Detección y cuantificación de emociones en población mexicana mediante el análisis de expresiones faciales y reconocimiento de voz

Francisco Emiliano Sánchez Callejas

2025

Universidad Autónoma de Querétaro

Facultad de Ingeniería

Detección y cuantificación de emociones en población mexicana mediante el análisis de expresiones faciales y reconocimiento de voz

Tesis

Que como parte de los requisitos para obtener el Grado de

Maestro en Ciencias (Mecatrónica)

Presenta

Ing. Francisco Emiliano Sánchez Callejas

Dirigido por:
Dr. Irving Armando Cruz Albarrán

Codirector:
Dr. Luis Alberto Morales Hernández

San Juan del Río, Qro. a 20 de enero de 2025

Universidad Autónoma de Querétaro
Facultad de Ingeniería
Maestría en Ciencias (Mecatrónica)

Detección y cuantificación de emociones en población mexicana mediante el análisis de expresiones faciales y reconocimiento de voz

Tesis

Que como parte de los requisitos para obtener el Grado de

Maestro en Ciencias
(Mecatrónica)

Presenta

Ing. Francisco Emiliano Sánchez Callejas

Dirigido por

Dr. Irving Armando Cruz Albarrán

Codirector

Dr. Luis Alberto Morales Hernández

Dr. Irving Armando Cruz Albarrán (15483)
Presidente

Dr. Luis Alberto Morales Hernández (6284)
Secretario

Dr. Emmanuel Reséndiz Ochoa (17415)
Vocal

Dr. Luis Morales Velázquez (6829)
Suplente

Dr. Carlos Andrés Pérez Ramírez (14209)
Suplente

San Juan del Río, Qro. México, 20 de enero de 2025

"Dreams save us. Dreams lift us up and transform us into something better."
- Superman.

Dedicatoria

A mi mamá, Ma. del Rosario Callejas Ríos, y a mi papá, Francisco Joel Sánchez Sánchez.
A mis hermanos Leonardo y Diego, y a mi hermana Helena. También a Valkiria, Frigga, Hera y Harley.
A las personas que me acompañaron durante este proceso, me apoyaron y me dieron aliento para iniciar, continuar y terminar.

A quien ha sido parte importante de mi vida.

Y en especial, a mi yo de 17 años que escogió una licenciatura pensando en hacer una maestría.

Agradecimientos

Agradezco a mis padres por la educación que tengo, tanto académica como personal, ya que es la base de mi persona.
A mis hermanos por escuchar sobre el proyecto, aunque no entendieran nada.
Agradezco a mis docentes por transmitirme los conocimientos necesarios que me permitieron profundizar en los temas necesarios para la realización de este proyecto.
Al Dr. Irving Armando Cruz Albarrán por su apoyo a lo largo de este proceso, por su amistad y sus consejos, por seguir creyendo en mí como lo hizo en la licenciatura y permitirme ser parte de su grupo de trabajo.
Al Dr. Luis Alberto Morales Hernández por su apoyo a lo largo de la licenciatura y la maestría. Así como por creer en mí y permitirme ser parte de su grupo de trabajo.
A mis sinodales, el Dr. Emmanuel, el Dr. Luis y el Dr. Carlos, por su tiempo y comentarios en la revisión de este proyecto.
A la Lic. Li Erandi y el Lic. Demian por su apoyo en la realización de las pruebas, así como a los participantes.
A mis amigos, quienes al pasar del tiempo hicieron de la maestría algo más sencillo.
A la Universidad Autónoma de Querétaro por permitirme ser parte de esta institución y a la Facultad de Ingeniería y la Dirección de Investigación y Posgrado por brindarme las herramientas necesarias para mi desarrollo profesional y académico.
Al CONAHCYT por la beca de manutención (CVU: 1276124) otorgada durante los estudios de posgrado.

Resumen

Las emociones son parte fundamental del ser humano y pueden transmitirse de diferentes maneras, por ejemplo, mediante expresiones faciales y el tono de voz. Esta información puede utilizarse para realizar un diagnóstico preciso de su estado emocional. En este contexto, se han desarrollado sistemas de reconocimiento de emociones utilizando métodos como los coeficientes cepstrales en la escala de Mel y la visión por computadora para la extracción de características y algoritmos como las redes neuronales convolucionales para la identificación tanto del rostro como de la voz. Los algoritmos utilizados en este tipo de sistemas de reconocimiento han demostrado su utilidad en la identificación de emociones al ser validados por medio del uso de bases de datos de expresiones faciales y voz, sin embargo, estas bases de datos suelen ser enfocadas en poblaciones extranjeras, por lo que las características de otra población, en este caso la mexicana, no son tomadas en cuenta. Además, los sistemas existentes únicamente detectan la emoción del individuo, sin embargo, no entregan un nivel emocional que permita al usuario identificar de manera correcta dicha emoción. El principal interés dentro de este trabajo de investigación es generar una herramienta que permita a profesionales de la salud mental satisfacer una necesidad para un mejor diagnóstico; por ello, se propone el desarrollo de un sistema embebido de clasificación y cuantificación de emociones en población mexicana adulta, mediante el análisis de expresiones faciales y reconocimiento de voz. El sistema se desarrolló utilizando coeficientes cepstrales en la escala de Mel y considerando los niveles de intensidad y herramientas de visión artificial para la extracción de características y redes neuronales convolucionales para la clasificación; y a través del entrenamiento del modelo basado en la concatenación de imágenes se logró obtener una precisión del 99.95 %, una pérdida de 0.02 %, mientras que, para la precisión en la validación, la exactitud, la sensibilidad y el valor F1 se obtuvo un 100 %, y la pérdida de la validación con un valor de 0 % para las emociones felicidad, tristeza y el estado neutral.

Palabras clave: expresiones faciales, tono de voz, cuantificación, clasificación, emociones, MFCC, CNN.

Abstract

Emotions constitute a fundamental aspect of the human condition and can be expressed in a variety of ways, including facial expressions and tone of voice. These forms of communication convey information about an individual's emotional state, which can be utilized to enhance diagnostic accuracy regarding an individual's mental health. In this context, emotion recognition systems have been developed using methods such as Mel's Frequency Cepstral Coefficients and computer vision for feature extraction, and algorithms such as convolutional neural networks for both face and voice identification. The algorithms utilized in this category of recognition systems have been validated through databases of facial expressions and voice, demonstrating their efficacy in identifying emotions. However, these databases are typically focused on foreign populations, with the characteristics of a specific population, in this case, the Mexican population, not being considered. Furthermore, the existing systems only detect the emotional state of an individual, lacking the capacity to provide an emotional level that allows the user to identify that emotion. The principal objective of this research project is to develop a tool that will assist mental health professionals in making more accurate diagnoses. To this end, the development of an embedded system for the classification and quantification of emotions in the adult Mexican population based on the analysis of facial expressions and voice recognition is proposed in this work. The system was developed using cepstral coefficients at the Mel scale, considering the intensity level and computer vision tools for feature extraction and convolutional neural networks for the classification and through model training based on image concatenation, resulting in an accuracy of 99.95 %, a loss of 0.02 % and 100 % validation accuracy, precision, sensitivity, and F1 value, with a loss of validation of 0 %. Additionally, the system incorporates a database focused on the adult Mexican population.

Keywords: facial expressions, voice tone, quantification, classification, emotions, MFCC, CNN.

Índice

Índice de tablas

Índice de figuras

Índice de tablas

Índice de figuras

1. Introducción

La comunicación del ser humano permite, entre otras cosas, dar un contexto del estado emocional. Este estado emocional puede transmitirse de manera verbal, es decir, por medio de la voz; y no verbal, a través de expresiones faciales y corporales. Tomando en cuenta esta información, existen sistemas de clasificación de emociones basados en algoritmos inteligentes, que, por medio de las características de la voz y las expresiones faciales, pueden dar información del estado emocional de una persona; lo que a su vez puede utilizarse.

Dichos sistemas suelen desarrollarse a partir de algoritmos como las máquinas de soporte vectorial, las redes neuronales convolucionales, las redes neuronales artificiales, el perceptrón multicapa, entre otros; los cuales permiten una clasificación de emociones a partir del entrenamiento de un modelo que utiliza una base de datos de imágenes o sonido.

