análisis cuantitativo detallado y riguroso de su comportamiento.

Además, se ha desarrollado e implementado un sistema CNN+LSTM funcional para su uso en tiempo real con entrada desde webcam. Este modelo ha sido adaptado a partir de la arquitectura propuesta por Elena Ryumina en su investigación sobre reconocimiento emocional en vídeo. Se ha modificado y ajustado dicho enfoque para adaptarlo a los requisitos del presente trabajo. Aunque no se ha podido evaluar cuantitativamente este modelo debido a la falta de anotaciones dinámicas específicas, su implementación ha sido validada empíricamente mediante observación de resultados coherentes en pruebas visuales.

Se han alcanzado también los objetivos específicos propuestos:

- Se ha integrado y preprocesado correctamente información visual en forma de vídeo dinámico, extrayendo regiones faciales relevantes mediante técnicas automatizadas.
- Se ha diseñado un pipeline modular para el entrenamiento, test y validación cruzada de modelos afectivos.
- Se ha comparado experimentalmente el rendimiento entre dos paradigmas representativos del DL, diferenciando entre precisión estadística y operatividad en tiempo real.
- Se ha adaptado el sistema al reconocimiento emocional dinámico, considerando los requisitos de entrada continua en tiempo real y la estabilidad del reconocimiento a través del tiempo.

En conjunto, el sistema desarrollado constituye una base robusta para el reconocimiento automático de emociones humanas, y permite su adaptación a distintos contextos como la interacción afectiva hombre-máquina o la monitorización emocional en entornos clínicos.

5.2. Conclusiones

El sistema propuesto ha demostrado que es posible aplicar arquitecturas modernas de Deep Learning al problema del reconocimiento emocional en vídeo facial

dinámico, combinando extracción de características espaciales con codificación temporal para capturar la evolución de la expresión emocional.

El modelo ViT-B/32 ha logrado un rendimiento elevado en emociones con rasgos prototípicos como “felicidad”, “ira” y “tristeza”, alcanzando un F1-score de hasta 91.04 % en pruebas sobre DFEW. Asimismo, su validación cruzada en MAFW ha permitido constatar una capacidad de generalización adecuada, incluso ante condiciones no vistas, expresiones espontáneas y mayor variabilidad gestual.

Por su parte, la arquitectura CNN+LSTM ha ofrecido un rendimiento funcionalmente válido en escenarios de vídeo en tiempo real, mostrando estabilidad, fluidez y precisión cualitativa aceptable en la identificación de emociones básicas. Su eficiencia computacional y su implementación ligera lo convierten en un candidato adecuado para aplicaciones embebidas o contextos interactivos. No obstante, la falta de evaluación cuantitativa limita su comparación directa con ViT-B/32.

Los resultados obtenidos reflejan las fortalezas de los sistemas actuales de reconocimiento afectivo, pero también las limitaciones persistentes. En particular, el modelo ViT-B/32 presenta dificultades notables en la clasificación de emociones menos representadas o ambiguas como “repulsión” y “miedo”, que suelen mostrar mayor solapamiento gestual o menor representación en los datasets. Esto evidencia la necesidad de mayor equilibrio y diversidad en los conjuntos de datos emocionales.

En definitiva, se ha conseguido desarrollar un sistema funcional y técnicamente avanzado para la determinación del estado emocional mediante vídeo facial, cumpliendo con los objetivos del TFG y sentando las bases para aplicaciones futuras tanto en investigación como en desarrollo tecnológico en campos como la salud emocional, la robótica social o la docencia inteligente.

5.3. Líneas futuras

A pesar de los avances alcanzados a lo largo del presente Trabajo Fin de Grado y del cumplimiento satisfactorio de los objetivos planteados, existen múltiples líneas de mejora y extensión que podrían ser abordadas en trabajos futuros. Una de las limitaciones más evidentes se encuentra en la capacidad del modelo para discriminar

emociones minoritarias o de expresión ambigua, como “repulsión” o “miedo”, cuya baja representación en los conjuntos de datos ha afectado negativamente a la robustez del sistema. Una solución clave a esta problemática consistiría en ampliar y diversificar los datasets empleados, incorporando bases de datos con un mayor equilibrio interclase y una representación cultural más heterogénea. Esto permitiría reducir los sesgos y aumentar la capacidad de generalización del sistema en contextos reales.

