GRADO EN INGENIERÍA MATEMÁTICA E INTELIGENCIA ARTIFICIAL

TRABAJO FIN DE GRADO

DETECCIÓN AUTOMÁTICA DE DOLOR EN PACIENTES NO COMUNICATIVOS MEDIANTE VISIÓN POR COMPUTADOR Y DATOS SINTÉTICOS GENERADOS CON INTELIGENCIA ARTIFICIAL

Autor: Mario Alonso Alonso

Director: David Contreras Bárcena

Madrid

Mayo de 2026

Declaro, bajo mi responsabilidad, que el Proyecto presentado con el título

“detección automática de dolor en pacientes no comunicativos mediante visión por computador y datos sintéticos generados con inteligencia artificial”

en la ETS de Ingeniería - ICAI de la Universidad Pontificia Comillas en el curso académico 2025/26 es de mi autoría, original e inédito y no ha sido presentado con anterioridad a otros efectos.

El Proyecto no es plagio de otro, ni total ni parcialmente y la información que ha sido tomada de otros documentos está debidamente referenciada.

Fdo.: Mario Alonso Alonso Fecha: 25/5/2026

Autorizada la entrega del proyecto

EL DIRECTOR DEL PROYECTO

Fdo.: David Contreras Bárcena Fecha: 25/5/2026

GRADO EN INGENIERÍA MATEMÁTICA E INTELIGENCIA ARTIFICIAL

TRABAJO FIN DE GRADO

DETECCIÓN AUTOMÁTICA DE DOLOR EN PACIENTES NO COMUNICATIVOS MEDIANTE VISIÓN POR COMPUTADOR Y DATOS SINTÉTICOS GENERADOS CON INTELIGENCIA ARTIFICIAL

Autor: Mario Alonso Alonso

Director: David Contreras Bárcena

Madrid

Agradecimientos

A mis abuelos por priorizar la educación

A mis padres por el bienestar y las oportunidades que me brindan

A mi hermana por confiar en mí

A Carmen por apoyarme durante todo el proceso

A mis compañeros por hacerme disfrutar de la universidad

DETECCIÓN AUTOMÁTICA DE DOLOR EN PACIENTES NO COMUNICATIVOS MEDIANTE VISIÓN POR COMPUTADOR Y DATOS SINTÉTICOS GENERADOS CON INTELIGENCIA ARTIFICIAL

Autor: Alonso Alonso, Mario.
Director: Contreras Bárcena, David.
Entidad Colaboradora: Escuela de Enfermería San Juan de Dios

RESUMEN DEL PROYECTO

El proyecto consiste en la creación de una base de datos sintética con vídeos que representen rostros sintiendo dolor, para después entrenar un modelo de visión por ordenador que combine precisión y eficiencia. Las muestras generadas cumplen con los requisitos de calidad y han demostrado ser útiles para diseñar un sistema robusto. El modelo ha aprendido a diferenciar las clases con facilidad en test, y se ha implementado un prototipo de asistencia al personal sanitario en un contexto hospitalario, notificando en tiempo real el nivel de dolor de los pacientes.

Palabras clave: Detección de dolor; Dataset sintético; Visión por ordenador; Tiempo real

1. Introducción

El trabajo tiene el objetivo de desarrollar un modelo de visión por ordenador capaz de clasificar el nivel de dolor de pacientes que no pueden comunicarse, pero que sin embargo son capaces de expresar sensaciones de incomodidad con su rostro. Actualmente existen pocas bases de datos de calidad, dada la difícil obtención de muestras, por lo que también se contribuirá en este sentido al producir vídeos sintéticos con inteligencia artificial generativa. Es fundamental el avance en este sector, ya que no existen mecanismos implementados en entornos reales de forma generalizada que puedan realizar esta tarea automáticamente. Si se consiguiera instalar y estandarizar herramientas de este tipo, se facilitaría el trabajo del personal sanitario al delegar, de manera parcial, la detección a sistemas perfeccionados que están disponibles en todo momento y que no sufren de la subjetividad propia de los humanos. Por todo esto, este proyecto pretende colaborar en la búsqueda de una atención más eficaz al paciente, incrementando su bienestar considerablemente.

2. Objetivos

En primer lugar, se buscará una inteligencia artificial generativa de última generación que pueda crear vídeos equivalentes a vídeos reales a ojos de un modelo de clasificación. Si las muestras generadas cumplen con este requisito, se solucionarían muchos problemas en el aspecto de la recopilación de datos, pues en cualquier contexto se podría atacar la falta de muestras con este método, un efecto más acentuado cuando los datos en cuestión son sensibles. Una vez producidos los vídeos sintéticos más satisfactorios posibles, se entrenará un modelo de clasificación capaz de procesar los aspectos espaciales, geométricos y temporales de las muestras. Dicho modelo deberá balancear la precisión de clasificación y la velocidad de inferencia, pues ambas características son fundamentales en el ámbito médico.

3. Descripción del modelo/sistema/herramienta

Para generar los vídeos se acudió a Veo3.1, la inteligencia artificial generativa de Google, especializada en realismo y que dispone de una API que permite la generación automatizada. Se generaron 3300 muestras de 4 segundos de duración perfectamente balanceadas entre 3 clases definidas en base a la escala ESCID (0 para no dolor, 1 para dolor moderado y 2 para dolor intenso).

Antes de entrenar o inferir, los vídeos deben ser procesados para que el modelo los pueda interpretar. De cada vídeo se extraen 16 fotogramas, los cuales se recortarán para solo capturar el rostro y se redimensionarán a imágenes de 224x224 píxeles.

El modelo de clasificación desarrollado se compone de varios módulos. Para la interpretación visual se han construido dos ramas paralelas, una que emplea el backbone de EfficientNet-B0 preentrenado en ImageNet para la interpretación espacial, y otra que usa 478 landmarks tridimensionales extraídas con Mediapipe (puntos que construyen una red en los rostros). Estas dos ramas se concatenan, otorgando una proporción bastante mayor a EfficientNet, y los vectores de características de los fotogramas de cada vídeo pasarán por una LSTM que capturará la dimensión temporal. Finalmente, el último estado oculto entrará en una cabeza clasificadora, que devolverá un vector de tres dimensiones, una para cada clase.

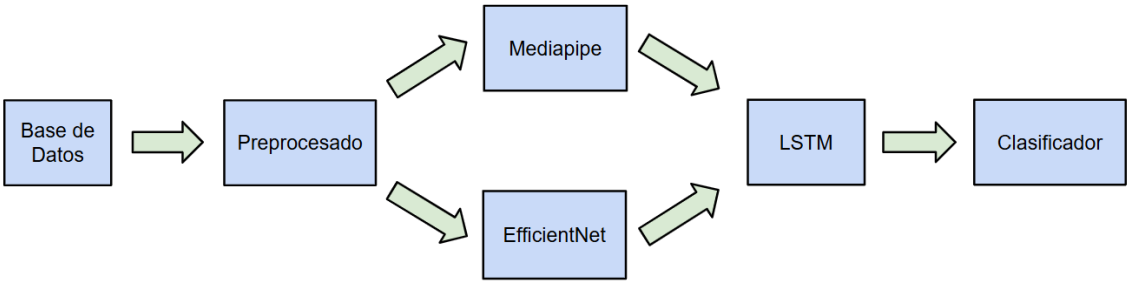

Ilustración 1. Arquitectura básica del sistema

En inferencia se recoge la señal captada por una cámara, preprocesa un fotograma cada 250ms y pasa por el modelo entrenado los últimos 16 fotogramas seleccionados. Al enfocar a un rostro, el usuario puede calibrar la posición del paciente y observar la probabilidad que el sistema asigna a cada clase en tiempo real.

4. Resultados

La base de datos ha sido validada por personal sanitario de la Escuela de Enfermería San Juan de Dios. La calidad de las muestras cumple los requisitos con holgura, y se calcula que recoge información útil para la gran mayoría de casos en un contexto real. Además, se ha usado OpenFace 2.0 como validación adicional, obteniendo unos resultados correlacionados con el índice PSPI, empleado en la mayoría de estudios del estado del arte.

Ilustración 2. Fotogramas extraídos de la base de datos sintética

Las métricas finales son excelentes: validación obtiene una pérdida inferior a 0,05 durante el entrenamiento, empleando entropía cruzada con doble penalización para las confusiones entre clase 0 y clase 2, y tanto validación como test resultan en precisión, recall y F1 Score en torno a 99% y AUC cercano a 1.

El prototipo en tiempo real funciona correctamente, y ha recibido la aprobación directa del personal sanitario. Aún está sujeto a condiciones favorables, como una posición de la cámara bien calibrada, una iluminación correcta, o un rango de expresiones faciales limitado, pero se trata de una versión inicial satisfactoria, que sienta el camino para desarrollos futuros prometedores.

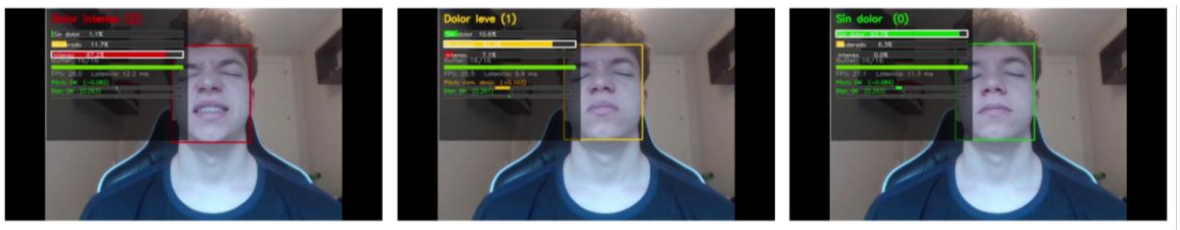

Ilustración 3. Muestra del funcionamiento del sistema de inferencia en tiempo real

5. Conclusiones

Se ha producido una base de datos extensa, validada, diversa, equilibrada y sin conflictos de privacidad, junto con una implementación capaz de generar más de 100 vídeos por hora, sin incluir paralelización. Además, se ha demostrado el potencial de este tipo de adquisición de datos; si es más fácil y rápido crear muestras de calidad, los modelos tendrán mejores condiciones para ser más precisos y generalizar mejor.

Por otro lado, se ha implementado un sistema capaz de clasificar dolor en tiempo real en un ordenador comercial, lo que, junto con la capacidad generativa de muestras sintéticas, invita al uso de modelos más pesados que puedan captar más información aún, preferiblemente con contexto temporal además de visual. Sin embargo, es importante mantener la capacidad de inferencia en tiempo real, para que estas soluciones puedan ser implementadas en los hospitales.

6. Referencias

[1] Latorre Marco, M. Solís Muñoz, T. Falero Ruiz, A. Larrasquitu Sánchez, A. B. Romay Pérez y I. Millán Santos, "Validación de la Escala de Conductas Indicadoras de Dolor para valorar el dolor en pacientes críticos, no comunicativos y sometidos a ventilación mecánica: resultados del proyecto ESCID," Enfermería Intensiva, vol. 22, no. 1, pp. 3–12, 2011, doi: 10.1016/j.enfi.2010.09.005.

[2] M. Tavakolian and A. Hadid, "A Spatiotemporal Convolutional Neural Network for Automatic Pain Intensity Estimation from Facial Dynamics," International Journal of Computer Vision, vol. 127, pp. 1413–1425, 2019, doi: 10.1007/s11263-019-01191-3.

[3] P. Rodriguez, G. Cucurull, J. Gonzàlez, J. M. Gonfaus, K. Nasrollahi, T. B. Moeslund, and F. X. Roca, "Deep Pain: Exploiting Long Short-Term Memory Networks for Facial Expression Classification," IEEE Transactions on Cybernetics, vol. 52, no. 5, pp. 3314–3324, 2022, doi: 10.1109/TCYB.2017.2662199

[4] X. L. Lin, S. Mehraban, A. Moturu, and B. Taati, "Pain in 3D: Generating Controllable Synthetic Faces for Automated Pain Assessment," arXiv preprint arXiv:2509.06727, 2025.

[5] B. Taati, M. Muzammil, Y. Zarghami, A. Moturu, A. Kazerouni, H. Reimer, A. Mihailidis, and T. Hadjistavropoulos, "SynPAIN: A Synthetic Dataset of Pain and Non-Pain Facial Expressions," arXiv preprint arXiv:2507.19673, 2025.

[6] G. D. De Sario, C. R. Haider, K. C. Maita, R. A. Torres-Guzman, O. S. Emam, F. R. Avila, J. P. Garcia, S. Borna, C. J. McLeod, C. J. Bruce, R. E. Carter, and A. J. Forte, "Using AI to Detect Pain through Facial Expressions: A Review," Bioengineering, vol. 10, no. 5, p. 548, 2023, doi: 10.3390/bioengineering10050548.