Estas bases de datos suelen contener información de personas de origen asiático, europeo o estadounidense, por lo que la precisión de estos sistemas puede variar al utilizarlas en otro tipo de población. Además, los sistemas existentes únicamente detectan la emoción del individuo; sin embargo, no entregan un nivel emocional que permita al usuario identificar de manera correcta dicha emoción.

El principal interés dentro de este trabajo de investigación por el nivel emocional es con el objetivo de poder generar una herramienta que permita a profesionales de la salud mental satisfacer una necesidad para un mejor diagnóstico. Por ello, se propone el desarrollo de un sistema embebido de clasificación y cuantificación de emociones en población mexicana adulta, mediante el análisis de expresiones faciales y reconocimiento de voz. Para lograrlo se utilizan algoritmos inteligentes como las redes neuronales convolucionales, un algoritmo para la segmentación automática en imágenes y la caracterización de las señales de audio a partir de los coeficientes cepstrales en la escala de Mel.

1.1. Antecedentes

El ser humano es capaz de comunicarse tanto de manera verbal y no verbal, pudiendo dar a entender sus emociones por medio de expresiones faciales y corporales, además de la voz (Brener et al. 2023). Las emociones desempeñan un papel crucial en la vida diaria de los seres humanos, ya que, como menciona Takahashi (2004), son fundamentales en la comunicación dentro de las interacciones sociales. Estas emociones se clasifican en dos principales tipos: las emociones primarias y las emociones secundarias. Las emociones primarias, tales como la ira, miedo, alegría, tristeza, asco y sorpresa, están presentes desde el nacimiento. Por otro lado, las emociones secundarias resultan de combinaciones de las primarias, como la vergüenza, la culpa, el orgullo, el placer, los celos, entre otras. A su vez, ambos tipos de emociones se pueden clasificar en positivas y negativas dependiendo el sentimiento con el que se relacione; en el caso de las positivas se relacionan con sentimientos agradables, mientras que las negativas con sentimientos desagradables (Takahashi, 2004).

YYa que las emociones se pueden transmitir de forma corporal y verbal, existen investigaciones concentradas en poder detectarlas a través de expresiones faciales o reconocimiento de la voz. Puesto que la cara es la parte más expresiva al transmitir emociones de manera no

verbal, se han desarrollado diferentes estudios para el reconocimiento de expresiones faciales (*Facial Expression Recognition*, FER) enfocadas en esta zona, permitiendo identificar y conocer las variables que influyen en este tipo de sistemas, a modo de generar un resultado preciso. Zhao y Zhang (2016) indican que el FER permite de manera eficiente y precisa reconocer el estado emocional del ser humano a través de sus expresiones.

Además, Revina y Emmanuel (2021) mencionan que la cara ofrece tres tipos de señales de interés: estáticas, lentas y rápidas; y son elementos esenciales en el reconocimiento de expresiones faciales. Las señales estáticas incluyen atributos como el tono de piel, la forma de la cara y el tamaño de los ojos. Dentro de las señales lentas se consideran cuestiones como las arrugas, mientras que las señales rápidas se consideran aspectos como el movimiento de la boca o las cejas, debido a que pueden cambiar constantemente y por lo tanto son especialmente influyentes en la expresión de emociones.

Para generar sistemas de FER, Zhao y Zhang (2016) indican que la extracción de características para el reconocimiento facial puede hacerse tanto para imágenes estáticas, como para imágenes dinámicas. Por su parte, Revina y Emmanuel (2021) sugieren emplear el preprocesamiento de imágenes, para ajustar los atributos de éstas, de modo que el reconocimiento se haga de manera correcta.

En la literatura se han utilizado diversos algoritmos para generar sistemas de clasificación de emociones a través de expresiones faciales. Uno de ellos es el de Apatean et al. (2009) quienes demostraron que algoritmos como las máquinas de soporte vectorial (*Support Vector Machine*, SVM) obtienen resultados de entre el 85.2 % al 90.3 % de precisión. En contraste, métodos como el k-ésimo vecino más próximo (*K-Nearest Neighbors Algorithm*, KNN) su precisión ronda entre el 84 %. Otro trabajo centrado en los sistemas FER es el presentado por Zhao y Zhang (2016) donde comparan diferentes métodos, como las redes neuronales artificiales (*Artificial Neural Network*, ANN), los modelos ocultos de Markov (*Hidden Markov Models*, HMM) y las SVM, entre otros, obteniendo una precisión de 60.09 %, 78.64 % y 79.88 %, respectivamente.

Una investigación adicional que expone el uso de redes neuronales convolucionales para identificar emociones es el que presentan Bhagat et al. (2024), donde utilizan la base de datos FER-2013 (*Facial Expression Recognition 2013 Dataset*), y a través de una DCNN (*Deep Convolutional Neural Network*) y modelos preentrenados como *EfficientNet, ResNet, VGGNet* y un clasificador facial basado en Haar Cascade obtienen una precisión del 82 %.

En el ámbito del reconocimiento de emociones a través de la voz (*Speech Emotion Recognition*, SER), El Ayadi et al. (2011) afirman su utilidad en cuestiones de seguridad, entretenimiento, traducción, atención al cliente y como herramienta de diagnóstico terapéutico, entre otros. Sin embargo, Ferreiros et al. (1998) destacaron que, los principales problemas para hacer reconocimiento por voz son cuestiones como la pronunciación y el tono, ya que pueden variar, aunque se trate del mismo individuo. A causa de esto se han generado múltiples investigaciones donde se utilizan diferentes tipos de algoritmos para el análisis y clasificación de las muestras de sonido. Por ejemplo, García Guajardo (2011) presenta el uso de coeficientes cepstrales en las escala de Mel (*Mel-Frequency Cepstral Coefficients*, MFCC) para el reconocimiento de voz, donde se validaron sus resultados por medio de una ANN. Por otro lado, Datta Rakshith et al. (2021) utilizaron métodos como los MFCC para la extracción de las

características de la voz, y el aprendizaje por cuantificación de vectores (*Vector Quantization*, VQ) para la clasificación; donde obtuvieron un desempeño promedio del 93.3 %

Además, la literatura incluye métodos como los HMM utilizados por Coto-Jiménez et al. (2014) donde realizan un análisis basado en el español mexicano, con el objetivo de comparar los parámetros de tono y fluctuación de las vocales. Por otra parte Matveev et al. (2022), compararon el uso de un perceptrón multicapa (*Multi-layer Perceptron*, MLP) y una SVM, para realizar la clasificación de emociones de niños hablantes del idioma ruso; este trabajo muestra una precisión del 84.6 % para la SVM, mientras que un 83.3 % para el MLP.

Existen diferentes bases de datos para investigaciones basadas en FER, sin embargo, la mayoría de estas bases de datos son enfocadas en poblaciones europeas y estadounidenses (Li et al. 2022). Una de las bases de datos más utilizadas es la *Japanese Female Facial Expression* (JAFFE), que presentan Lyons et al. (1998), la cual cuenta con un total de 219 imágenes divididas en seis expresiones faciales emocionales y una neutral. Por otro lado, el trabajo de Lucey et al. (2010), describe la *Extended Cohn-Kanade Dataset* (CK+), donde se cuenta con siete emociones para clasificar a través de 593 imágenes en escala de gris y a color. Por otra parte, la base de datos de imágenes FER-2013 (*Facial Expression Recognition 2013 Dataset*), utilizada en los trabajos que presentan Saurav et al. (2021) y Shi et al. (2021), es una base de datos de uso libre que contiene las categorías enojo, disgusto, miedo, felicidad, tristeza, sorpresa y neutral; a partir de imágenes de 48x48 pixeles en escala de gris.

Con respecto a los sistemas de SER, existen múltiples bases de datos, descritas en la investigación de Abbaschian et al. (2021), como la *Berlin Database of Emotional Speech* (EMO-DB) para el idioma alemán, y contiene 700 muestras de audio; o la *Danish Emotional Speech Database* (DES), desarrollada para identificar 5 emociones totales, basada en el idioma danés.