Asimismo, uno de los retos técnicos pendientes ha sido la evaluación cuantitativa del modelo CNN+LSTM desarrollado. Aunque se ha verificado su funcionalidad en tiempo real mediante cámara webcam, no ha sido posible establecer una comparación objetiva con ViT-B/32 debido a la ausencia de un conjunto de vídeos dinámicos anotado cuadro a cuadro. La elaboración de un dataset adaptado a las exigencias del análisis secuencial permitiría no solo validar empíricamente el rendimiento de dicho modelo, sino también consolidar su uso en entornos prácticos que requieran procesamiento emocional en línea.

Otra línea prometedora reside en la optimización del modelo CNN+LSTM para su integración en dispositivos embebidos de bajo consumo como *Raspberry Pi, Jetson Nano* o plataformas móviles. Dada su eficiencia computacional, esta arquitectura se presenta como candidata ideal para aplicaciones portátiles de detección emocional en tiempo real, especialmente útiles en contextos educativos, de atención sanitaria domiciliaria o asistencia personalizada.

En paralelo, se propone extender la arquitectura actual hacia un enfoque multimodal. La incorporación de señales fisiológicas (como frecuencia cardíaca, actividad electrodermal o tono de voz) podría complementar la información visual, incrementando así la precisión del sistema en casos de neutralidad expresiva o expresiones sutiles. Esta fusión de modalidades favorecería un reconocimiento emocional más robusto y adaptado a la complejidad del comportamiento humano.

También se sugiere explorar la evolución del sistema hacia un análisis afectivo continuo. En lugar de realizar una clasificación puntual por clip, el objetivo sería captar y modelar la evolución temporal del estado emocional de una persona, detectando patrones afectivos sostenidos que puedan tener relevancia clínica o psicológica. Este tipo de enfoque sería especialmente útil en procesos de terapia, monitorización del bienestar o evaluación de la carga emocional en entornos laborales.

Finalmente, sería de gran valor aplicar el sistema propuesto en escenarios reales mediante estudios piloto, con el fin de evaluar su rendimiento operativo, su aceptación por parte de los usuarios y su posible integración en plataformas clínicas, educativas o de atención al cliente. Esta validación práctica contribuiría significativamente a determinar la utilidad real del sistema más allá del laboratorio, abriendo nuevas posibilidades de innovación en la interacción humano-máquina.

BIBLIOGRAFÍA

[1] C. Darwin, *The Expression of the Emotions in Man and Animals*, London, UK: John Murray, 1872.

[2] D. Graziotin, X. Wang, and P. Abrahamsson, “Do feelings matter? On the correlation of affects and the self-assessed productivity in software engineering,” *Journal of Software: Evolution and Process*, vol. 27, no. 7, pp. 467–487, Jul. 2015.

[3] P. Ekman and W. V. Friesen, “Constants across cultures in the face and emotion,” *Journal of Personality and Social Psychology*, vol. 17, no. 2, pp. 124–129, 1971.

[4] P. Ekman and W. V. Friesen, *Facial Action Coding System: A technique for the measurement of facial movement*, Consulting Psychologists Press, 1978.

[5] P. Ekman, *Emotions Revealed: Recognizing Faces and Feelings to Improve Communication and Emotional Life*, Times Books, 2003.

[6] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, MIT Press, 2016.

[7] G. E. Hinton, S. Osindero, and Y. W. Teh, “A fast learning algorithm for deep belief nets,” *Neural Computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[8] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[9] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed., MIT Press, 2018.

[10] F. Carceller Llorens, “Detección de estados de ánimo usando técnicas de machine learning,” TFM, Universitat Politècnica de València, 2023.

[11] W. S. McCulloch and W. Pitts, “A logical calculus of the ideas immanent in nervous activity,” The Bulletin of Mathematical Biophysics, vol. 5, no. 4, pp. 115–133, 1943.

[12] D. O. Hebb, The Organization of Behavior: A Neuropsychological Theory. New York, NY, USA: Wiley, 1949.