[7] N. Ben Aoun, "A Review of Automatic Pain Assessment from Facial Information Using Machine Learning," Technologies, vol. 12, no. 6, p. 92, 2024, doi: 10.3390/technologies12060092.

[8] M. Cascella, D. Esposito, M. R. Muzio, V. Cascella y V. Cerrone, "Artificial intelligence for pain assessment via facial expression recognition (2015–2025): a systematic review," Exploration of Medicine, vol. 6, art. no. 1001370, 2025, doi: 10.37349/emed.2025.1001370.

[9] J. Huo, Y. Yu, W. Lin, A. Hu y C. Wu, "Application of AI in multilevel pain assessment using facial images: systematic review and meta-analysis," Journal of Medical Internet Research, vol. 26, e51250, Apr. 2024, doi: 10.2196/51250.

[10] H. El-Ghaish, M. Y. Al-Basiouny y M. A. M. Alshewimy, "Enhanced deep learning framework for real-time pain assessment using multi-modal fusion of facial features and video streams," Scientific Reports, vol. 15, art. no. 18970, 2025, doi: 10.1038/s41598-025-03362-5.

[11] C. W. Tan, T. Du, J. C. Teo, D. X. H. Chan, W. M. Kong y B. L. Sng, “Automated pain detection using facial expression in adult patients with a customized spatial temporal attention long short-term memory (STA-LSTM) network,” Scientific Reports, vol. 14, art. no. 21778, 2024, doi: 10.1038/s41598-024-72427-4.

[12] P. Sankoh, A. Raza, K. Parwez, W. Shishah, A. Alharbi, M. Javed y M. Bilal, “Automated facial pain assessment using dual-attention CNN with clinically calibrated high-reliability and reproducibility framework,” IEEE Access, vol. 12, pp. 118746–118760, 2024, doi: 10.1109/ACCESS.2024.3445808.

[13] T. Alghamdi y G. Alaghband, “Facial expressions based automatic pain assessment system,” Healthcare, vol. 11, no. 3, art. no. 403, 2023, doi: 10.3390/healthcare11030403

[14] F. Alhamdoosh, P. Pala y S. Berretti, “Pain level estimation from videos by analyzing the dynamics of facial landmarks with a spatio-temporal graph neural network,” IEEE Access, vol. 12, pp. 27877–27890, 2024, doi: 10.1109/ACCESS.2024.3367684.

AUTOMATIC DETECTION OF PAIN IN NON-COMMUNICATIVE PATIENTS USING COMPUTER VISION AND SYNTHETIC DATA GENERATED WITH ARTIFICIAL INTELLIGENCE

Author: Alonso Alonso, Mario.
Supervisor: Contreras Bárcena, David
Collaborating Entity: Escuela de Enfermería San Juan de Dios

ABSTRACT

The project consists of creating a synthetic video database representing facial expressions of pain, in order to train a computer vision model that combines accuracy and efficiency. The generated samples meet the required quality standards and have proven useful for designing a robust system. The model has learned to distinguish between classes with ease during testing, and a prototype assistance system for healthcare personnel has been implemented in a hospital context, providing real-time notifications of patients' pain levels.

Keywords: Pain Detection; Synthetic Dataset; Computer Vision; Real-Time

1. Introduction

The aim of this work is to develop a computer vision model capable of classifying the pain level of patients who cannot communicate verbally, but are still able to express discomfort through facial expressions. Currently, there are very few high-quality datasets available due to the difficulty of obtaining such samples, so this project also contributes in this regard by producing synthetic videos using generative artificial intelligence. Advancement in this field is essential, as there are currently no widely implemented mechanisms in real-world environments capable of performing this task automatically. If tools of this kind could be installed and standardized, healthcare professionals' workload could be eased by partially delegating detection tasks to highly optimized systems that are always available and free from the subjectivity inherent to humans. For all these reasons, this project aims to contribute to more efficient patient care and significantly improve patient well-being.

2. Objectives

First, the project seeks to identify a state-of-the-art generative artificial intelligence system capable of creating videos that are equivalent to real videos from the perspective of a classification model. If the generated samples satisfy this requirement, many data collection issues could be solved, as the lack of samples in any context could be addressed using this method, especially when dealing with sensitive data. Once the most satisfactory synthetic videos possible have been produced, a classification model capable of processing the spatial, geometric, and temporal aspects of the samples will be trained. This model must balance classification accuracy and inference speed, as both characteristics are essential in the medical field.

3. Description of the proposed model/system/tool

To generate the videos, Veo3.1 was used, Google's generative artificial intelligence specialized in realism and equipped with an API that enables automated generation. Slightly over 3,300 samples of 4-second duration were generated, perfectly balanced across three classes defined according to the ESCID scale (0 for no pain, 1 for moderate pain, and 2 for severe pain).

Before training or inference, the videos must be processed so that the model can interpret them. From each video, 16 frames are extracted, cropped to capture only the face, and resized to 224×224 pixel images.

The developed classification model consists of several modules. For visual interpretation, two parallel branches were built: one uses an EfficientNet-B0 backbone pretrained on ImageNet for spatial interpretation, while the other uses 478 three-dimensional landmarks extracted with Mediapipe (points forming a mesh over the face). These two branches are concatenated, assigning a significantly greater proportion to EfficientNet, and the feature vectors from each video frame are passed through an LSTM that captures the temporal dimension. Finally, the last hidden state is fed into a classification head that outputs a three-dimensional vector, one for each class.

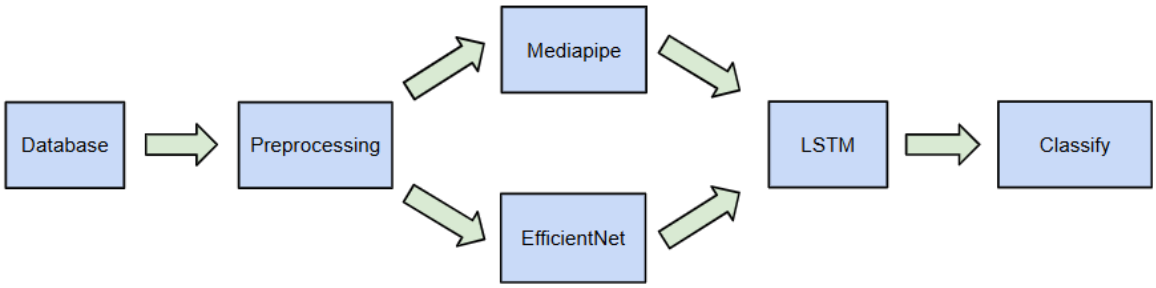

Figure 1. Basic architecture of the system

During inference, the signal captured by a camera is processed by selecting one frame every 250 ms, and the trained model analyzes the latest 16 selected frames. When a face is detected, the user can calibrate the patient's position and observe in real time the probability assigned by the system to each class.

4. Results

The dataset was validated by healthcare personnel from the San Juan de Dios School of Nursing. The quality of the samples comfortably meets the required standards, and it is estimated that the dataset captures useful information for the vast majority of real-world cases. Additionally, OpenFace 2.0 was used as an extra validation method, obtaining results correlated with the PSPI index, which is widely used in state-of-the-art studies.

Figure 2. Frames extracted from the synthetic dataset

The final metrics are excellent: validation achieved a loss below 0.05 during training, using cross-entropy with double penalization for confusions between classes 0 and 2, while both validation and test achieved accuracy, recall, and F1-score values around 99%, with an AUC close to 1.

The real-time prototype operates correctly and has received direct approval from healthcare personnel. It is still subject to favorable conditions, such as proper camera calibration, correct lighting, and a limited range of facial expressions, but it represents a satisfactory initial version that paves the way for promising future developments.

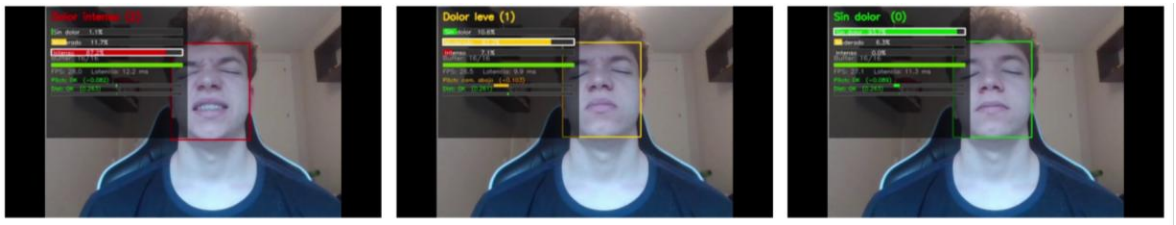

Figure 3. Example of the real-time inference system in operation

5. Conclusions

An extensive, validated, diverse, balanced, and privacy-safe dataset has been produced, together with an implementation capable of generating more than 100 videos per hour without parallelization. Furthermore, the potential of this type of data acquisition has been demonstrated: if it becomes easier and faster to create high-quality samples, models will have better conditions to achieve higher accuracy and improved generalization.

In addition, a system capable of classifying pain in real time on a commercial computer has been implemented. Combined with the ability to generate synthetic samples, this encourages the use of larger models capable of capturing even more information, preferably with both temporal and visual context. However, maintaining real-time inference capability remains essential so that these solutions can eventually be deployed in hospitals.

6. References

[1] Latorre Marco, M. Solís Muñoz, T. Falero Ruiz, A. Larrasquitu Sánchez, A. B. Romay Pérez y I. Millán Santos, "Validación de la Escala de Conductas Indicadoras de Dolor para valorar el dolor en pacientes críticos, no comunicativos y sometidos a ventilación mecánica: resultados del proyecto ESCID," Enfermería Intensiva, vol. 22, no. 1, pp. 3–12, 2011, doi: 10.1016/j.enfi.2010.09.005.

[2] M. Tavakolian and A. Hadid, "A Spatiotemporal Convolutional Neural Network for Automatic Pain Intensity Estimation from Facial Dynamics," International Journal of Computer Vision, vol. 127, pp. 1413–1425, 2019, doi: 10.1007/s11263-019-01191-3.

[3] P. Rodriguez, G. Cucurull, J. Gonzàlez, J. M. Gonfaus, K. Nasrollahi, T. B. Moeslund, and F. X. Roca, "Deep Pain: Exploiting Long Short-Term Memory Networks for Facial Expression Classification," IEEE Transactions on Cybernetics, vol. 52, no. 5, pp. 3314–3324, 2022, doi: 10.1109/TCYB.2017.2662199

[4] X. L. Lin, S. Mehraban, A. Moturu, and B. Taati, "Pain in 3D: Generating Controllable Synthetic Faces for Automated Pain Assessment," arXiv preprint arXiv:2509.06727, 2025.

[5] B. Taati, M. Muzammil, Y. Zarghami, A. Moturu, A. Kazerouni, H. Reimer, A. Mihailidis, and T. Hadjistavropoulos, "SynPAIN: A Synthetic Dataset of Pain and Non-Pain Facial Expressions," arXiv preprint arXiv:2507.19673, 2025.

[6] G. D. De Sario, C. R. Haider, K. C. Maita, R. A. Torres-Guzman, O. S. Emam, F. R. Avila, J. P. Garcia, S. Borna, C. J. McLeod, C. J. Bruce, R. E. Carter, and A. J. Forte, "Using AI to Detect Pain through Facial Expressions: A Review," Bioengineering, vol. 10, no. 5, p. 548, 2023, doi: 10.3390/bioengineering10050548.

[7] N. Ben Aoun, "A Review of Automatic Pain Assessment from Facial Information Using Machine Learning," Technologies, vol. 12, no. 6, p. 92, 2024, doi: 10.3390/technologies12060092.

[8] M. Cascella, D. Esposito, M. R. Muzio, V. Cascella y V. Cerrone, "Artificial intelligence for pain assessment via facial expression recognition (2015–2025): a systematic review," Exploration of Medicine, vol. 6, art. no. 1001370, 2025, doi: 10.37349/emed.2025.1001370.

[9] J. Huo, Y. Yu, W. Lin, A. Hu y C. Wu, "Application of AI in multilevel pain assessment using facial images: systematic review and meta-analysis," Journal of Medical Internet Research, vol. 26, e51250, Apr. 2024, doi: 10.2196/51250.

[10] H. El-Ghaish, M. Y. Al-Basiouny y M. A. M. Alshewimy, "Enhanced deep learning framework for real-time pain assessment using multi-modal fusion of facial features and video streams," Scientific Reports, vol. 15, art. no. 18970, 2025, doi: 10.1038/s41598-025-03362-5.

[11] C. W. Tan, T. Du, J. C. Teo, D. X. H. Chan, W. M. Kong y B. L. Sng, “Automated pain detection using facial expression in adult patients with a customized spatial temporal attention long short-term memory (STA-LSTM) network,” Scientific Reports, vol. 14, art. no. 21778, 2024, doi: 10.1038/s41598-024-72427-4.