Por otra parte, en el trabajo presentado por Pan et al. (2024),se utilizan diferentes modelos de detección de audio y texto para validar una base de datos denominada *Spanish MEACorpus 2023*, que contiene 13.6 horas de audio obtenidas a partir de videos de YouTube. Sin embargo, al hablar del español de México, Duville et al. (2021) presentan la base de datos con nombre *Mexican Emotional Speech Database* (MESD), la cual contiene 864 grabaciones de voz, para enojo, desagrado, miedo, felicidad, tristeza y neutral; además de tres categorías importantes: mujer adulta, hombre adulto y niños, y se validó por medio de una SVM.

Respecto a la cuantificación emocional se han desarrollado modelos para investigaciones enfocadas en la conversión emocional de la voz (*Emotional Voice Conversion*, EVC), las cuales son técnicas de procesamiento de voz para convertir una voz neutral a una emoción especifica. Este tipo de técnicas se basan en modificar las características acústicas, como la frecuencia, la intensidad y el timbre, para reflejar la emoción deseada (Zhou et al. 2022). Dentro del uso de la intensidad como referencia para el estado emocional se encuentra EmoVox, el cual es un sistema que ajusta la intensidad de la emoción en función de las características del audio como los MFCC, lo que permite cuantificar los niveles emocionales del enojo, felicidad y el estado neutral a través de texto y audio (Zhou et al. 2023).

Como se ha expuesto, existen trabajos por separado para el SER y el FER, sin embargo, se han desarrollado investigaciones que las conjuntan, como por ejemplo, la de Wang et al. (2013), quienes generaron sistemas por separado a través de la utilización de un modelo

de mezcla gaussiana (*Gaussian Mixture Model*, GMM); donde para el FER se obtuvo una precisión del 40 %, mientras que para el SER se logró una precisión del 88.8 %. Sin embargo, al combinarlos el resultado fue una precisión del 90.5 %. Mientras que, Ristea et al. (2019) demostraron que al combinar los métodos de reconocimiento, se consiguió una precisión del 69.42 % al utilizar una Red Neuronal Convolucional (*Convolutional Neural Network*, CNN).

EEs importante mencionar que los trabajos descritos hasta el momento han logrado clasificar emociones, ya sea a través de un sistema FER, un sistema SER, o la combinación de estos. Esta clasificación a través de algoritmos inteligentes demuestra que es posible generar un algoritmo que sea capaz de detectar emociones, permitiendo que, las técnicas de procesamiento de información obtengan las características necesarias para los sistemas de clasificación, sin embargo, en el desarrollo de este trabajo de investigación no se encontró información dentro de la literatura por la cuantificación del nivel emocional de los individuos de población mexicana, así como del desarrollo de algún sistema que unifique hardware y software en un sistema embebido. Por ello, este trabajo de investigación se enfoca en el desarrollo de un sistema para la clasificación de emociones y la cuantificación de su nivel a través de expresiones faciales y voz, utilizando algoritmos de inteligencia artificial; lo cual brindará una herramienta complementaria a los profesionales de la salud mental.

1.2. Objetivos

1.2.1. Objetivo general

Detectar y cuantificar las emociones de felicidad, tristeza y el estado neutral en individuos de la población mexicana adulta mediante el análisis de expresiones faciales y voz, utilizando técnicas de procesamiento de imágenes, análisis de señales y Redes Neuronales Convolucionales, a partir de la implementación y calibración de los modelos en un sistema embebido.

1.2.2. Objetivos particulares

- Implementar un modelo para la detección de expresiones faciales mediante visión artificial y redes neuronales convolucionales, capaz de clasificar las emociones de felicidad, tristeza y neutral.
- Implementar un modelo de reconocimiento de emociones, tales como felicidad, tristeza y neutral, mediante el uso de señales de voz y, haciendo uso de algoritmos como MFCC y redes neuronales convolucionales, enfocado en bases de datos de población mexicana.
- Fusionar los sistemas de detección y clasificación de expresiones faciales y de reconocimiento de emociones a través de la voz en un modelo funcional.
- Calibrar y reentrenar el sistema de detección y clasificación de emociones de felicidad, tristeza y neutral por medio de las características de individuos de población mexicana.

- Desarrollar un sistema embebido, basado en el uso de un GPU, que conjunte los elementos necesarios para visualizar la detección y cuantificación de los estados tristeza, felicidad y neutral, mediante el uso de expresiones faciales y voz.

1.3. Hipótesis

El uso de técnicas de procesamiento de imágenes, análisis de señales y redes neuronales convolucionales aplicadas al reconocimiento de expresiones faciales y voz permitirá detectar y cuantificar de manera precisa los estados emocionales de felicidad, tristeza y neutral. Esta precisión se logra a través de un sistema embebido que incorpora la calibración y reentrenamiento de los modelos de detección, utilizando datos de audio y video provenientes de individuos de población mexicana.

1.4. Planteamiento del Problema

Ya que las emociones son una parte fundamental de la psicología humana, es importante saber identificarlas de manera correcta, sobre todo cuando se trata de emociones negativas. Por ello, en 2017, el Instituto Nacional de Estadística y Geografía (INEGI), a través de la Encuesta Nacional de los Hogares (ENH), realizó la consulta en la República Mexicana para conocer el número de integrantes del hogar de 12 años y más que se ha sentido deprimido (INEGI, 2017); obteniendo los resultados que se muestran en la Figura 1.

Entidad federativa	Sexo	2014			
		Total	Se han sentido preocupados o nerviosos	Nunca se han sentido preocupados o nerviosos	No saben si se han sentido preocupados o nerviosos
Estados Unidos Mexicanos	Total	100.0	51.1	48.4	0.5
	Hombres	100.0	57.7	42.0	0.3
	Mujeres	100.0	49.6	50.1	0.4
Entidad federativa	Sexo	2015			
		Total	Se han sentido preocupados o nerviosos	Nunca se han sentido preocupados o nerviosos	No saben si se han sentido preocupados o nerviosos
Estados Unidos Mexicanos	Total	100.0	46.2	53.3	0.4
	Hombres	100.0	52.7	47.0	0.3
	Mujeres	100.0	51.8	47.6	0.6
Entidad federativa	Sexo	2017			
		Total	Se han sentido preocupados o nerviosos	Nunca se han sentido preocupados o nerviosos	No saben si se han sentido preocupados o nerviosos
Estados Unidos Mexicanos	Total	100.0	47.8	51.5	0.7
	Hombres	100.0	55.5	44.0	0.5
	Mujeres	100.0	30.5	69.4	0.1

Figura 1: Encuesta Nacional de los Hogares (INEGI, 2017).

Estos resultados indican que existe un 48.33 % de la población mexicana, en promedio, que ha identificado este tipo de emociones; sin embargo, también existe un 0.37 % que no puede identificarlo.

Los sistemas actuales de FER y SER están enfocados en su mayoría a población europea, asiática o estadounidense, por lo que, al utilizarlos en otro tipo de población, los resultados pueden verse sesgados. Si bien los sistemas FER y SER son capaces de clasificar las emociones, la carga computacional tiende a ser alta, impidiendo en la mayoría de estos un procesamiento en línea, por la complejidad de los algoritmos de inteligencia artificial que se utilizan. En consecuencia, es difícil implementar dichos sistemas en un área donde tengan una aplicación real.

Por otra parte, dichos sistemas utilizan algoritmos inteligentes para clasificar las emociones con un nivel de precisión alto; sin embargo, no permiten cuantificar el nivel emocional al que se encuentra el sujeto. Además, los sistemas de clasificación actuales tienen un enfoque general, por lo que no consideran ciertas características de los individuos, las cuales pueden afectar el resultado.