[13] F. Rosenblatt, “The perceptron: A probabilistic model for information storage and organization in the brain,” Psychological Review, vol. 65, no. 6, pp. 386–408, 1958.

[14] M. Minsky and S. Papert, Perceptrons: An Introduction to Computational Geometry. Cambridge, MA, USA: MIT Press, 1969.

[15] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning representations by back-propagating errors,” Nature, vol. 323, pp. 533–536, Oct. 1986.

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in Proc. Advances in Neural Information Processing Systems, 2012, vol. 25, pp. 1097–1105.

[17] IBM, "¿Qué es una red neuronal?", [Online]. Available: https://www.ibm.com/think/topics/neural-networks. Fecha último acceso: 26.04.2025.

[18] J. Schmidhuber, "Deep learning in neural networks: An overview," Neural Networks, vol. 61, pp. 85–117, 2015.

[19] M. A. Nielsen, Neural Networks and Deep Learning. Determination Press, 2015. [Online]. Available: https://neuralnetworksanddeeplearning.com. Fecha último acceso: 26.04.2025.

[20] L. Alzubaidi et al., "Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions," Journal of Big Data, vol. 8, no. 1, p. 53, 2021. DOI:10.1186/s40537-021-00444-8.

[21] M. E. Shiri, A. Nabizadeh, and M. A. Nematbakhsh, "A review of recent advances in CNN interpretability and visualization," Artificial Intelligence Review, vol. 56, pp. 3735–3762, 2023. DOI:10.1007/s10462-022-10246-z.

[22] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A database for facial expression, valence, and arousal computing in the wild," IEEE Transactions on Affective Computing, vol. 10, no. 1, pp. 18–31, Jan.–Mar. 2019, doi: 10.1109/TAFFC.2017.2740923..

[23] D. Das, A. Das, and C. K. Panigrahi, "A review on recent advances in RNN-based deep learning for sequential data," Neural Comput. Appl., vol. 35, pp. 11999–12018, 2023. DOI:10.1007/s00521-022-07314-0.

[24] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," IEEE Trans. Neural Netw., vol. 5, no. 2, pp. 157–166, 1994. DOI:10.1109/72.279181.

[25] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with LSTM," Neural Computation, vol. 12, no. 10, pp. 2451–2471, Oct. 2000. DOI: 10.1162/089976600300015015.

[26] K. Cho et al., "Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation," arXiv preprint, 2014. [Online]. Available: https://arxiv.org/abs/1406.1078. Fecha último acceso: 16.05.2025.

[27] X. Chen, Y. Lin, and J. Zhao, "Recurrent Neural Networks for Sequence Learning: A Survey," Mathematics, vol. 8, no. 10, p. 1683, 2020. doi: 10.3390/math8101683.

[28] A. Vaswani et al., "Attention is all you need," in Advances in Neural Information Processing Systems, vol. 30, pp. 5998–6008, 2017.

[29] A. Dosovitskiy et al., “An image is worth 16x16 words: Transformers for image recognition at scale,” in Proc. Int. Conf. Learn. Represent., 2021.

[30] ResearchGate, “Transformer models: limitations and future directions,” [Online]. Available: https://www.researchgate.net/publication/361422805. Fecha último acceso: 18.05.2025.

[31] Z. Zhang et al., “From facial expression recognition to interpersonal relation prediction,” Int. J. Comput. Vis., vol. 126, no. 5, pp. 550–569, 2018.

[32] V. Alarcón-Aquino et al., “Deep learning-based emotion recognition: A survey,” Appl. Sci., vol. 13, no. 2, p. 849, 2023.

[33] Y. Fan et al., “Video-based emotion recognition using CNN-RNN and C3D hybrid networks,” in Proc. ICMI, 2016.

[34] B. Hasani and M. H. Mahoor, “Facial expression recognition using enhanced deep 3D convolutional neural networks,” in Proc. CVPRW, 2017.

[35] Y. Zhao et al., “Learning emotion features with spatial attention in facial expression recognition,” Sensors, vol. 21, no. 4, p. 1292, 2021.

[36] P. Zhou, Y. Zhao, and M. Xu, “FER-Former: Facial Expression Recognition with Transformers,” IEEE Trans. Affective Comput., 2023.