[12] P. Sankoh, A. Raza, K. Parwez, W. Shishah, A. Alharbi, M. Javed y M. Bilal, “Automated facial pain assessment using dual-attention CNN with clinically calibrated high-reliability and reproducibility framework,” IEEE Access, vol. 12, pp. 118746–118760, 2024, doi: 10.1109/ACCESS.2024.3445808.

[13] T. Alghamdi y G. Alaghband, “Facial expressions based automatic pain assessment system,” Healthcare, vol. 11, no. 3, art. no. 403, 2023, doi: 10.3390/healthcare11030403

[14] F. Alhamdoosh, P. Pala y S. Berretti, “Pain level estimation from videos by analyzing the dynamics of facial landmarks with a spatio-temporal graph neural network,” IEEE Access, vol. 12, pp. 27877–27890, 2024, doi: 10.1109/ACCESS.2024.3367684.

Índice de la memoria

Índice de figuras

Índice de tablas

Capítulo 1. INTRODUCCIÓN

1.1 CONTEXTO Y MOTIVACIÓN

Muchos pacientes se encuentran en situaciones en las que su condición física no les permite comunicar el dolor que sienten al exterior de forma directa. A pesar de no estar en un estado de consciencia completo, es posible que pacientes anestesiados, sedados, comatosos, o en otros estados, puedan sentir dolor. Además, en ocasiones, tienen la capacidad de realizar expresiones faciales, voluntaria o involuntariamente, que nos permiten detectar estas sensaciones externamente.

Es en este contexto donde se fundamenta uno de los pilares de este proyecto, realizado en colaboración con la Escuela de Enfermería San Juan de Dios. Se tratará de implementar un sistema basado en visión por ordenador capaz de detectar, de forma automática y en tiempo real, el dolor en los pacientes a partir de los gestos en sus rostros. Esto supondrá aportaciones en el desarrollo de una herramienta que no requiera intervención externa, con el objetivo futuro de cumplir la tarea de forma más rápida, eficiente y precisa que incluso los propios humanos. En consecuencia, el proyecto no solo supondrá avances en el área de la inteligencia artificial, sino que también puede contribuir a lo que sería una mejora notable en el bienestar del paciente, pues el equipo sanitario podría conocer en todo momento qué estímulos le causan incomodidad, para así poder evitarlos y ofrecer una mejor atención médica.

Por otro lado, la creación de nuevos datos es fundamental tanto para desarrollar un trabajo propio, como para que otros investigadores puedan realizar sus propios avances. Esto se torna más importante aún en un área donde se trata con datos tan sensibles y difíciles de obtener y manejar, como lo es la medicina, y más aún si se trabaja con rostros, la parte más reconocible del cuerpo. Teniendo en cuenta que el objetivo es detectar dolor, el problema se vuelve mucho más sensible, porque existe un sufrimiento directo de por medio. Sin embargo, la mayor dificultad reside en el hecho de que los pacientes no se pueden comunicar, por lo que, si no se hace previamente, no se les podrá pedir consentimiento para usar sus datos. Es

por ello que se optará por generar los datos de manera sintética a través de inteligencia artificial generativa comercial de alta calidad, cumpliendo de manera más sencilla con las restricciones éticas y de privacidad, al mismo tiempo que se investiga una nueva manera de crear datos, aplicable a otras muchas áreas de la tecnología.

1.2 OBJETIVOS

Uno de los ejes principales del proyecto, y el primero en orden cronológico, es la creación de datos sintéticos depurados, con la calidad suficiente como para emular fielmente la realidad. Se busca crear una base de datos de videos a través de una inteligencia artificial generativa que sea capaz de ofrecer un alto realismo anatómico. Tras ello, con la ayuda de un equipo médico, se garantizará que las muestras cumplan con los requisitos de calidad, y que sus etiquetas se correspondan con la escala ESCID.

Otro objetivo fundamental es la implementación de una arquitectura de aprendizaje profundo que cumpla con precisión su tarea: detectar el nivel de dolor que un rostro humano puede comunicar al exterior. Esto se debe realizar en tiempo real, por lo que resulta fundamental dar con un modelo que optimice el balance entre precisión y velocidad. Para ello, se utilizará transferencia de aprendizaje, aplicando sistemas especializados y modificarlos para cumplir nuestros requisitos específicos.

Como última meta fundamental, se sitúa la creación de un prototipo operativo capaz de comunicarse con el equipo sanitario para ofrecer una mejor atención al paciente. El sistema debe poder transmitir de forma palpable la información que ofrezca el modelo, para que los sanitarios puedan tomar las decisiones correspondientes. Por tanto, la puesta en marcha y la utilización del sistema no debe ser complicada, pues será utilizado por personas que no están familiarizadas con el sistema que se está ejecutando detrás de la interfaz de usuario.

1.3 ALINEACIÓN CON LOS ODS

Entre los Objetivos de Desarrollo Sostenible, el proyecto contribuye en mayor magnitud en cuatro de ellos. En primer lugar, el trabajo aporta en el objetivo número tres "Salud y Bienestar", gracias al enfoque médico central. También se contribuye en gran medida al objetivo número ocho "Trabajo Decente y Crecimiento Económico", pues al automatizar procesos y mejorar su eficiencia se pueden reducir costes, y así ofrecer mejores servicios a un precio más accesible. En consecuencia, también se alinea con el objetivo número diez "Reducción de las Desigualdades", ya que democratizar la medicina de alta calidad para que cualquier persona, independientemente de su situación económica, pueda acceder a ella, ofrece oportunidades a todos los miembros de una sociedad para vivir una vida digna. Por último, el proyecto colabora en el objetivo nueve "Industria, Innovación e Infraestructura", como consecuencia de su naturaleza innovadora, empleando y colaborando en el avance de técnicas de tecnología punta.

1.4 ESTRUCTURA DEL TRABAJO

En consonancia con los objetivos, el trabajo se estructurará en tres secciones principales que explicarán la implementación realizada para cumplir cada uno de ellos, junto con la posterior exposición de los resultados obtenidos. Además, se hará una breve introducción al estado del arte dentro del sector, observando tendencias que pongan en contexto las decisiones tomadas en este proyecto. Por último, se sacarán conclusiones de lo analizado, sugiriendo algunas líneas de investigación futura que pudieran contribuir a avanzar en una dirección acertada.

Capítulo 2. ESTADO DEL ARTE

La detección de dolor mediante visión por ordenador es un campo activo desde hace varios años. Ha habido grandes avances, pero siguen existiendo limitaciones importantes, algunas de las cuales serán tratadas de resolver en el estudio.

Escalas de dolor

El Facial Action Coding System (FACS) establece una taxonomía de los movimientos faciales descomponiéndolos en unidades individuales denominadas Action Units (AUs), cada una asociada a la contracción de un músculo o grupo muscular específico. Basándose en FACS, Prkachin y Solomon (2008) identificaron un subconjunto específico de AUs consistentemente asociadas al dolor y formularon el Prkachin and Solomon Pain Intensity Index (PSPI). Esta escala toma valores discretos entre 0 y 16, ha sido validada clínicamente en múltiples estudios y constituye el estándar de referencia en investigación sobre detección automática de dolor facial.

La escala ESCID es una herramienta clínica utilizada por el personal de enfermería para evaluar la presencia y la intensidad del dolor en pacientes que no pueden comunicarse verbalmente. Diferencia cinco indicadores de dolor, que toman un valor entre 0, si no se da el indicador, 1, si se da moderadamente, o 2, si aparece de forma intensa. La musculatura facial es uno de estos indicadores, y será el centro de investigación en este trabajo. Se planteará un modelo de clasificación que diferencie directamente entre estas tres clases, en lugar de clasificar en función de la escala PSPI para después calibrar umbrales que traduzcan los valores obtenidos a la escala ESCID.

Bases de datos de personas reales

La UNBC-McMaster Shoulder Pain Expression Archive Database es la más empleada en el campo. Sin embargo, solamente incluye a 25 personas distintas, por lo que sufre de falta de

diversidad, sobre todo por edad, al no incluir a ancianos. Además, tiene un gran desbalance en los datos, sobrerrepresentando rostros que no muestran dolor.

Otra base de datos ampliamente usada es BioVid Heat Pain Database, que además incluye otras métricas biológicas como electrocardiogramas. Sin embargo, este tipo de datos no sirven en casos donde el foco se centra exclusivamente en el campo de la visión por ordenador. Por otro lado, BioVid también incluye sesgos, ya que incluye a 90 sujetos distintos, que sigue siendo un número muy bajo.

El contexto general del resto del campo es parecido: pocos datos, desbalances, diversidad baja. Es aquí donde surge la inteligencia artificial generativa como una alternativa seria, pues requiere menos consideraciones éticas, permite la generación masiva de datos, y dispone de una alta adaptabilidad a los requisitos que se le planteen. Se debe tener en cuenta que el realismo absoluto siempre va a ser una limitación a tener en cuenta; por ejemplo, los vídeos capturados en la mayoría de bases de datos del estado del arte se dan en situaciones planificadas y no en escenarios médicos reales, ofreciendo resultados no del todo genuinos. Además, aplicar dolor artificialmente también debe enfrentarse a restricciones éticas, ya que no se le pueden administrar estímulos excesivamente dolorosos a los participantes. Por otro lado, no en todos los casos hay un equipo médico que evalúe las muestras y, aunque lo haya, la medición del dolor siempre será subjetiva, incluso con el reporte de los propios pacientes, ya que cada persona tiene criterios distintos. Es por todas estas razones que cualquier avance en el campo de las bases de datos de este sector es positivo.

Bases de datos sintéticas

Recientemente, para lidiar con los problemas mencionados de las bases de datos del sector, han aparecido varios trabajos que exploran la calidad de las muestras sintéticas que la tecnología actual es capaz de generar.

3DPain (Lin et al., 2025) crea modelos faciales en tres dimensiones en los que modifican directamente los valores de las AUs para generar expresiones de dolor parametrizadas. Así,

obtuvieron una base de datos totalmente equilibrada en cuanto a diversidad y clases. No obstante, las muestras son imágenes estáticas que no ofrecen una dimensión temporal.

SynPAIN (Taati et al., 2025) innova al utilizar una inteligencia artificial generativa comercial controlada mediante consultas en lenguaje natural describiendo las expresiones de dolor. El resultado fue satisfactorio, obteniendo muestras no solo diversas, sino también validadas con métricas parecidas a las que ofrecen otras muestras reales. Sin embargo, SynPAIN también se centró más en la creación de imágenes, dejando el vídeo en un segundo plano.

Aunque aún no es muy extensa, se observa una base prometedora en la utilidad de las muestras artificiales, incluso las generadas con herramientas comerciales. El enfoque es mayoritario en imágenes estáticas, a pesar de que una base de datos de vídeos pueda ofrecer datos más informativos aún. Dada esta carencia en el estado del arte, las muestras generadas en este proyecto serán vídeos.

Modelos de clasificación

Los primeros enfoques para detectar dolor en expresiones faciales empleaban técnicas de aprendizaje automático clásico, extrayendo características faciales de forma manual mediante descriptores como AAM, que luego se clasificaban con SVM o KNN. Con la llegada del aprendizaje profundo se obtuvieron mejoras sustanciales, al permitir que la red aprendiese directamente las representaciones relevantes. En este sector destaca Alghamdi y Alaghband (2022), con una de las mayores precisiones registradas en UNBC-McMaster.

Sin embargo, los primeros sistemas de aprendizaje profundo trabajaban mayoritariamente sobre fotogramas individuales, ignorando la dinámica temporal de la expresión de dolor. Posteriormente se desarrollaron modelos híbridos que combinan redes convolucionales para la extracción de características espaciales con otros módulos como las LSTM, mejorando notablemente los resultados. Rodriguez et al. (2022) emplea este enfoque en concreto.

En los últimos años se está explorando el uso de atención y de Vision Transformers, mostrando resultados prometedores, aunque aún con algunas limitaciones. En muchos casos

tienen un coste computacional elevado para inferencia en tiempo real, lo que los hace menos adecuados para entornos clínicos con recursos limitados. Además, requieren grandes volúmenes de datos de preentrenamiento para evitar el sobreajuste, dada la alta cantidad de parámetros que incluyen. Paralelamente, los enfoques multimodales que integran expresión facial con audio, geometría, señales fisiológicas, o movimiento corporal obtienen de forma consistente los mejores resultados. Alhamdoosh et al. (2025) trabaja con vídeos y usa landmarks faciales para recoger información geométrica.

Por todo esto, se ha decidido implementar un modelo híbrido que se apoye la base sólida que ofrece la combinación del Deep Learning y la temporalidad. Por otro lado, dado que el trabajo se centra exclusivamente en el reconocimiento facial, no se pueden implementar muchos de los sistemas multimodales. Sin embargo, sí es posible incluir una fuente de información geométrica complementaria.

Otra de las mayores carencias del campo es la falta de soluciones que clasifiquen en tiempo real. El-Ghaish et al. (2025) combina información temporal, eficiencia y uso de landmarks, siendo uno de los trabajos con un enfoque más cercano al que se presenta en este proyecto.