Es importante mencionar que una de las principales razones para desarrollar este proyecto de investigación es dar una solución a una necesidad que tienen los profesionales en el área de la salud mental, ya que, en su proceso de diagnóstico para consultas de atención psicológica, muchos de los individuos o pacientes no suelen identificar de manera correcta sus emociones, por lo que la información puede ser imprecisa y, por lo tanto, afectar de manera negativa en el diagnóstico. Derivado de lo mencionado anteriormente, surge la siguiente pregunta de investigación: ¿Será posible utilizar algoritmos inteligentes, como las Redes Neuronales Convolucionales, para clasificar y cuantificar emociones en población mexicana adulta a través de un sistema embebido?

1.5. Justificación

En este trabajo se plantea el desarrollo de un sistema embebido capaz de clasificar y cuantificar el nivel emocional de felicidad, tristeza y neutral en individuos de población mexicana adulta, a través de reconocimiento de expresiones faciales y señales de voz, basándose en redes neuronales convolucionales para obtener un resultado en línea que considere las características del individuo al realizar una calibración previa. Todo ello como solución a una necesidad dada por profesionales de la salud mental, los cuales requieren este tipo de herramienta para dar un diagnóstico más preciso.

El sistema que se desarrolla es no invasivo, inocuo y amigable con el usuario. Además, permite clasificar y cuantificar, a través de una escala, el nivel de emoción, con la finalidad de que pueda ser utilizado en protocolos y consultas realizadas por profesionales en el área de la salud mental, permitiendo que les brinde información adicional sobre el estado del individuo. La implementación del sistema se desarrolla teniendo en cuenta la carga computacional para permitir el análisis de la información. La cuantificación del estado emocional de los individuos es una característica innovadora debido a que, en el desarrollo de este trabajo de investigación no se han encontrado sistemas FER y SER, capaces de generar una cuantificación, sino que, únicamente clasifican las emociones. Así mismo, el uso de metodologías y algoritmos basados

en la inteligencia artificial permite encontrar nuevas soluciones o alternativas a diferentes problemas, como es el caso de este proyecto de investigación.

Por otra parte, con el enfoque en la población mexicana adulta se pretende utilizar una base de datos de voz con el fin de cumplir con el objetivo de este trabajo de investigación. Con esto se potencializarán los beneficios, ya que estará enfocada directamente a la población que se pretende tomar en cuenta.

Este proyecto de investigación se basa en las necesidades planteadas en los Programas Nacionales Estratégicos del CONAHCYT (PRONACES), específicamente en el Proyecto Nacional de Investigación e Incidencia (PRONAII) Ciencia de Datos y Salud, donde su objetivo es utilizar la información y las tecnologías de la ciencia de datos, la inteligencia artificial, en conjunto con la medicina, las ciencias sociales, la biomédica, entre otras; para obtener indicadores en materia de salud. Además del PRONAII Salud Mental y Adicciones, ya que el desarrollo de este proyecto de investigación permitirá a los profesionales de la salud mental dar un mejor diagnóstico, lo que contribuye a la prevención y atención de la salud mental, el cual es uno de los objetivos del PRONAII.

Es importante mencionar que el desarrollo de la presente investigación es viable, ya que se cuenta con los conocimientos necesarios en el área de ingeniería, así mismo, se cuenta con colaboración con profesionales del área psicológica, lo cual permitirá validar los resultados obtenidos. De igual forma, se cuenta tanto con la tecnología como con los componentes electrónicos necesarios para el sistema embebido, así como para el diseño e impresión de carcasa.

1.6. Ética del estudio

A continuación, se muestran las leyes mexicanas vigentes a la fecha que se consideran para el desarrollo de este trabajo de investigación:

1.6.1. Ley Federal de Protección de Datos Personales en Posesión de los Particulares

Esta ley tiene como finalidad regular el uso de la información personal, para garantizar la privacidad y el derecho a la autodeterminación informativa de las personas, de una forma legítima, informada y controlada. La ley menciona que se debe informar al titular de la información y obtener su consentimiento para el uso de los datos, además de garantizar su derecho al acceso, la rectificación, la cancelación y la oposición, para el uso de su información (D.O.F., 2014).

Además, se consideran los principios éticos del Informe de Belmont. Referente a la investigación médica en seres humanos, se considera la declaración de Helsinki y los puntos básicos del Código de Núremberg.

Debido a esto, se extenderá a los participantes una Carta de Consentimiento Informado (Anexo 7.1) y una Carta de Confidencialidad de Datos (Anexo 7.2), donde se informará la finalidad de las pruebas y recolección de los datos, así como el resguardo de dicha información.

1.7. Normas y estándares industriales

A continuación se describe la normativa que se consideró para el desarrollo de este proyecto, la cual está enfocada en la fabricación de dispositivos médicos y su uso en investigación.

1.7.1. NORMA Oficial Mexicana NOM-241-SSA1-2021, Buenas prácticas de fabricación de dispositivos médicos

Esta normativa aborda las buenas prácticas en la manufactura de dispositivos médicos. Su propósito es establecer los requisitos mínimos que deben cumplirse durante el diseño, producción, almacenamiento y distribución de estos dispositivos, considerando su nivel de riesgo. El objetivo principal es garantizar que los dispositivos cumplan con estándares de calidad, seguridad y funcionalidad, permitiendo un uso seguro por parte de los pacientes o consumidores finales. Su cumplimiento es obligatorio en todo México y aplica tanto a los establecimientos que fabrican dispositivos médicos como a los almacenes de acondicionamiento, depósito y distribución (D.O.F., 2021).

Conforme a la norma, en el desarrollo de este proyecto no se expone a ningún riesgo a los sujetos al considerar métodos no invasivos e inocuos.

2. Fundamentación Teórica

Para el desarrollo de este proyecto de investigación se tienen en cuenta diferentes cuestiones teóricas que permitirán cumplir los objetivos que se plantean. Debido a que se desarrollan sistemas basados en algoritmos inteligentes a través del procesamiento de señales de audio y el procesamiento de imágenes, es importante contar con las herramientas necesarias.

Al hablar del procesamiento de imágenes, automáticamente se habla de la visión por computadora, siendo una parte importante de este proyecto; ya que, sus herramientas permitirán identificar las características necesarias de las imágenes. Además, el uso de redes neuronales convolucionales se suele hacer con información de imágenes para la clasificación.

Por otra parte, para el procesamiento de las señales de audio se utilizarán herramientas como los MFCC, ya que que permiten obtener datos e información de la señal basándose en la percepción humana y son muy utilizados en el campo de reconocimiento de voz.

Cada uno de estos conceptos que serán utilizados a lo largo del desarrollo de este proyecto de investigación se explican a mayor detalle a continuación.

2.1. Emociones

Al ser una parte fundamental de los seres humanos, las emociones han desempeñado un papel importante en investigaciones relacionadas con la psicología. Una de estas investigaciones define la Teoría de las emociones básicas de Ekman (Pan et al. 2024), donde mostró imágenes de expresiones faciales a individuos de diferentes culturas y evaluó si podían identificar dichas emociones. Las emociones que define Ekman son la felicidad, la ira, la sorpresa, la tristeza, el asco y el miedo (Figura 2); sin embargo, esta investigación únicamente se enfoca en la felicidad y la tristeza, así como en el estado neutral. Dando como resultado que, independientemente del entorno cultural, estas emociones pueden ser identificadas por los individuos; por ello, Ekman las define como emociones básicas universales (Ekman y Friesen, 1971).

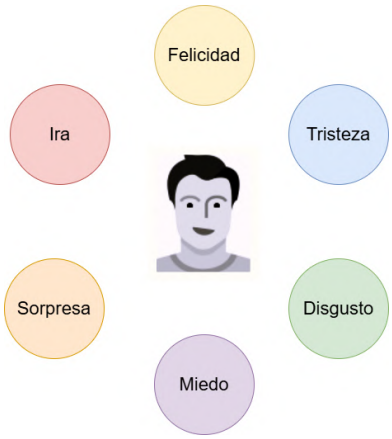

Figura 2: Emociones básicas universales (Autoría propia, 2024).