[37] R. Pereira et al., “Systematic review of emotion detection with computer vision and deep learning,” Sensors, vol. 23, no. 16, p. 7092, 2023. DOI:10.3390/s23167092.

[38] C. Halkiopoulos, E. Gkintoni, A. Aroutzidis, and H. Antonopoulou, “Advances in Neuroimaging and Deep Learning for Emotion Detection: A Systematic Review of Cognitive Neuroscience and Algorithmic Innovations,” Diagnostics, vol. 15, no. 4, p. 456, 2025. DOI:10.3390/diagnostics15040456.

[39] M. Mattioli and F. Cabitza, “Not in My Face: Challenges and Ethical Considerations in Automatic Face Emotion Recognition Technology,” Mach. Learn. Knowl. Extr., vol. 6, no. 4, pp. 2201–2231, 2024. DOI:10.3390/make6040109.

[40] H.-C. Yang, A. R. Rahmanti, C.-W. Huang, and Y.-C. Li, “How Can Research on Artificial Empathy Be Enhanced by Applying Deepfakes?” J. Med. Internet Res., vol. 24, no. 3, p. e29506, 2022. DOI: 10.2196/29506.

[41] X. Jiang, Y. Huang, B. Ni, J. Yang, W. Zou, and L. Wang, “DFEW: A Large-Scale Database for Recognizing Dynamic Facial Expressions in the Wild,” in Proc. of the 28th ACM Int. Conf. on Multimedia, Seattle, WA, USA, 2020, pp. 2881–2889. DOI:10.1145/3394171.3413891.

[42] Y. Liu, Y. Zhang, B. Jiang, W. Lin, Y. Wang, and X. Li, “MAFW: A Multimodal Database for Affective Faces in the Wild,” in Proc. ACM Multimedia, 2022.

[Online]. Available: http://mafw-database.github.io. Fecha último acceso: 11.06.2025.

[43] Z. Zhao and I. Patras, “Prompting Visual-Language Models for Dynamic Facial Expression Recognition,” in Proc. BMVC, 2023.

[44] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 770–778.

[45] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, “HMDB: A large video database for human motion recognition,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), pp. 2556–2563, 2011.

[46] E. Ryumina, D. Dresvyanskiy, and A. Karpov, “In Search of a Robust Facial Expressions Recognition Model: A Large-Scale Visual Cross-Corpus Study,” Neurocomputing, 2022. DOI:10.1016/j.neucom.2022.10.013.

[47] S. Sokolova and G. Lapalme, “A systematic analysis of performance measures for classification tasks,” Information Processing & Management, vol. 45, no. 4, pp. 427–437, 2009.

[48] C. Busso et al., “IEMOCAP: Interactive emotional dyadic motion capture database,” Language Resources and Evaluation, vol. 42, no. 4, pp. 335–359, 2008.

[49] H. Kaya, F. Gürpınar, and A. A. Salah, “Video-based emotion recognition in the wild using deep transfer learning and score fusion,” Image and Vision Computing, vol. 65, pp. 66–75, 2017.

[50] X. Li et al., “Deep Expression: A Multimodal Database for Dynamic Facial Expression Recognition in the Wild,” in Proc. IEEE Int. Conf. on Multimedia and Expo (ICME), 2017.

[51] X. Jiang et al., “DFEW: A Large-Scale Database for Recognizing Dynamic Facial Expressions in the Wild,” in Proc. ACM Multimedia (MM), 2020.

[52] Y. Liu et al., “MAFW: A Multimodal Database for Dynamic Facial Expressions in the Wild,” in Proc. ACM International Conference on Multimedia (ACM MM), 2022.

[53] What is Supervised Learning — The Foundation of Machine Learning — Day 5 | by Khushichoudhary | Medium. Accedido: 26.06.2025. [En línea]. Disponible en: https://medium.com/@khushichoudhary1020/what-is-supervised-learning-the-foundation-of-machine-learning-c746d88a63d3. Fecha último acceso: 26.06.2025.