Capítulo 3. BASE DE DATOS SINTÉTICA

3.1 PLANTEAMIENTO DEL PROBLEMA

El primer paso para construir un modelo de clasificación preciso es disponer de una base de datos de calidad. Considerando las dificultades expuestas en el apartado anterior, la premisa del proyecto es generar una base de datos sintética a través de inteligencia artificial generativa. Esto puede resolver varios problemas, pero también supone varios retos. Las muestras deben ser lo suficientemente realistas y diversas, para asegurar que el modelo pueda emplearse con rostros humanos reales, sin importar su etnia, sexo, o edad. Por otro lado, la generación es costosa, sobre todo para vídeo. También existen restricciones éticas a la hora de generar muestras artificiales, por lo que solicitar videos de rostros sintiendo dolor necesita de un lenguaje específico que el modelo generativo escogido pueda procesar sin problema.

Para validar el sistema, la escala ESCID no ofrece criterios enteramente objetivos, pero sí algunas referencias fácilmente observables para distinguir entre niveles de dolor, detallados en la siguiente figura:

ESCID	0	1	2
MUSCULATURA FACIAL	Relajada	En tensión, ceño fruncido/gesto de dolor	Ceño fruncido de forma habitual, dientes apretados
TRANQUILIDAD	Tranquilo, relajado, movimientos normales	Movimientos ocasionales, inquietud y/o posición	Movimientos frecuentes, incluyendo cabeza o extremidades
TONO MUSCULAR	Normal	Aumento de la flexión de dedos de manos y/o pies	Rígido
ADAPTACIÓN A VENTILACIÓN MECÁNICA	Tolera la ventilación mecánica	Tose, pero tolera la ventilación mecánica	Lucha con el respirador
CONFORTABILIDAD	Confortable, tranquilo	Se tranquiliza al tacto y/o a la voz. Fácil de distraer	Difícil de controlar al tacto o hablándole

Ilustración 1. Descripción de la escala ESCID (Enfermería Creativa)

La puntuación final para medir el dolor sería la suma del valor entre 0 y 2 de las 5 categorías, variando finalmente entre los 0 y 10 puntos. Este trabajo se centra en la categoría de musculatura facial, que distingue entre un rostro relajado, en tensión o nivel alto de dolor.

Por tanto, las muestras generadas deben poder ser validadas a simple vista por un humano, comparando con las pautas descritas. Además, se emplearán técnicas automatizadas como complemento para verificar la validez de la base de datos.

Aunque no se vaya a incluir por el momento, el indicador "Adaptación a la ventilación mecánica" también podría ser estudiado con la información facial, mientras que las otras categorías necesitarían de vídeos que capturen el cuerpo entero o sensores que capten otro tipo de información para ser evaluadas con éxito.

3.2 DISEÑO DE LA SOLUCIÓN

La inteligencia artificial generativa escogida es Veo3.1, de Google. Se trata de su modelo generativo de vídeo más avanzado, conocido por su fotorrealismo e incluso por su capacidad cinematográfica. Según la propia descripción de Google, "Veo 3.1 está diseñado para satisfacer las exigencias de las aplicaciones del mundo real" y para "necesidades de producción de alta calidad". La coherencia temporal, el realismo físico y la iluminación avanzada dotan a las muestras de la calidad suficiente y una semejanza al mundo real satisfactoria. También destaca por el sonido y la generación de diálogos, pero no son complementos útiles en este caso.

A un lado de la alta calidad, Veo3.1 ofrece una accesibilidad difícilmente superable. Google Cloud ofrece servicios gratuitos con valor de 300$ por cada cuenta de Google diferente, lo que permite generar varios centenares de vídeos y almacenarlos en la nube fácilmente. Además, Google ha desarrollado Vertex AI API, que permite automatizar de manera sencilla la creación masiva de vídeos, requisito indispensable para construir una base de datos con suficientes muestras.

El resto de inteligencias artificiales generativas estudiadas no han llegado a cumplir los requisitos de calidad y de capacidad de generación. Por ejemplo, al investigar sobre el famoso modelo Sora, se descubrió que su soporte no iba a ser continuado en el futuro. Por otro lado, otros modelos menos conocidos que prometían calidad, no disponían del

ecosistema tan completo con el que cuenta Google. Veo3.1 era sin duda la mejor opción por su calidad, facilidad de uso, y las funcionalidades de generación y almacenamiento.

3.3 IMPLEMENTACIÓN

Si la generación y almacenamiento de vídeos se iba a realizar en el ecosistema de Google, la plataforma Google Colab era idónea para desarrollar el código de automatización de estas tareas, ya que el intercambio de información entre unos servicios y otros se volvía altamente sencillo.

Dada la complejidad de escenas que Veo3.1 es capaz de generar, también debe soportar consultas elaboradas y detalladas. Aunque las consultas se han escrito en inglés, esta es la estructura final que siguen:

Una toma vertical fija y estática desde una cámara montada en el techo, sin movimiento de cámara ni zoom óptico. **{sexo} {edad} {etnia} {forma del rostro} {rasgos faciales} {descripción del pelo}***, boca arriba en una cama de hospital. Encuadre fijo que incluye los bordes de la almohada y muestra el rostro completo con amplios márgenes. La persona presenta* **{descripción de las expresiones faciales en función del grado de dolor seleccionado}** *de forma constante durante todo el video. Entorno clínico médico.* **{iluminación} {enfoque} {textura de la piel}**

Como se puede observar, la consulta final consta en primer lugar de dos frases introductorias, donde se especifica el sexo, la edad y la etnia de la persona en concreto, junto con el contexto fundamental de la escena. Se han incluido tres rangos de edad (entre 20 y 35 años, entre 45 y 60, y entre 70 y 85) y seis etnias generalistas (piel blanca, rasgos sudamericanos, piel oscura, rasgos árabes, rasgos asiáticos del este, y rasgos sudasiáticos). Después, se intenta detallar lo máximo posible el nivel de dolor oportuno, y se continúa con una descripción de los aspectos estéticos de la escena. La descripción de cada uno de los tres niveles de dolor es la siguiente:

- **Dolor nulo:** *La persona mantiene una expresión facial relajada y neutral. Ceño relajado, ojos gentilmente cerrados, boca cerrada de forma natural. Movimientos naturales de forma constante durante todo el vídeo.*
- **Dolor moderado:** *La persona mantiene una expresión facial de malestar físico moderado. Ceño ligeramente fruncido, ojos cerrados con fuerza moderada, nariz arrugada, boca cerrada firmemente con los labios apretados. La persona parece incómoda y molesta de forma constante durante todo el vídeo.*
- **Dolor severo:** *La persona mantiene una expresión facial altamente tensa. Ceño fuertemente fruncido, ojos intensamente cerrados, arrugamiento de la nariz, dientes fuertemente apretados, labio superior levantado ocasionalmente. La persona parece estar sufriendo un malestar físico severo de forma constante durante todo el vídeo.*

Se han evitado emplear palabras sensibles como “dolor”, eludiendo errores a la hora de generar el vídeo por no cumplir con las bases éticas de Veo3.1. Las descripciones se fundamentan en unas fotografías de ejemplo suministradas por la Escuela de Enfermería San Juan de Dios, y en las Action Units. Prkachin y Solomon describieron un subconjunto de AUs que se manifiesta de forma consistente ante el dolor, independientemente de la causa, la etnia, o la edad. Estas AUs son:

- **AU4**: Produce el conocido “ceño fruncido”. Es la AU más consistente y fiable como indicador de dolor.
- **AU6**: Eleva las mejillas y produce las arrugas conocidas como "patas de gallo".
- **AU7**: Tensa los párpados. Suele ir de la mano con la AU6.
- **AU9**: Arruga la nariz.
- **AU10**: Eleva el labio superior exponiendo los dientes superiores. Suele ir de la mano con la AU9. Se ha incluido solamente en la categoría de dolor severo y de forma ocasional, pues en las fotografías de ejemplo suministradas la boca abierta aparece solamente en el dolor severo. En el caso del dolor moderado, se ha descrito una mandíbula apretada en su lugar.

- **AU43**: Cierre completo de los párpados. En este caso, nos interesa que el paciente tenga los ojos cerrados independientemente del nivel de dolor, ya que en la gran mayoría de casos será así en los casos reales, tal y como lo demuestran las fotografías de ejemplo.

Además, Veo3.1 también admite consultas negativas, es decir, explicaciones de aquello que no se desea que aparezca en el vídeo. En todos los casos, la consulta negativa, traducida al castellano, fue:

Vista lateral, vista de perfil, rostro parcial, rostro recortado, equipo médico, movimiento de cámara, paneo, inclinación de cámara, zoom, seguimiento de cámara, anime, dibujo animado, ilustración, marca de agua, texto, logotipo, deformado, gafas, sombrero, maquillaje, joyas.

Para llegar a esta estructura, se ha hecho un proceso continuo de adaptación a las necesidades del modelo para que pueda capturar toda la información sin excepción. Como punto de partida, se realizó una consulta muy detallada siguiendo la estructura ordenada que proporciona Google: definición de la composición y de la cinematografía, identificación del sujeto principal, descripción de la acción que realiza el sujeto, descripción del entorno y los elementos de la escena, y la especificación de la iluminación y la ambientación. Esta versión inicial constaba de unas 250 palabras, una cantidad demasiado grande para la complejidad no tan elevada de la tarea. A pesar de ello, el resultado era bastante satisfactorio, pero quedaban cosas por corregir.

Aunque se dejaba claro en la consulta que se deseaba un plano totalmente estático y que mostrara la cara del paciente en su totalidad, la cámara a veces se movía y hacía zoom, o se situaba demasiado cerca del rostro. Por ello, se trató de reducir la extensión de la consulta, eliminando detalles que el modelo ya asumía a la perfección con el contexto. Como ejemplo, el fondo no necesitaba especificación adicional, ya que con nombrar que el sujeto estaba tumbado en una cama de hospital era suficiente. Por otro lado, se vio que era más importante describir un rasgo en concreto con claridad, antes que repetirlo varias veces a lo largo de la consulta para que el modelo no se olvidara de incluirlo. Por ejemplo, el hecho de mostrar el rostro entero sin recortar nada lo ejecutaba mejor añadiendo que se viera la almohada, antes

que remarcar que no se recortaran bajo ningún concepto ni las orejas, ni la frente, ni la barbilla. Tras estos retoques, se consiguió reducir la consulta a una longitud de entre 100 y 130 palabras, dentro del rango óptimo recomendado por Google, comprendido entre 100 y 200. No eran necesarias más palabras, ya que la secuencia no incluye muchos elementos, por lo que situarse en el rango bajo es suficiente.

Tras esto, era necesario afinar la descripción de la clase 1 (dolor moderado), pues al situarse en un punto intermedio entre la clase 0 y la clase 2, había que asegurar que no se acercara demasiado a ninguna de ellas, y que así se mantuviera fácilmente diferenciable. Aun así, era beneficioso que hubiera algo de variabilidad, cosa que Veo3.1 presumiblemente podía aportar. Mientras que la clase 0 era bastante monótona, las clases 1 y 2 eran ligeramente variables, representando distintos niveles de dolor, aunque casi siempre dentro de su propio rango. Esto es positivo, ya que el resultado buscado es que ante una cara totalmente neutra el modelo lo notifique como clase 0, ante un mínimo atisbo de incomodidad, la secuencia será catalogada como clase 1, y en caso de superarse determinado umbral de incomodidad, pasará a pertenecer a la clase 2. Queremos que solo se detecte que no hay dolor cuando sea muy obvio que no lo hay, y que ante la mínima sospecha de presencia de dolor se notifique. Además, se ha asumido que los pacientes, en caso de no sentir dolor, siempre mantendrán una expresión neutra, y que su estado no les permitirá hacer muecas de alegría, sorpresa, u otras sensaciones distintas al dolor.

Posteriormente, se añadieron varios rasgos aleatorios en cada muestra, para asegurar que cada sujeto fuera diferenciable. La etnia, la edad, el sexo y el grado de dolor son deterministas para asegurar el mismo número de muestras de cada categoría, lo que provocaba que Veo3.1 generara rostros muy parecidos para las mismas combinaciones. La solución propuesta añade elementos aleatorios a cada muestra generada, incluyendo 37 estilos de peinado distintos que varían en longitud, color y texturas, e incluyendo la ausencia de pelo, 6 formas de la cara (triangular, triangular inversa, circular, ovalada, cuadrada y alargada) y 15 detalles faciales que pueden incluir ligero vello facial, imperfecciones, o tamaño de la frente, labios, nariz, cejas o mejillas. Todas estas variaciones incluyen implícitamente rasgos como el peso, pues en algunos casos el modelo asume que una cara circular o unas mejillas llenas implican un peso elevado, lo cual no es negativo, ya que aporta

más variabilidad aún a las muestras. En cada caso se le dota al sujeto de un peinado concreto, una forma facial, y un solo detalle adicional, todos ellos al azar.