2.2. Visión por Computadora

La visión por computadora o visión artificial es el conjunto de herramientas que permiten adquirir, procesar y analizar imágenes o vídeos, con el fin de obtener información de éstas, tal como lo hace la visión humana a través de la vista (Figura 3). Dentro del procesamiento de las imágenes se utilizan herramientas para mejorar la calidad, reducir el ruido, identificar características, entre otras (Szeliski, 2011). Por otra parte, en la caracterización de las imágenes se considera el color y la forma. Finalmente, el análisis de la imagen permite utilizar técnicas de clasificación, detección de objetos y reconocimiento; todo a través de la información adquirida de la imagen.

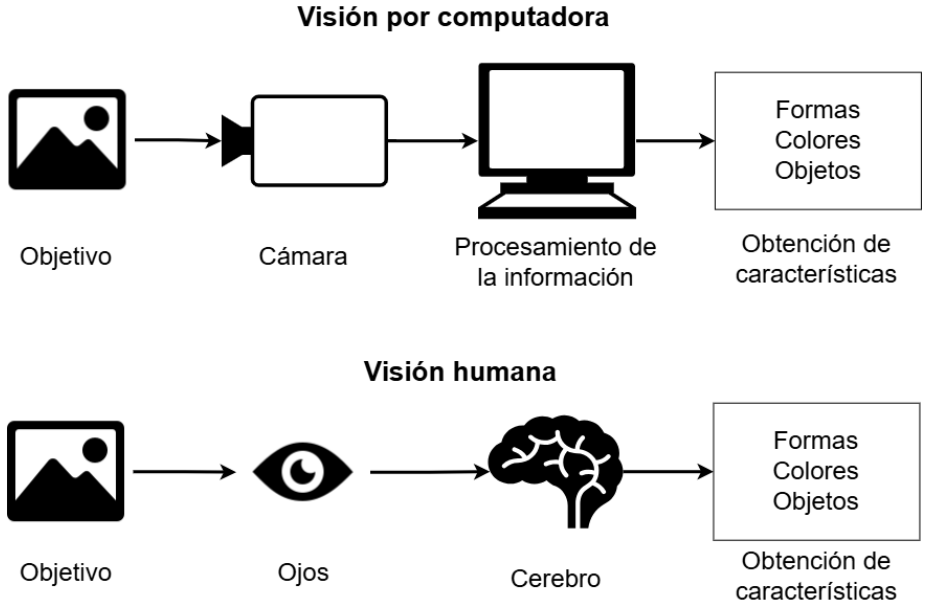

Figura 3: Visión por computadora (Autoría propia, 2024).

En cuanto a las aplicaciones de la Visión por Computadora se encuentran ámbitos como el reconocimiento de objetos, el reconocimiento facial, robótica, aplicaciones médicas, vehículos autónomos, agricultura, entre otros (Shreya et al. 2023). Además, la visión por computadora evoluciona tan rápidamente que al unificarla con técnicas como el aprendizaje profundo (*Deep Learning*, DL), o el procesamiento en tiempo real, permite generar continuamente nuevos desarrollos (Chai et al. 2021).

2.3. Redes Neuronales Artificiales

Las redes neuronales artificiales (ANN) son modelos computacionales basados en el comportamiento de las neuronas biológicas, donde el modelo puede aprender a través de las entradas y salidas del sistema (Samane Sharifi y Rada, 2024). Las ANN se dividen en tres principales arreglos de neuronas denominadas capas, la capa de entrada que proporciona la información inicial a la red, la capa de salida, que muestra los resultados del aprendizaje de

la red, y la capa oculta que a su vez puede estar dividida en múltiples capas que generan el aprendizaje a través de valores obtenidos por la retropropagación de datos denominados pesos (Figura 4). Cada capa de la ANN tiene un número definido de neuronas, dependiendo de los requerimientos de la tarea a realizar (Qamar y Ali Zardari, 2023).

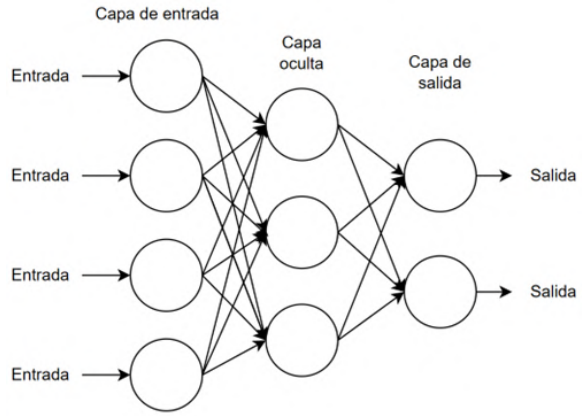

Figura 4: Arquitectura de la ANN (Autoría propia, 2024).

2.4. Redes Neuronales Convolucionales

Las redes neuronales convolucionales (CNN) son algoritmos similares a las redes neuronales artificiales, ya que se componen por neuronas que permiten al algoritmo aprender a través de la retropropagación de la información; sin embargo, la principal diferencia es que las CNN se utilizan para el reconocimiento de patrones en imágenes (Zhao et al. 2024). Las CNN cuentan con diferentes elementos dentro de su arquitectura que permiten modificar y controlar su comportamiento y, al combinarlos, pueden obtener diferentes resultados; sin embargo, la arquitectura más simple de una CNN se puede observar en la Figura 5.

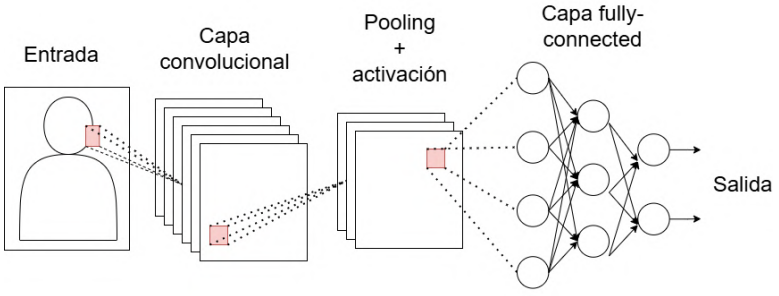

Figura 5: Arquitectura de la CNN (Autoría propia, 2024).

Dentro de la arquitectura de la CNN (Figura 5) se encuentran cuatro principales capas que permiten el funcionamiento del algoritmo (Goodfellow et al. 2016). La capa de entrada de la CNN obtiene los valores de los píxeles de la imagen. La capa convolucional permite

determinar la salida de las neuronas a través de la multiplicación escalar de los pesos de la región de la imagen con una matriz de pesos denominada *kernel*; y por medio de un conjunto de kernels se filtra la imagen, obteniendo sus características más importantes. Por otra parte, la función de activación utilizada en las CNN permite añadir no linealidad al modelo. Mientras que la capa de agrupación o *pooling* permite reducir las representaciones obtenidas a modo de disminuir la carga computacional; donde, dependiendo de la configuración del pooling, puede tomar el promedio o el valor máximo de dichas representaciones. Finalmente, la capa totalmente conectada funciona como una ANN, permitiendo obtener las clases necesarias para la clasificación. El proceso de convolución y pooling se puede generar tantas veces sea necesario, dependiendo de la aplicación de la CNN y la capacidad computacional que se tenga (Saxena, 2022).

2.5. Elementos de las Redes Neuronales

Algunos de los elementos más importantes de las redes neuronales convolucionales, como los hiperparámetros, los parámetros, las métricas de rendimiento y las capas, descritos por Goodfellow et al. (2016) se explican a continuación.

2.5.1. Hiperparámetros

Estos parámetros son configurables y permiten optimizar el rendimiento del sistema, este tipo de parámetros no son generados a través del aprendizaje de los datos, si no que influyen directamente en la capacidad de la red para aprender. Algunos de los hiperparámetros más importantes son el tamaño de lote (*batch size*), la tasa de aprendizaje (*learning rate*), las épocas (*epochs*), entre otros.

Tamaño de lote (Batch size): El tamaño de lote determina el número de ejemplos que el modelo procesa antes de actualizar sus parámetros, esto afecta el costo computacional y la estabilidad del modelo, por lo que, dependiendo de la aplicación del modelo, puede ser ajustado de tal manera que se sacrifique la estabilidad para un entrenamiento más rápido, o se eleve el costo computacional para elevar la precisión del modelo.