[54] What is Unsupervised Learning? - GeeksforGeeks. [En línea]. Disponible en: https://www.geeksforgeeks.org/machine-learning/unsupervised-learning/. Fecha último acceso: 26.06.2025.

[55] What is Machine Learning? Let's see what is Machine Learning and… | by Prabhavi Jayanetti | Analytics Vidhya | Medium. [En línea]. Disponible en: https://medium.com/analytics-vidhya/what-is-machine-learning-446d570cadab. Fecha último acceso: 26.06.2025.

[56] An Introduction to Deep Feedforward Neural Networks | by Reza Bagheri | TDS Archive | Medium. [En línea]. Disponible en: https://medium.com/data-science/an-introduction-to-deep-feedforward-neural-networks-1af281e306cd. Fecha último acceso: 26.06.2025.

[57] Intro to Deep Learning. By Kevin David Farinango | by Bootcamp AI | Medium. [En línea]. Disponible en: https://bootcampai.medium.com/intro-to-deep-learning-d2d1ef64473e. Fecha último acceso: 26.06.2025.

[58] GitHub - axelvanherle/imgConvolutionCuda-C: Written by Sem Kirkels, Nathan Bruggeman and Axel Vanherle. Grayscales an image, applies convolution, maximum pooling and minimum pooling. [En línea]. Disponible en: https://github.com/axelvanherle/imgConvolutionCuda-C. Fecha último acceso: 26.06.2025.

[59] ¿Qué es una red neuronal recurrente (RNN)? | IBM. Disponible en: https://www.ibm.com/es-es/think/topics/recurrent-neural-networks. Fecha último acceso: 26.06.2025.

[60] Introducing GRU, RNN, and LSTM: A Beginner's Guide to Understanding these Revolutionary Deep Learning Models | by Jyoti Dabass, Ph.D. | Python in Plain English. [En línea]. Disponible en: https://python.plainenglish.io/introducing-gru-rnn-and-lstm-a-beginners-guide-to-understanding-these-revolutionary-deep-35b509a34a5a. Fecha último acceso: 26.06.2025.

[61] Understanding Transformers: The Role of Attention Mechanisms in Modern NLP | by Romy Adi Rotbar | Medium. [En línea]. Disponible en: https://medium.com/@romybuch22/understanding-transformers-the-role-of-attention-mechanisms-in-modern-nlp-ba2895f3377b. Fecha último acceso: 26.06.2025.

[62] (a) The interdisciplinary background of emotion recognition; | Download Scientific Diagram. [En línea]. Disponible en: https://www.researchgate.net/figure/a-The-interdisciplinary-background-of-emotion-recognition-bmodalities-used-to_fig1_358515968. Fecha último acceso: 26.06.2025.

[63] Vision Transformer (ViT). Transformers have already… | by Rishabh Singh | Medium. [En línea]. Disponible en: https://medium.com/@RobuRishabh/vision-transformer-vit-39f627d04b2a. Fecha último acceso: 26.06.2025.

[64] C. Ferrari, J. Baptista Cardia Neto, H.-T. Wang, J.-L. Lyu, y S. Hui-Lin Chien, Dynamic Emotion Recognition and Expression Imitation in Neurotypical Adults and Their Associations with Autistic Traits, *Sensors 2024, Vol. 24, Page 8133*, vol. 24, n.º 24, p. 8133, dic. 2024. DOI:10.3390/S24248133.

[65] MAFW/academics at main · MAFW-database/MAFW · GitHub. [En línea]. Disponible en: https://github.com/MAFW-database/MAFW/tree/main/academics. Fecha último acceso: 26.06.2025.

[66] DFER-CLIP：革新的なビジュアル言語モデルによる動的顔表情認識 | AI-SCHOLAR | AI：(人工知能)論文・技術情報メディア. [En línea]. Disponible en: https://ai-scholar.tech/en/articles/large-language-models%2Fdfer-clip. Fecha último acceso: 26.06.2025.

[67] A. Pandey *et al.*, A Deep Learning-Based Hybrid CNN-LSTM Model for Location-Aware Web Service Recommendation, *Neural Processing Letters 2024 56:5*, vol. 56, n.º 5, pp. 1-25, sep. 2024. DOI:10.1007/S11063-024-11687-W.