Para finalizar la consulta, se ha intentado aumentar la capacidad de generalización del modelo ante las situaciones cambiantes que se pueden dar en un hospital, con 7 tipos de iluminación distintos, comprendiendo desde una iluminación brillante de quirófano con una natural proveniente de una ventana. A su vez, aunque se asume que las cámaras empleadas serán de buena calidad, se han incluido 6 diferentes enfoques de cámara, intentando incluir muestras con alta capacidad de detección de detalles y otras con un enfoque más suave. También se han tenido en cuenta 8 diferentes texturas de la piel, representando rostros pálidos, enrojecidos, oleosos, etc.

Gracias a que Veo3.1 está especializado en cine, todas las especificaciones mencionadas (movimientos de la cámara, posición, iluminación, enfoque…) son captadas con facilidad por el modelo si son formuladas con un lenguaje técnico y claro.

Con la consulta ya construida en su totalidad, se procedía a realizar la solicitud de generación al modelo con toda la información adicional. Se seleccionó la versión Fast de Veo3.1, que prometía cumplir los mismos requisitos de calidad que la versión completa, con la diferencia de que era idónea para escenas con pocos elementos y nulo movimiento de cámara. Comparando la salida de ambas versiones se podía observar que la diferencia era mínima, por lo que la versión Fast resultaba la más conveniente, ya que ofrecía más capacidad de generación, al ser más veloz y tener un coste menor. Se fijo la resolución a 720p en horizontal (1280 píxeles en la dimensión horizontal y 720 en la vertical), una calidad de vídeo más que suficiente sabiendo que más tarde los fotogramas serían redimensionados más adelante a una resolución posiblemente menor. Por defecto, se representaron 24 fotogramas por cada segundo, una buena cantidad para cualquier investigación del campo, que podrían optar por usar todos, o también por emplear solo una fracción de estos. Se suprimió el audio y se impuso una duración de cuatro segundos. Al ser secuencias simples, no hacía falta más duración para representar transiciones capaces de expresar dolor inequívocamente.

En torno al 2% de las ocasiones, en lugar de generarse el vídeo saltaba un mensaje de error informando sobre la alta demanda que tenía Veo3.1 en ese momento. Como solución, las

consultas de estos casos eran guardadas para después reintentar su ejecución, pudiendo ser generadas en su práctica totalidad.

Las muestras se generaban en tandas de 108 (todas las combinaciones posibles de los rasgos deterministas: 2 sexos, 6 etnias, 3 rangos de edad y 3 grados de dolor) porque Google Colab tiene tiempos limitados antes de apagar el entorno de ejecución por inactividad que hacían imposible generar la base de datos en tandas más amplias con seguridad, y mucho menos generarla completa en una sola ejecución. Así pues, usar la misma semilla para cada ronda generaría los mismos números aleatorios usados para seleccionar los rasgos no deterministas. Esto quizás podría haber sido solucionado empleando una semilla distinta concreta para cada generación; sin embargo, el propio modelo de Veo3.1 crea secuencias distintas dada la misma consulta. En consecuencia, aunque parecidas, dos bases de datos generadas con las mismas semillas nunca habrían sido idénticas, por lo que finalmente se prescindió de usar semillas en esta parte del proyecto.

Esta variabilidad característica de las inteligencias artificiales generativas puede ser negativa en algunos casos concretos al aparecer elementos no deseados en la escena, pero en la mayoría de ocasiones fueron beneficiosas para este proyecto. Las muestras finales incluyen factores variables como camas y almohadas de distinto color, o cámaras situadas a distintas distancias y con perspectivas no siempre orientadas desde el mismo punto. La generalización del clasificador que posteriormente se entrenará depende de este tipo de cambios, lo que presenta una ventaja sobre las bases de datos no sintéticas, ya que estas requieren muchos recursos para crear un volumen de datos amplio con muestras suficientemente distintas entre sí.

Por último, la base de datos puede ser almacenada en Google Drive, Google Cloud, y ser descargada en disco local. La estructura de salida elegida ha sido una carpeta con 3 subcarpetas nombradas “0”, “1” y “2”, integrando cada una de ellas los vídeos de las clases 0, 1 y 2 respectivamente. Se han generado 1100 vídeos por clase, haciendo un total de 3300 en formato .mp4.

Capítulo 4. SISTEMA DESARROLLADO

4.1 PLANTEAMIENTO DEL PROBLEMA

Disponiendo de todos los vídeos generados, se desea entrenar un modelo que sea capaz de, ante una secuencia nueva, clasificar el nivel de dolor que se muestre en ella. Para ello, primero se deben procesar las muestras de la base de datos, para convertir los vídeos en un formato que pueda ser procesado por el modelo. Más tarde, será necesaria una arquitectura especializada en la extracción de rasgos faciales y que interprete correctamente la sucesión temporal de fotogramas.

Es fundamental que el modelo creado rinda con una precisión alta, pues en este contexto los errores de clasificación en el mundo real pueden acarrear consecuencias graves. Por otro lado, para poder llegar a tiempo ante un problema, el modelo debe ser capaz de procesar con baja latencia las nuevas muestras que se introduzcan, para así hacer posible la futura inferencia en tiempo real. También sería preferible que el entrenamiento esté optimizado en función de los recursos del entorno donde se despliegue e incluya el guardado de puntos de control, con el objetivo de invertir el menor tiempo posible y poder hacer pruebas fácilmente de cada una de las fases sin tener que entrenar el modelo en su totalidad.

4.2 DISEÑO DE LA SOLUCIÓN E IMPLEMENTACIÓN

El primer paso es el preprocesado. Como punto de partida, se extraen los fotogramas del vídeo con ayuda de la librería OpenCV, para después seleccionar solamente un número determinado de ellos. En un video con muchos fotogramas por segundo (en este caso 24) la diferencia de un fotograma a otro es mínima, por lo que conviene, y más en un contexto donde la velocidad de inferencia es crucial, sustraer fotogramas más espaciados entre sí. Por defecto, el modelo extrae 16 fotogramas a partir de un vídeo de 4 segundos, es decir, en torno a un fotograma cada 250 milisegundos. Este valor permite capturar correctamente los cambios espaciales a lo largo del tiempo y también asegura un buen rendimiento.

A continuación, con la librería Mediapipe se extraen 478 landmarks (puntos tridimensionales que construyen una red en un rostro) de cada uno de los fotogramas. A partir de las landmarks situadas en los extremos (izquierda, derecha, arriba y abajo), se recorta la cara y se redimensiona a una resolución estandarizada, en este caso a 224x224 píxeles. Gracias a esto, el modelo recibirá siempre muestras del mismo tamaño y donde aparezca solamente el rostro, eliminando la mayoría de elementos del fondo. En consecuencia, las landmarks, que en este momento están ubicadas en coordenadas relativas al fotograma completo, deben ser corregidas para que representen las coordenadas relativas al recorte de la cara realizado. En el caso en el que Mediapipe no detecte ninguna cara en el primer fotograma extraído, por simplicidad el vídeo correspondiente no será utilizado. Se ha comprobado que este suceso ocurre con una probabilidad menor al 1%, por lo que no se trata de una pérdida significativa. Por otro lado, en el caso de que no se detecte una cara en un fotograma distinto al primero, se utilizará el fotograma anterior en el que sí se había detectado la presencia de un rostro.

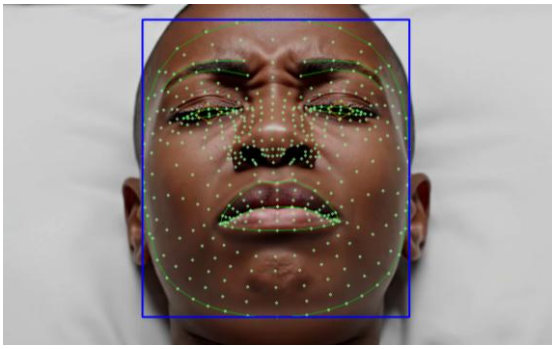

Ilustración 2. Landmarks de Mediapipe y recorte extraído

Ilustración 3. Fotogramas recortados y redimensionados

Una vez se dispone de un formato que almacene los datos de una manera que un modelo pueda procesar, se procede a construir la arquitectura. El primer bloque, centrado en detectar los fenómenos espaciales, consta de dos ramas paralelas. Por un lado, la rama visual utiliza EfficientNet-B0 preentrenado en ImageNet, una red convolucional profunda con una cantidad de parámetros (4.007.548) idónea para asegurar una precisión satisfactoria al mismo tiempo que se cumplen con los requisitos de eficiencia. Además, para el número de muestras del que se dispone, es menos probable que el modelo sobreajuste en comparación con otros modelos más pesados. Por otro lado, la cabeza clasificadora de EfficientNet no nos serviría de ninguna forma, ya que este módulo diferencia entre las 1000 categorías de Imagenet, y porque todavía no queremos clasificar. Lo que queremos es un vector de características previo a la clasificación que reúna toda la información extraída de un fotograma, que en este caso es de 1280 dimensiones.

La otra rama de este bloque (LandmarkEncoder) captura la geometría del rostro. Los 478 landmarks (1434 valores por tener tres dimensiones) pasan a través de un perceptrón de dos capas. La primera capa proyecta a 512 dimensiones, y va seguida de Batch Normalization, GELU y Dropout de 0,4. La segunda capa proyecta a 256 dimensiones, de nuevo seguida de Batch Normalization y GELU. Finalmente obtenemos un vector de 256 dimensiones que recoge la información geométrica del rostro de forma compacta. Esta rama tiene una ventaja considerable con respecto a la visual, y es que es invariante al color y la iluminación. El modelo aprende a interpretar la posición de cada una de las landmarks y sus relaciones entre sí, desarrollando por sí solo la habilidad de interpretar las Action Units. La gran variedad de muestras permite al modelo generalizar, detectando patrones de posiciones que conducen a la interpretación correcta.

El resultado final del bloque espacial, tras procesar paralelamente los 16 fotogramas seleccionados de cada vídeo, son 16 vectores de 1536 dimensiones (fruto de la concatenación del vector visual de 1280 dimensiones y el geométrico de 256 dimensiones). Así, por cada fotograma obtenemos una representación que combina el aprendizaje que las dos ramas han llevado a cabo independientemente.

Estos 16 vectores pasan por una LSTM unidireccional de 2 capas con 512 unidades ocultas y Dropout con 0,4 de probabilidad entre las capas (y no después de la segunda capa porque ya se hace Dropout en el clasificador) para reducir el sobreajuste. La intención es que la primera capa aprenda dinámicas temporales de bajo nivel, para que la segunda capa pueda detectar patrones de más alto nivel sobre estas dinámicas. Una capa solamente no captaría todo el contexto que se desea, mientras que más de dos capas podrían ser demasiadas para una secuencia de únicamente 16 vectores.

Tras esto, se extrae el último estado oculto. Aunque se espera que el último estado capte la información de toda la secuencia, tiene un sesgo a priorizar la información más reciente, lo cual puede ser positivo en este contexto, donde el objetivo final es la detección en tiempo real, y por tanto se debe priorizar lo más cercano al presente. Esta también es la razón por la que la LSTM es unidireccional, llevando la información de todos los vectores anteriores hacia el estado oculto correspondiente al último fotograma.

Dado que la base de datos no destaca por ser excesivamente extensa, y que se busca poder ejecutar el modelo en tiempo real, se han descartado otros modelos temporales más pesados como los Visual Transformers.

El vector que sale de la LSTM pasa por una cabeza clasificadora de tres capas. La primera proyecta de 512 a 256 dimensiones, seguida de Batch Normalization, GELU y Dropout de 0,4. La segunda capa pasa de 256 a 128 dimensiones, también con Batch Normalization y GELU, pero con Dropout de 0,2, ya que las capas finales necesitan más presición. La última capa proyecta linealmente a 3 dimensiones para producir los logits.

Todos estos módulos combinados otorgan al modelo interpretabilidad geométrica, espacial y temporal, para finalmente realizar una clasificación que pueda detectar el dolor.

Para comenzar el entrenamiento, primero debemos dividir las muestras en tres grupos: entrenamiento, validación y test. En este caso se ha optado por una distribución estándar del 70% para entrenamiento y un 15% para cada uno de los otros grupos. Es importante que esta división se realice de forma equilibrada entre clases, es decir, que estos porcentajes se apliquen en cada nivel de dolor, para que la proporción en cada uno de los grupos sea idéntica. Como tenemos el mismo número de vídeos para cada clase, la base de datos está

perfectamente equilibrada, por lo que cada muestra procesada tiene el mismo valor en el modelo y no hay necesidad de balancear.