Épocas (Epochs): El número de épocas determina las veces que el modelo repite todo el proceso de aprendizaje durante su entrenamiento, esto permite a la retro propagación de la información generar el aprendizaje del modelo al guardar los parámetros obtenidos en la época anterior como parte de las entradas del sistema en la época actual. El número de épocas es un hiperparámetro que modifica directamente el comportamiento de la CNN ya que se deben definir las suficientes para que el modelo se estabilice.

Tasa de aprendizaje (Learning rate): La tasa de aprendizaje o learning rate es un hiperparámetro que determina el tamaño de los pasos que toma el algoritmo de optimización al actualizar los parámetros del modelo, por lo que un valor alto del learning rate podría acelerar el proceso de entrenamiento, pero afectar al rendimiento del modelo de tal modo que no converja, es decir, que el modelo se estabilice para obtener predicciones precisas.

Funciones de activación: Las funciones de activación permiten introducir no linealidad en el modelo entre las capas, existen diferentes funciones de activación dependiendo del

objetivo que se tenga, tales como *ReLU, ELU, Sigmoid, Tanh, Softmax,* entre otras.

Ponderación de datos (Data weighting): La ponderación de datos permite dividir el uso de la información de cada categoría en los procesos de entrenamiento, validación y pruebas para el modelo de CNN.

2.5.2. Parámetros

A los valores que el modelo genera a través del entrenamiento se les conoce como parámetros, estos datos, como los pesos y sesgos, se actualizan durante todo el aprendizaje modificando directamente las métricas obtenidas por el modelo.

Pesos (Weight): Los pesos son un tipo de parámetro que representan las conexiones entre las neuronas de diferentes capas y determinan la influencia de las entradas hacia las salidas, este tipo de parámetro se actualiza durante todo el proceso de aprendizaje y se ajustan conforme al comportamiento de la predicción del modelo y el error. Mientras mayor sea el peso, más influencia tiene sobre el resultado de la salida.

2.5.3. Capa convolucional (Convolutional layer)

La convolución (Ecuación 1), es una operación que superpone una función de probabilidad w sobre una función x, generando una tercera función que permite extraer características locales importantes de la información, mientras mantiene la estructura de los datos.

$$s = x * w \tag{1}$$

Dentro de las redes neuronales la función x se denomina entrada, mientras que la función w se le denomina filtro o *kernel*, obteniendo a su vez una función denominada mapa de características. La Ecuación 2 define la operación de convolución para una imagen de entrada I con un filtro K de tamaño $m \times n$, donde $S(i,\ j)$ es la salida en la posición (i, j), mientras que la Figura 6 muestra la obtención del mapa de características.

$$(i,\ j) = \sum_m \sum_n I\left(i + m - 1,\ j + n - 1\right)\ K(m, n) \tag{2}$$

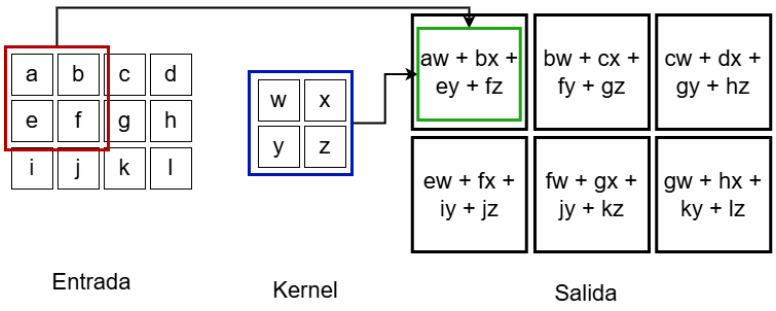

Figura 6: Generación de mapa de características (Autoría propia, 2024).

Dentro de algunos elementos de la configuración de la capa convolucional se encuentra el relleno o *padding*, el cual es un método que permite agregar pixeles alrededor de los bordes de la entrada con el fin de controlar la salida del filtro convolucional y evitar la reducción del tamaño de la imagen después de varias capas. El paso o *stride*, es un parámetro que indica el número de pixeles que se desplaza el proceso de convolución a lo largo de la imagen, por lo que afecta directamente el tamaño del mapa de características (Figura 6).

Otro elemento importante para una capa convolucional es el número de filtros, el cual determina cuantos mapas de características se generan y a su vez se concentran en una característica específica. Por otra parte, el kernel es una matriz cuadrada de pesos el cual realiza el producto punto entre los pixeles de la imagen y los valores del kernel mientras se actualizan para aprender las características de la imagen (Lecun et al. 1998).

De esta forma se obtienen las características necesarias para el reconocimiento de patrones, bordes y objetos dentro de las imágenes.

2.5.4. Capa de agrupación (Pooling layer)

La capa de agrupación o *pooling layer* (Figura 7) permite reducir la dimensionalidad de los mapas de características que se generan a través de los kernels de las capas convolucionales. Existen diferentes tipos de capas de agrupación como el *max pooling* que extrae el valor máximo de una ventana, mientras que el *average pooling* utiliza el valor promedio de dicha ventana.

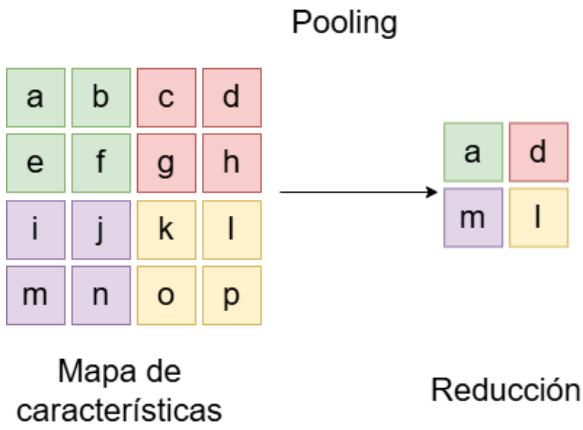

Figura 7: Capa de agrupación (Autoría propia, 2024).

2.5.5. Capa de aplanado (Flatten layer)

Las capas convolucionales y de agrupación generan salidas multidimensionales en forma de tensores que contienen mapas de características, la capa de aplanado permite transformar esta información en un vector unidimensional que pueda ser utilizado como entrada para las capas totalmente conectadas (Figura 8).

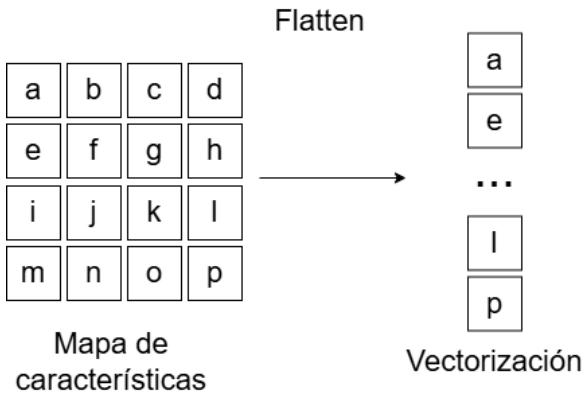

Figura 8: Capa de aplanado (Autoría propia, 2024).

2.5.6. Capa totalmente conectada (Fully-connected)

Dentro de una capa totalmente conectada todas las neuronas de la capa están conectadas a las neuronas de la capa anterior (Figura 9), el número de neuronas de estas capas se definen de tal forma que permiten hacer la predicción y la última capa *fully-connected* realiza la clasificación.

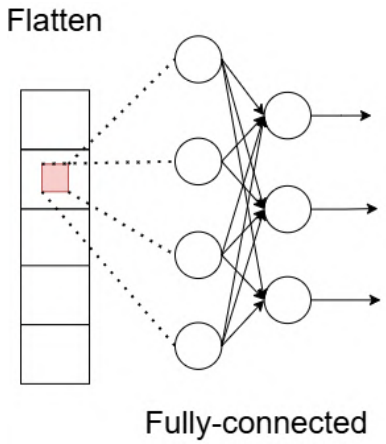

Figura 9: Capa de totalmente conectada (Autoría propia, 2024).

2.5.7. Función ReLU

La función ReLU, del inglés *Rectified Linear Unit* es una función de activación usada comúnmente para el desarrollo de CNN ya que permite introducir no linealidad en la red neuronal permitiendo al modelo aprender representaciones complejas. La Ecuación 3 define el comportamiento de la función ReLU, donde $f(x)$ es la salida de la función ReLU y x es la entrada, usualmente la suma ponderada de las activaciones de la neurona. Esto indica que, si la entrada es mayor que cero, la salida obtiene el valor de x, mientras que, si es menor o

igual a cero, la salida es cero.

$$f(x) = \text{máx}(0, x) \tag{3}$$

Esta función disminuye el costo computacional al no utilizar cálculos complejos como las utilizadas en las funciones sigmoide o tangente hiperbólica, además de evitar problemas de saturación por la reducción de sus gradientes permitiendo que el modelo pueda entrenarse más rápidamente y obtenga mejores resultados y, por otra parte, al únicamente utilizar entradas positivas, esta función desactiva las neuronas que reciben entradas negativas haciendo la red más eficiente. Sin embargo, el desactivar neuronas puede afectar también la precisión de la red al disminuir la capacidad de aprendizaje de la red y afectar la convergencia del modelo.

Sin embargo, el desactivar neuronas puede afectar también la precisión de la red al disminuir la capacidad de aprendizaje de la red y afectar la convergencia del modelo.

2.5.8. Función ELU

La función *Exponential Linear Unit (ELU)*, unidad lineal exponencial, es una función de activación que a diferencia de la función ReLU permite el uso de valores negativos para entradas menores a cero, lo que permite estabilizar las activaciones de las capas. La Ecuación 4 define el comportamiento de la función ELU, donde α es un parámetro que controla el valor de salida para entradas negativas.

$$f(x) = \begin{cases} x & x > 0 \\ \alpha(\exp(x) - 1) & x \leq 0 \end{cases} \tag{4}$$

Al permitir valores negativos esta función puede acelerar el aprendizaje y a su vez evitar que algunas neuronas sean desactivadas beneficiando a que el modelo converja. Sin embargo, esto eleva el costo computacional y una mala elección de α puede afectar el comportamiento del modelo de forma negativa.

2.5.9. Función Softmax

La función Softmax es una función de activación utilizada en la última capa de una red neuronal de clasificación multiclase, de tal modo que convierte las salidas numéricas de las neuronas en probabilidades. La ecuación que define el comportamiento de esta función se muestra en la Ecuación 5.

$$f(z_i) = \frac{e^{z_i}}{\sum_{i=1}^{K} e^{z_i}} \tag{5}$$

Donde z_i es la entrada para la clase i, K es el número total de clases, e^{z_i} es la exponencial de la entrada que asegura que los valores sean positivos, mientras que la sumatoria convierte la salida en una distribución de probabilidad, lo que permite al modelo predecir a partir de una probabilidad la clasificación de los datos.

2.5.10. Entropía cruzada categórica (Categorical Cross-Entropy)

Este tipo de función, conocida como función de pérdida, es utilizada en problemas de clasificación multiclase. La Ecuación 6 define el comportamiento de esta función por medio del logaritmo de la probabilidad predicha para la clase correcta. Si el modelo predice probabilidades bajas para la clase correcta, la pérdida es alta y viceversa.

$$CCE = -\sum_{i=1}^{K} y_i \log\left(p_i\right) \tag{6}$$

Donde y_i es la etiqueta real para la clase i, esta etiqueta se define como *one-hot*, es decir, la clase correcta tiene un valor de 1, mientras que las clases restantes tienen valor de 0. Por otra parte K es el número total de clases y p_i es la probabilidad predicha por el modelo de la clase i, es decir, la probabilidad calculada por la función de activación en la última capa de la red neuronal.

2.5.11. Normalización por lotes (Batch normalization)

La normalización por lotes o *batch normalization* es una técnica para normalizar las activaciones de cualquier capa de una red neuronal a partir del cálculo de la media (Ecuación 7) y la desviación estándar (Ecuación 8) de un lote $\mathbf{H}$ más pequeño o mini-batch, permitiendo regularizar su comportamiento evitando el sobreajuste.

$$\mu = \frac{1}{m}\sum_{i} \mathbf{H}_i \tag{7}$$

$$\sigma = \sqrt{\delta + \frac{1}{m}\sum_{i} \left(\mathbf{H} - \mu\right)_i^2} \tag{8}$$

Mientras que la normalización del mini-batch $\mathbf{H}$ se define como $\mathbf{H\prime}$ y se obtiene a partir de la Ecuación 9.

$$\mathbf{H}' = \frac{\mathbf{H} - \mu}{\sigma} \tag{9}$$

Donde μ y σ son vectores que contienen la media y la desviación estándar de cada unidad respectivamente, por otra parte, δ es un valor positivo pequeño, conocido como bias, que evita obtener una división por cero y m es el número total de activaciones.

Una vez normalizado, se aplican dos parámetros γ y β, conocidos como escala y desplazamiento, los cuales permiten que la red aprenda a ajustar las activaciones por medio de la Ecuación 10.

$$\mathbf{Y}_i = \gamma\mathbf{H}' + \beta \tag{10}$$

El uso de la normalización por lotes permite regularizar el comportamiento del modelo durante el entrenamiento y aumentando su estabilidad, de modo que puede obtener mejores resultados, sin embargo, esto puede aumentar el costo computacional (Ioffe & Szegedy, 2015).

2.5.12. Aumento de datos (Data augmentation)

Es una técnica para aumentar el tamaño del conjunto de datos utilizados para el entrenamiento a través de transformaciones de las imágenes, como rotaciones, cambios de escala, espejado, entre otros, de tal forma que el modelo pueda identificar de mejor manera las categorías al generalizar su información.

2.5.13. Descarte (Dropout)

Es una técnica de regularización que permite apagar de manera aleatoria un porcentaje de las neuronas de una red durante el entrenamiento, evitando problemas de sobreajuste y mejorando la generalización del modelo.

2.5.14. Métricas

Precisión (Acuraccy): Es la proporción de predicciones correctas sobre el total de predicciones realizadas (Ecuación 11), se define comúnmente a través de un porcentaje.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{11}$$

Donde TP son los verdaderos positivos, TN los verdaderos negativos, FP los falsos positivos y FN los falsos negativos.

Pérdida (Loss): Es el resultado de la función de pérdida al obtener el error entre las probabilidades predichas por el modelo contra las probabilidades correctas. Al realizar problemas con clasificación multiclase, se suele utilizar la entropía cruzada categórica para obtener la pérdida (Ecuación 12).

$$Loss = -\frac{1}{N}\sum_{i=1}^{C}\sum_{c=1}^{N} y_{i,c}\log\left(p_{i,c}\right) \tag{12}$$

Donde N es el número total de elementos, C es el número de clases, $p_{i,c}$ es la probabilidad de la clase C para el elemento i, mientras que $y_{i,c}$ es el valor verdadero.

Exactitud (Precision): Es la proporción de predicciones correctas con respecto a las predicciones positivas realizadas, lo que permite conocer la cantidad de falsos positivos (Ecuación 13).

$$Precision = \frac{TP}{TP + TN} \tag{13}$$

Sensibilidad (Recall): Es la proporción de predicciones positivas correctas sobre los datos realmente correctos, permite obtener la cantidad de falsos negativos (Ecuación 14).

$$Recall = \frac{TP}{TP + FN} \tag{14}$$

Valor-F1 (F1-Score): El valor-F1 es la media armónica entre la exactitud y el recall (Ecuación 15).

$$F1Score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{15}$$

2.6. Coeficientes Cepstrales en la Escala de Mel

Los coeficientes cepstrales en la escala de Mel (MFCC) son un método de obtención de características de audio basados en la percepción humana a través de una escala logarítmica o Cepstrum (García Guajardo, 2011). Para generar los MFCC, es necesario seguir la metodología que se describe en la Figura 10.