Una vez construida la arquitectura del modelo y teniendo los datos preparados, se procede a comenzar el entrenamiento, que se realizará en dos fases bien diferenciadas. La primera fase se realiza con EfficientNet congelado, pues se trata de un modelo preentrenado, y en consecuencia con pesos bien aprendidos, a diferencia del LandmarkEncoder, la LSTM y el clasificador, que se inicializan con pesos aleatorios. Si no hiciéramos esto, los pesos arbitrarios producirían gradientes muy grandes, fruto de las malas predicciones propias del comienzo de los entrenamientos, y podrían destruir los pesos que sí son valiosos de EfficientNet. Las características extraídas por EfficientNet sabemos que son fiables, y el objetivo es que ayuden a que el resto de módulos aprendan desde el principio con información útil que los lleve a realizar clasificaciones razonables. Por otro lado, al congelar EfficientNet podemos computar los vectores de características que se obtienen al pasar todas las muestras por dicho backbone antes de comenzar el entrenamiento. Solo queremos que se entrenen el LandmarkEncoder, la LSTM y el clasificador, lo que hace que esta fase no requiera de mucho tiempo para completarse. Sin embargo, con este enfoque no se puede aplicar aumentación (con la excepción de Mixup, que se ha aplicado ya que funciona sobre los vectores de características) de datos en esta fase, pero al tratarse de una preparación de los pesos antes de comenzar el entrenamiento completo, no supone ningún impedimento.

En la segunda fase se descongela EfficientNet para entrenar el modelo en su totalidad, partiendo del mejor modelo obtenido en la fase anterior. Este backbone es un gran extractor de características, pero no se debe olvidar que fue entrenado con imágenes muy diversas, y no con rostros humanos expresando dolor. Por consiguiente, es necesario ajustar sus pesos para que se especialice en el dominio especifico que nos interesa, aprovechando toda su capacidad. Por tanto, en esta fase sí que podemos hacer aumentación de datos, lo cual es especialmente importante en este contexto, puesto que el fin es que el modelo aprenda a interpretar dolor en entornos reales, y no sintéticos, donde se pueden producir muestras con rasgos o patrones ocultos que pueden sesgar mucho el entrenamiento. Además, si se quiere implementar un modelo de estas características de forma estandarizada en la medicina, debe ser capaz de generalizar a cualquier tipo de calidad de imagen, enfoque óptico, o

iluminación. Aunque en la propia generación de la base de datos se haya tenido en cuenta este requisito, con aumentación ayudamos enormemente al modelo a generalizar, sin depender de la propia inteligencia artificial generativa. Por todos estos motivos, se ha optado por una aumentación muy agresiva tanto en el aspecto geométrico como el fotométrico.

La aumentación geométrica se ejecuta sobre los fotogramas y las landmarks, ya que deben guardar coherencia entre sí. Es preferible hacerlo de esta manera en lugar de aplicar la aumentación a los fotogramas y luego extraer landmarks del nuevo fotograma, porque para extraer el recorte del fotograma ya hemos computado las landmarks, haciendo preferible realizar la transformación sobre ellas. Las técnicas usadas incluyen el volteo horizontal el 50% de las ocasiones, rotación aleatoria entre 10 y -10 grados, y traslación aleatoria de máximo el 10% de la imagen tanto en el eje x como en el eje y.

La aumentación fotométrica se aplica solamente a los fotogramas. Se añade ruido gaussiano sutil para no degradar demasiado la información visual, variaciones aleatorias de brillo y contraste con probabilidad del 50%, oclusiones en posiciones aleatorias de un ancho y alto variables aplicado también en la mitad de las ocasiones, y desenfoque gaussiano con probabilidad del 30%, empleando un kernel de 3x3 e intensidades variables dentro de un rango bajo, garantizando que la bajada de calidad de la secuencia sea leve.

También se ha implementado Mixup, una técnica que ayuda a clasificar los casos más ambiguos. Este método debe aplicarse tras el resto de aumentaciones, ya que consiste en combinar dos muestras del lote para que el modelo construya fronteras de decisión suaves al observar casos que no pertenecen enteramente a una clase en concreto. Para el ojo humano, estas combinaciones no tienen sentido, pero en realidad resultan informativas al entrenar el modelo.

Prácticamente todas las aumentaciones descritas se aplican a una secuencia íntegra, es decir, los 16 fotogramas extraídos de cada vídeo recibirán las mismas transformaciones, aunque distintas a las de otras muestras del lote. La única excepción de esta norma es el ruido gaussiano, que actúa como regularizador al añadir ligeras variaciones en la textura para cada fotograma, y que así el modelo no sobreajuste a posibles patrones intrínsecos a las muestras sintéticas.

La función de perdida es la Entropía Cruzada, con el matiz de que, cuando se confunde la clase 0 con la clase 2 o viceversa, la penalización se duplica, pues se trata de un error más grave que confundir alguna de estas categorías con la clase 1.

El uso de semillas permite que, si se entrena el sistema con la misma base de datos, el modelo resultante será el mismo.

Capítulo 5. INFERENCIA EN TIEMPO REAL

5.1 PLANTEAMIENTO DEL PROBLEMA

Disponiendo de un modelo clasificador, se desea realizar inferencia en un contexto real. Esto se debe hacer bajo la suposición de una velocidad de cómputo suficiente por dicho modelo, para que así pueda procesar en tiempo real las secuencias que le lleguen a través de una cámara.

Para medir el rendimiento se debe comparar las predicciones generadas con las valoraciones humanas. El número de clases con las que estamos trabajando es muy reducido, por lo que es posible encontrar casos claros en cada una de las categorías, donde el modelo debería inferir sin mucha duda, al igual que los humanos, el nivel de dolor presente en la secuencia. Además, es necesaria la aprobación del personal sanitario confirmando que el prototipo funciona según se espera.

5.2 DISEÑO DE LA SOLUCIÓN

Dado que todo el entrenamiento se ha realizado con muestras de 16 fotogramas equiespaciados por 250ms, el enfoque en tiempo real ha sido el mismo. En consecuencia, cada 250ms se hará una predicción en base a los fotogramas extraídos durante los últimos 4 segundos, teniendo que realizarse la primera predicción pasado este tiempo para poder reunir suficientes fotogramas.

El camino que siguen los fotogramas para preprocesarse es el mismo que siguen las muestras de la base de datos antes de usarse para entrenar el modelo: Mediapipe detecta la cara, devolviendo 478 landmarks; con esas landmarks se hace el recorte de la cara, que se redimensiona; y se renormalizan los landmarks a las posiciones relativas de la nueva imagen recortada. Con dos buffers circulares (uno para las landmarks y otro para las imágenes recortadas y redimensionadas) se garantiza una predicción continua y con información

reciente, donde se elimina la información del fotograma más antiguo cada vez que uno nuevo se muestrea.

Recordando que el modelo tiende a dar prioridad al último fotograma, puede ocurrir que este de pie a una predicción errónea en un momento dado, cambiando la confianza en cada clase de manera brusca. Para reducir este efecto, se ha implementado la Media Móvil Exponencial (EMA), un método que pondera las probabilidades obtenidas en las predicciones anteriores junto con las actuales, usando un solo parámetro que mide si se da mucha o poca importancia a las predicciones pasadas. Se ha usado un valor de 0,35, lo que quiere decir las probabilidades actuales tienen el 35% de importancia. La importancia va decreciendo a medida que las predicciones son más antiguas, de forma que, en este caso, las predicciones del último segundo reúnen más del 82% de la importancia.

Por otro lado, para mantener consistencia en las predicciones ante el mismo tipo de expresiones se ha calibrado la inclinación a la que debe estar la cámara. Como los vídeos generados no muestran un plano perfectamente vertical en la mayoría de casos, tenemos que saber cuál ha sido la inclinación que el modelo ha aprendido a interpretar. Para calcular el punto de referencia se ha estudiado la distribución de dicha variable en la base de datos restando la profundidad (coordenada en z) de la barbilla y la profundidad de la frente, obteniendo los siguientes resultados:

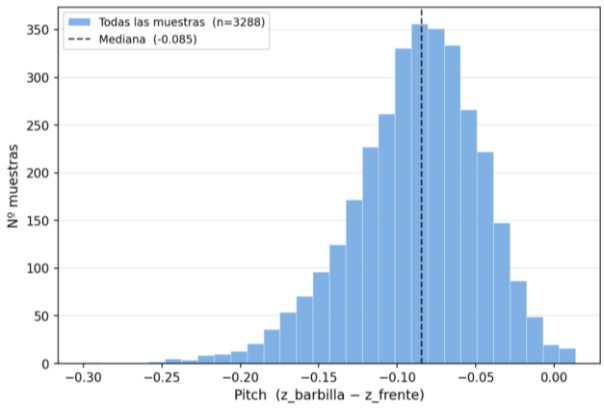

Ilustración 4. Inclinación media de cada vídeo

Como la profundidad tiene un valor más pequeño a medida que un punto se acerca a la cámara, sabemos que si la mayoría de muestras obtienen un resultado negativo en esta resta,

entonces en la mayoría de casos la cámara estará situada en una posición más cercana a la barbilla. Si se inspeccionan los vídeos generados se puede apreciar cómo, efectivamente, casi siempre esto es así. Se escoge la mediana, de valor -0,085, aunque la media también sería un buen enfoque al tratarse de una calibración orientativa con margen de desviación.

5.3 IMPLEMENTACIÓN

En este caso las pruebas se han realizado en un entorno local, pero la adaptación a cualquier cámara externa a la del dispositivo es fácilmente implementable. Para ofrecer al usuario información interpretable, la idea es mostrar en pantalla la grabación captada por la cámara, y sobre esa grabación se muestra la forma del recorte que se realiza a partir de las landmarks de Mediapipe, las probabilidades de cada clase, el nivel de dolor detectado en su color correspondiente (verde para ausencia de dolor, naranja para dolor moderado y rojo para dolor intenso), y una barra representando la desviación de la inclinación de la cámara con respecto al punto de referencia. Además, se incluyen los fotogramas por segundo que están siendo captados por la cámara, que son los que se proyectan por pantalla a pesar de que las predicciones solo se actualicen cada 250ms.

En el espacio tridimensional hay tres ejes, pero solo se ha tenido en cuenta, tal y como se ha explicado anteriormente, la inclinación de la cámara (eje y desde la perspectiva de la propia cámara). No se ha incluido calibración para la orientación de la cabeza (eje x), ya que se presupone una perspectiva centrada. Por otro lado, aunque sí se ha implementado la distancia a la cámara (eje z) como orientación, en realidad la extracción de la bounding box y el redimensionado decrementan el efecto de este eje.

Sin embargo, cabe recalcar que la distancia y la resolución sí que guardan entre sí una relación que puede llevar a problemas. Un rostro a la misma distancia de dos cámaras puede dar resultados algo distintos si la resolución de estos dispositivos es distinta. Si la resolución es baja, como es el caso de la webcam con la que se han realizado las pruebas (640 x 480 píxeles), es posible que se pierda información al redimensionar el recorte de la cara. En este caso, el recorte tiene una anchura de 224 píxeles, de forma que, a no ser que el paciente esté lo suficientemente cerca como para que el ancho de su cara ocupe más un tercio de la anchura

de la resolución (224/640=0,35), se perderá información que el modelo sería capaz de detectar. De la misma forma, si la resolución es muy alta, al redimensionar se interpolarán píxeles, que no siempre ofrece resultados satisfactorios. En consecuencia, la reproducibilidad de este apartado es parcial, puesto que con el mismo modelo se obtendrán resultados similares, que en parte variarán por la resolución de la cámara empleada. Es decir, si el modelo recibe la misma información en dos ocasiones, ofrecerá los mismos resultados.

Capítulo 6. RESULTADOS

6.1 VÍDEOS SINTÉTICOS

La base de datos ha recibido buena retroalimentación del personal sanitario de la Escuela de Enfermería San Juan de Dios, indicando que cumple con los requisitos de calidad y realismo con holgura, y que la perspectiva cenital de las muestras representa el 80% de los casos a los que se quiere atender. Dado que el paciente debe ser movido o girado, en algunas ocasiones dejará de mostrar el rostro a la cámara, casos en los que se tendría que atender a otros factores, como la frecuencia cardiaca u otros indicadores de la escala ESCID que se puedan continuar midiendo a través de otros sensores. Además, se ha puntualizado que la ventilación mecánica empleada actualmente en un contexto médico real no ocluye ninguna parte relevante para la detección del dolor, de forma que el enfoque actual de generación podría ser suficiente para generalizar sobre escenarios donde se utilicen. Por ejemplo, que la clase 2 muestre los dientes puede clasificar bien escenas en las que el paciente muerde el respirador, un indicador de nivel 2 en la escala ESCID en el apartado "Adaptación a la Ventilación Mecánica". A continuación, se presentan algunos fotogramas de muestra:

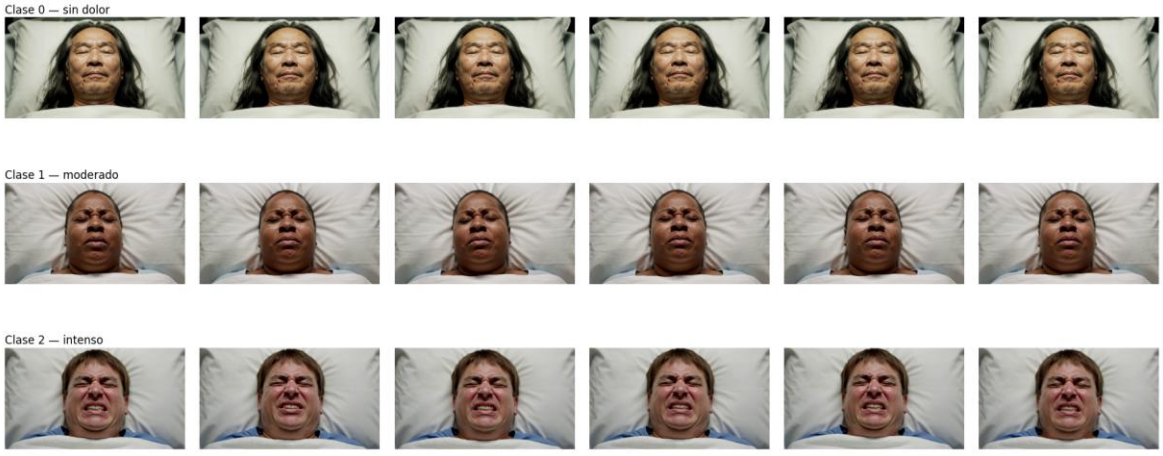

Ilustración 5. Fotogramas extraídos de la base de datos sintética

Como herramienta adicional de validación se ha empleado Openface 2.0, de libre acceso y bastante extendida en el campo. Openface mide todas las AUs de todos los fotogramas, aunque en vez de medir la AU43, mide la AU45. La diferencia es que la AU43 mide el cierre de los párpados, mientras que la AU45 se fija únicamente en los músculos relacionados con el parpadeo. Haciendo la media de todos los valores de las AUs de interés para cada vídeo, se obtienen los siguientes histogramas al evaluar en 300 vídeos aleatorios de cada clase:

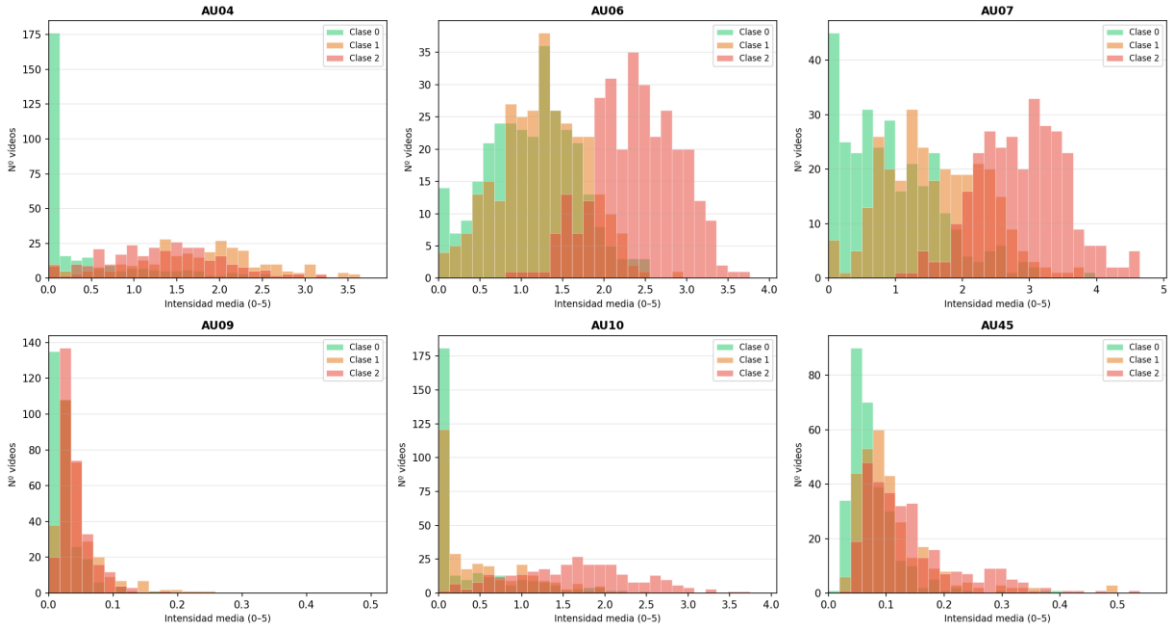

Ilustración 6. Valor de las AUs relacionadas con el dolor de los vídeos, diferenciados por clase

La AU4, que representa el ceño fruncido, toma valores altos para las clases 2 y 1, mientras que para la clase 0 la mayoría se sitúa en 0. Este comportamiento es el esperado, pues la escala ESCID especifica que, si hay dolor, debe haber ceño fruncido. La AU6, que se fija en la elevación de las mejillas, toma valores altos para todas las clases, aunque la clase 2 se diferencia de las otras dos. Es posible que Openface detecte elevación de mejillas en la clase 0 por la perspectiva ligeramente inclinada de algunos vídeos. La AU7, relacionada con la tensión en los párpados, es muy consistente, ofreciendo métricas bien diferenciadas para cada clase. Dado que solo se selecciona la AU con mayor presencia entre la AU6 y la AU7, siendo la AU7 la que tiene mayores valores generales para las clases 1 y 2, la diferenciación entre clases es aún más probable. La AU9, que representa el arrugamiento de la nariz, toma

valores muy bajos, lo que quizás indica una mala representación de Veo3.1 en este aspecto. Sin embargo, en la fórmula la AU9 no se tiene en cuenta si la AU10, activada por la exposición de los dientes, toma un valor mayor. Se puede observar que en el caso de la clase 2 siempre se muestran los dientes, mientras que en la clase 1 aparecen menos veces y que en la clase 0 no son visibles prácticamente en ningún momento. Por último, la AU45 toma valores muy bajos, debidos con bastante seguridad a que en realidad la AU que se debería activar es la AU43. No obstante, en la fórmula original del PSPI la AU43 es un valor discreto entre 0 (ojos abiertos) y 1 (ojos cerrados), y como en este caso todas las muestras tienen los ojos cerrados, no supondría ninguna diferenciación. Al aplicar la fórmula del PSPI sobre estas AUs, obtenemos:

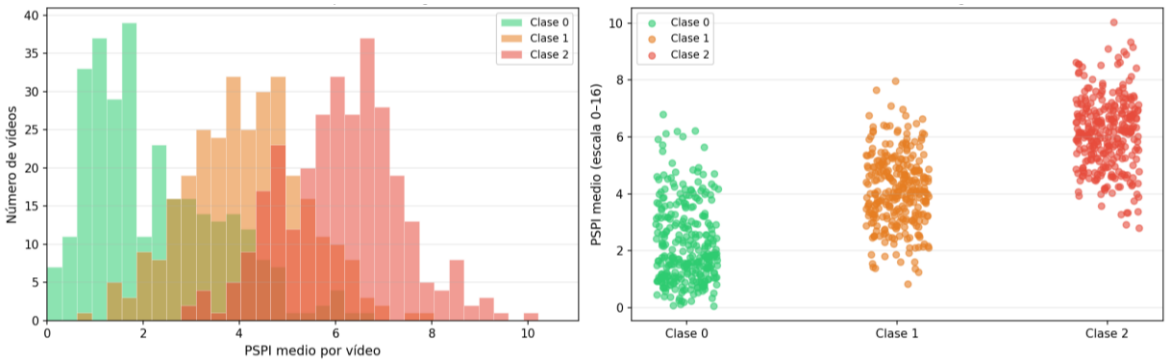

Ilustración 7. PSPI medio de los vídeos, diferenciando por clase

Las gráficas muestran bloques bien diferenciados por clase, lo que sugiere una exitosa representación del dolor por parte de Veo3.1.

6.2 *ENTRENAMIENTO DEL MODELO*

En cuanto al entrenamiento, estas han sido las métricas de aprendizaje:

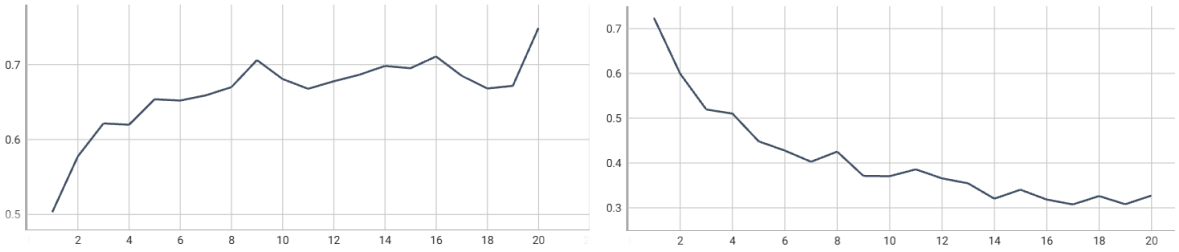

Ilustración 8. Precisión y pérdida en fase 1 de entrenamiento

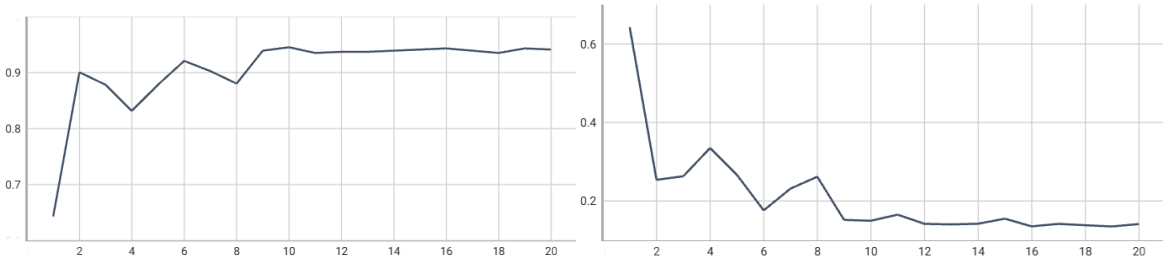

Ilustración 9. Precisión y pérdida de validación en fase 1

En la primera fase, con EfficientNet congelado, se realizaron 20 épocas con early stopping de 5, aunque no se activó. Se aprecia cómo la precisión del entrenamiento crece, aunque no toma valores muy altos por la influencia de MixUp, cuya función es mostrar al modelo casos ambiguos para obligarle a construir fronteras de decisión suaves. Por otro lado, la validación, donde no se realiza ningún tipo de aumentación, alcanza una precisión mayor al 90% rápidamente, mientras la pérdida decrece en consonancia, lo que presume la usencia de sobreajuste. Todo esto confirma que EfficientNet, aún sin hacer ajuste fino, captura suficiente información para distinguir niveles de dolor. Al mismo tiempo, el resto de módulos son capaces de adaptarse al problema y aprender pesos que realicen su función con éxito.

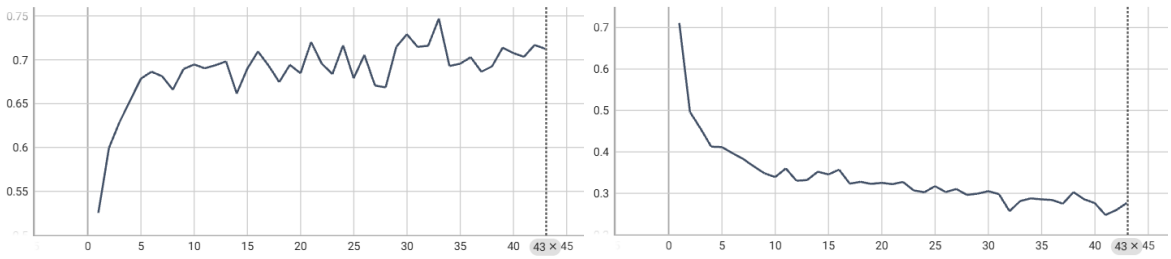

Ilustración 10. Precisión y pérdida en fase 2 de entrenamiento

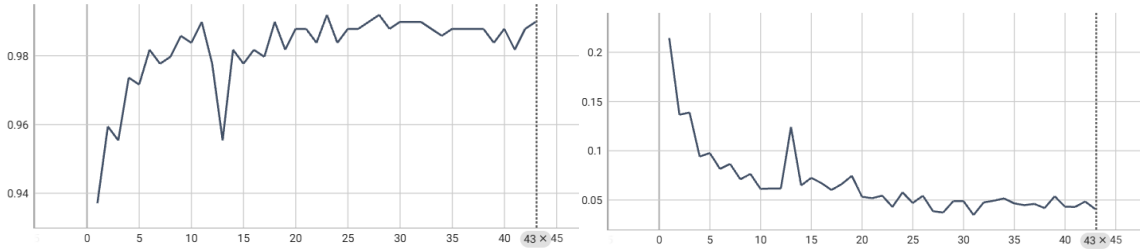

Ilustración 11. Precisión y pérdida de validación en fase 2

En la segunda fase EfficientNet se descongela, aunque se reduce el learning rate para ajustar sus pesos sin riesgo a destruir su aprendizaje. Con 50 épocas y un early stopping de 12, el entrenamiento acaba a las 43 épocas, es decir, obtiene su menor pérdida en validación en la época 31. En este caso se añaden muchas más técnicas de aumentación, lo que provoca que al principio el modelo tenga que adaptarse de nuevo a estas muestras tan distintas. Aún así, al igual que en la fase 1, tanto la precisión de entrenamiento como la de validación crece en pocas épocas al tiempo que sus respectivas pérdidas van bajando. En entrenamiento se mantiene, como es esperable, el efecto de Mixup, mientras que en validación se observa una notable mejoría de las métricas hasta una precisión superior al 99% y una pérdida inferior a 0,05.

Tabla 1. Métricas de validación

	Precision	*Recall*	*F1-Score*
Clase 0	0,994	0,992	0,993

Clase 1	0,988	0,987	0,988
Clase 2	1	0,994	0,997

Tabla 2. Métricas de test

	Precision	*Recall*	*F1-Score*
Clase 0	0,994	0,994	0,994
Clase 1	0,982	0,982	0,982
Clase 2	0,988	0,988	0,988

Conociendo la precisión que la validación alcanzaba durante el entrenamiento, no sorprenden unas métricas finales tan altas en este apartado. Por otro lado, el resultado del test nos reafirma que el sistema aprende correctamente. Todos los pacientes son distintos entre sí, y no existe ningún elemento común entre clases, de forma que el modelo se ha visto obligado a diferenciar las expresiones de dolor. Además, como sanity check, se ha intentado realizar el entrenamiento con muestras cambiadas arbitrariamente de clase, obteniendo como resultado unas métricas pésimas, que indican que el modelo sí está aprendiendo a diferenciar una clase de otra. Es importante reconocer que estos resultados no sorprenden tanto al estudiar el estado del arte: los modelos actuales aprenden a la perfección a adaptarse a la base de datos correspondiente, pero después suelen sufrir al enfrentarse a otros entornos, algo que se comprobará en la siguiente sección sobre inferencia.

6.3 *INFERENCIA EN TIEMPO REAL*

El personal sanitario ha indicado que el sistema es idóneo. El rendimiento y la interfaz de usuario cumple las expectativas marcadas, lo que abre camino a continuar con el desarrollo de una herramienta aún más sofisticada a partir de este prototipo. A continuación, se aportan algunos fotogramas representativos del funcionamiento del sistema:

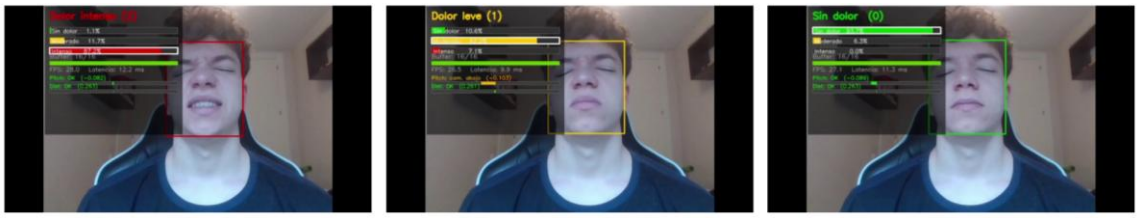

Ilustración 12. Muestra del sistema de inferencia en tiempo real

Cabe recalcar que el funcionamiento, aunque correcto, es limitado. Ante situaciones parecidas a las que incluye la base de datos responde mucho mejor que ante otras menos favorables. Una buena iluminación, orientación del paciente, y expresiones de dolor parecidas a las que Veo3.1 tiende a generar, son factores que elevan sustancialmente el potencial del modelo. También es importante remarcar que no se pretende detectar el dolor fingido o el dolor crónico, que necesitan otro tipo de técnicas y enfoques para ser interpretados correctamente.

Capítulo 7. CONCLUSIONES Y TRABAJOS FUTUROS

En primer lugar, se ha comprobado que la generación de vídeos sintéticos es una fuente de creación de datos válida, lo que podría suponer un avance mucho más acelerado en ámbitos donde la extracción de datos de calidad sea un problema. En la medicina, disponer de muestras sin límite práctico en cantidad, rápidas y sencillas de crear, y que cumplan las características concretas que se requieran, podría suponer la intrusión de la inteligencia artificial como elemento real de apoyo para una mejor atención al paciente, prevenir enfermedades, y diagnósticos precisos y eficientes.

Por otro lado, se ha observado la alta calidad de los modelos de visión por ordenador actuales. Con un hardware comercial al uso, se puede entrenar en pocas horas un sistema que clasifique con precisión secuencias de varios fotogramas de rostros sintiendo dolor, al igual que permite la inferencia en tiempo real con latencia imperceptible. Esto remarca un margen de maniobra para usar modelos más exigentes y pesados, junto con un mayor volumen de datos, lo que presumiblemente mejoraría aún más el desempeño de la solución actual. Además, el hecho de que el campo de la aplicación del aprendizaje profundo a la detección de dolor esté activo y con nuevas líneas de investigación resulta aún más prometedor, pues es el camino para que en un futuro próximo esta tecnología sea introducida en contextos médicos reales.

El proyecto, entonces, aporta varios elementos al sector:

- Una base de datos de 3300 vídeos de 4 segundos validada, diversa, equilibrada, y sin conflictos de privacidad.
- Una implementación capaz de generar más de 100 vídeos por hora, que incluso puede ser paralelizada.

- Un modelo híbrido que captura elementos espaciales, geométricos y temporales especializado en la clasificación de vídeos de rostros sintiendo dolor.
- Un sistema fácilmente adaptable preparado para mostrar en tiempo real las predicciones del modelo.
- Una estructura end-to-end modularizada, permitiendo la adición, eliminación o modificación de bloques, así como la transferencia de aprendizaje.

El resultado final ha resultado del agrado del personal sanitario de la Escuela de Enfermería San Juan de Dios, concertando una reunión con Ignacio Latorre, desarrollador de la escala ESCID, para abrir líneas de investigación que continúen con el proyecto hacia una versión cada vez más versátil y adaptada al mundo real. El principal punto para abordar es la detección de dolor en los momentos en los que no sea posible captar el rostro del paciente. Una idea podría ser continuar usando visión por ordenador para detectar otros indicadores de la escala ESCID, como la tranquilidad, el tono muscular, o la confortabilidad, y añadir otras variables como el ritmo cardiaco o el sonido.

En relación con la creación de muestras sintéticas, es necesario un equipo médico experto que desarrolle una serie de descripciones de expresiones de dolor. De este modo, se podrían realizar consultas detalladas y diversas a la inteligencia artificial generativa, y obtener una base de datos que permita la máxima generalización posible a los modelos de visión por ordenador. La colaboración directa entre el personal sanitario y los desarrolladores de los sistemas es fundamental para crear un círculo de retroalimentación que eleve la calidad de las muestras.

Dado que se ha observado un buen rendimiento con recursos de hardware limitados, es viable emplear modelos más pesados, como los Vision Transformers, lo que también sugiere la necesidad de un gran volumen de datos. No obstante, es preferible que las soluciones en este campo no ignoren la cuestión de la inferencia en tiempo real. El objetivo final es la implementación en los hospitales, donde una respuesta rápida es tan importante como una respuesta precisa.

Finalmente, como técnicas adicionales, se sugiere:

- Utilizar la funcionalidad de Veo3.1 que permite la creación de vídeos a partir de imágenes iniciales. Esto incluiría más complejidad al sistema, pero en caso de obtener imágenes que cumplan con los requisitos de consultas muy específicas, se podría ofrecer a Veo3.1 un punto de partida muy informativo para que solo se tenga que encargar de añadir la dinámica temporal.

- Añadir técnicas de explicabilidad que informen sobre los motivos que llevan al modelo a realizar una clasificación. La implementación en los hospitales requiere esta función para que la recepción del público sea positiva.

- Incluir técnicas de validación avanzadas para asegurar la precisión de los modelos. Dada la carencia de muestras reales, resulta fundamental el desarrollo de técnicas que aseguren la máxima generalización, como evaluar y/o entrenar el modelo en otras bases de datos rigurosas, sean sintéticas o no.

Capítulo 8. BIBLIOGRAFÍA

[1] M. Tavakolian and A. Hadid, "A Spatiotemporal Convolutional Neural Network for Automatic Pain Intensity Estimation from Facial Dynamics," International Journal of Computer Vision, vol. 127, pp. 1413–1425, 2019, doi: 10.1007/s11263-019-01191-3.

[2] Latorre Marco, M. Solís Muñoz, T. Falero Ruiz, A. Larrasquitu Sánchez, A. B. Romay Pérez y I. Millán Santos, "Validación de la Escala de Conductas Indicadoras de Dolor para valorar el dolor en pacientes críticos, no comunicativos y sometidos a ventilación mecánica: resultados del proyecto ESCID," Enfermería Intensiva, vol. 22, no. 1, pp. 3–12, 2011, doi: 10.1016/j.enfi.2010.09.005.

[3] P. Rodriguez, G. Cucurull, J. Gonzàlez, J. M. Gonfaus, K. Nasrollahi, T. B. Moeslund, and F. X. Roca, "Deep Pain: Exploiting Long Short-Term Memory Networks for Facial Expression Classification," IEEE Transactions on Cybernetics, vol. 52, no. 5, pp. 3314–3324, 2022, doi: 10.1109/TCYB.2017.2662199

[4] X. L. Lin, S. Mehraban, A. Moturu, and B. Taati, "Pain in 3D: Generating Controllable Synthetic Faces for Automated Pain Assessment," arXiv preprint arXiv:2509.06727, 2025.

[5] B. Taati, M. Muzammil, Y. Zarghami, A. Moturu, A. Kazerouni, H. Reimer, A. Mihailidis, and T. Hadjistavropoulos, "SynPAIN: A Synthetic Dataset of Pain and Non-Pain Facial Expressions," arXiv preprint arXiv:2507.19673, 2025.

[6] G. D. De Sario, C. R. Haider, K. C. Maita, R. A. Torres-Guzman, O. S. Emam, F. R. Avila, J. P. Garcia, S. Borna, C. J. McLeod, C. J. Bruce, R. E. Carter, and A. J. Forte, "Using AI to Detect Pain through Facial Expressions: A Review," Bioengineering, vol. 10, no. 5, p. 548, 2023, doi: 10.3390/bioengineering10050548.

[7] N. Ben Aoun, "A Review of Automatic Pain Assessment from Facial Information Using Machine Learning," Technologies, vol. 12, no. 6, p. 92, 2024, doi: 10.3390/technologies12060092.

[8] M. Cascella, D. Esposito, M. R. Muzio, V. Cascella y V. Cerrone, "Artificial intelligence for pain assessment via facial expression recognition (2015–2025): a systematic review," Exploration of Medicine, vol. 6, art. no. 1001370, 2025, doi: 10.37349/emed.2025.1001370.

[9] J. Huo, Y. Yu, W. Lin, A. Hu y C. Wu, "Application of AI in multilevel pain assessment using facial images: systematic review and meta-analysis," Journal of Medical Internet Research, vol. 26, e51250, Apr. 2024, doi: 10.2196/51250.

[10] H. El-Ghaish, M. Y. Al-Basiouny y M. A. M. Alshewimy, "Enhanced deep learning framework for real-time pain assessment using multi-modal fusion of facial features and video streams," Scientific Reports, vol. 15, art. no. 18970, 2025, doi: 10.1038/s41598-025-03362-5.

[11] C. W. Tan, T. Du, J. C. Teo, D. X. H. Chan, W. M. Kong y B. L. Sng, "Automated pain detection using facial expression in adult patients with a customized spatial temporal attention long short-term memory (STA-LSTM) network," Scientific Reports, vol. 14, art. no. 21778, 2024, doi: 10.1038/s41598-024-72427-4.

[12] P. Sankoh, A. Raza, K. Parwez, W. Shishah, A. Alharbi, M. Javed y M. Bilal, "Automated facial pain assessment using dual-attention CNN with clinically calibrated high-reliability and reproducibility framework," IEEE Access, vol. 12, pp. 118746–118760, 2024, doi: 10.1109/ACCESS.2024.3445808.

[13] T. Alghamdi y G. Alaghband, "Facial expressions based automatic pain assessment system," Healthcare, vol. 11, no. 3, art. no. 403, 2023, doi: 10.3390/healthcare11030403

[14] F. Alhamdoosh, P. Pala y S. Berretti, "Pain level estimation from videos by analyzing the dynamics of facial landmarks with a spatio-temporal graph neural network," IEEE Access, vol. 12, pp. 27877–27890, 2024, doi: 10.1109/ACCESS.2024.3367684.