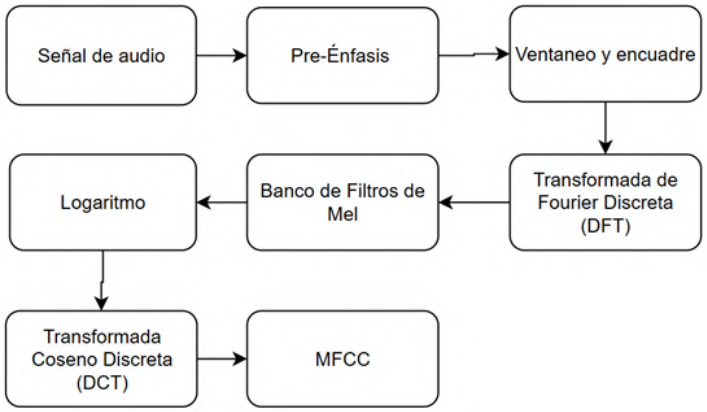

Figura 10: Proceso para la obtención de los MFCC (Autoría propia, 2024).

En primera instancia se debe obtener la señal de audio que se desea procesar, para someterla a un filtro de preénfasis donde se realzan las frecuencias altas, a continuación de deben generar los cuadros o *frames* dividiendo la señal para poder aplicarles una función ventana, comúnmente la ventana de Hamming (Ecuación 16) para el análisis de lenguaje, con el fin de destacar la amplitud de parte central de cada cuadro y reducir la amplitud en los límites de estos.

$$v(n) = a_0 - a_1 \cos\left(\frac{2\pi n}{N-1}\right) \tag{16}$$

Para $a_0 = 0.53836$ y $a_1 = 0.46164$. Donde $v(n)$ es el valor de la ventana de Hamming, a_0 y a_1 son los coeficientes de Hamming, n es el índice de muestra y N el tamaño de la ventana. Una vez realizado el ventaneo se debe aplicar la transformada de Fourier discreta (DFT) como se muestra en la Ecuación 17, donde N representa el número total de muestras en la secuencia discreta, k es el índice de frecuencia en el dominio discreto de la DFT, y $X(k)$ es la representación en el dominio de la frecuencia de una señal discreta en función de la frecuencia angular $\omega = -\frac{2\pi jnK}{N}$ y $x(n)$ es la secuencia discreta en el dominio del tiempo que se va a transformar.

$$X(k) = \sum_{n=0}^{N-1} x_n e^{-\frac{2\pi jnk}{N}} \tag{17}$$

$$\text{Para } k = 1, 2, 3, \ldots, N-1$$

Dichos datos deben evaluarse a través del banco de filtros, el cual es generado por filtros triangulares, y por medio de la función de transferencia definida en la Ecuación 18, la información pasa a la escala de Mel.

$$H_m(k) = \begin{cases} 0 & k < f(m-1) \\ \dfrac{k - f(m-1)}{f(m) - f(m-1)} & f(m-1) \leq k < f(m) \\ 1 & k = f(m) \\ \dfrac{f(m+1) - k}{f(m+1) - f(m)} & f(m) < k \leq f(m+1) \\ 0 & f(m+1) < k \end{cases} \tag{18}$$

En dicha ecuación, $f(m)$ es la frecuencia central del filtro triangular, k representa la frecuencia en el dominio discreto, $H_m(k)$ es el valor de un filtro triangular de Mel en la frecuencia k, f representa la frecuencia en el dominio de las señales de audio o en el espectro de frecuencia, m es un índice que se utiliza para denotar un filtro específico en el banco de filtros Mel de $m = 0$ hasta $m = M - 1$, y el valor de $\sum_m^{M-1} H_m(k) = 1$.
La escala de Mel a partir de la frecuencia de respuesta se obtiene por medio de la Ecuación 19, y de forma inversa en la Ecuación 20, donde m es la escala de Mel y f es la frecuencia de respuesta.

$$m = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \tag{19}$$

$$f = 700 \cdot (10^m \cdot 2595 - 1) \tag{20}$$

Ya que la escala de Mel (Figura 11) se basa en el nivel de percepción del oído humano, esta escala se vuelve lineal para frecuencias debajo de 1kHz, mientras que, para frecuencias por encima de este valor, es de forma logarítmica.

Finalmente, se obtiene el logaritmo de la señal y se le aplica la transformada de coseno discreta (DCT) como se muestra en la Ecuación 21, donde x_n es la señal discreta en el tiempo, $X(k)$ es la señal discreta en la frecuencia, en función de la frecuencia angular $\omega = -\frac{2\pi jnK}{N}$, y N es la longitud de la señal. Lo que da como resultado el vector de coeficientes para cada cuadro (Martínez-Mascorro & Aguilar-Torres, 2013).

$$X(k) = \sum_{n=0}^{N-1} x_n \cos\left(\frac{2\pi jnk}{N}\right) \tag{21}$$

$$\text{Para } k = 1, 2, 3, \ldots, N-1$$

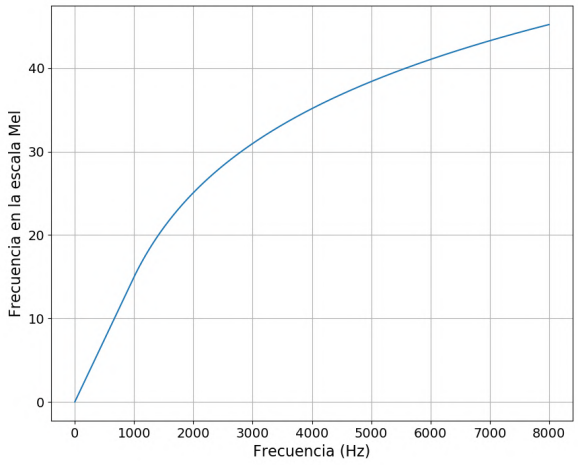

Figura 11: Escala de Mel (Autoría propia, 2024).

Es importante mencionar que los MFCC tienen aplicaciones industriales, médicas y de análisis acústico. Dentro de las aplicaciones industriales se encuentra el monitoreo del estado de rodamientos, turbinas, bombas y engranes, entre otros. Mientras que, para el análisis médico, se utilizan para identificar enfermedades, generar diagnósticos por electrocardiograma y análisis por electroencefalograma. Por otro lado, el análisis acústico permite aplicaciones enfocadas en el reconocimiento de voz, reconocimiento de emociones, dialecto y lenguaje, entre otras (Abdul & Al-Talabani, 2022).

2.7. Jetson Nano Developer Kit

La Jetson Nano Developer Kit (Figura 12) es un módulo de desarrollo que permite el procesamiento de algoritmos inteligentes para detección de objetos, clasificación de imágenes, procesamiento de voz, entre otros.

Figura 12: Jetson Nano Developer Kit (NVIDIA, 2019)

Este dispositivo cuenta con un procesador ARM Cortex-A57 Quad-Core de 64 bits y una GPU NVIDIA Maxwell de 128 núcleos, lo cual le permite tener un alto rendimiento para procesamiento en aplicaciones de visión por computadora y desarrollo de Inteligencia Artificial (IA), además, cuenta con una memoria RAM de 4GB para complementar al procesamiento (NVIDIA, 2019).

Por otra parte, permite adicionarle diferentes periféricos para los módulos de Raspberry Pi, tales como la cámara, la pantalla, entre otros; así como los necesarios para la comunicación inalámbrica, ya sea por WiFi o Bluetooth. Esta tarjeta incluye múltiples puertos USB, así como un puerto HDMI y un puerto RJ-45 para conexión Ethernet; además de pines configurables para el desarrollo. En la Tabla 1 se incluyen datos técnicos.

Tabla 1: Especificaciones técnicas de la Jetson Nano Developer Kit (NVIDIA, 2019).

Característica	**Descripción**
Memoria	4GB 64 bits LPDDR4 25.6 GB/s
Almaenamiento	Por medio de MicroSD
CPU	ARM-57 de 4 núcleos
GPU	Maxwell de 128 núcleos
USB	2 puertos 3.0, 1 puerto 2.0 Micro-B
Periféricos	GPIO, I2C, I2S, SPI, UART
Alimentación	5V - 4A
Largo	100 mm
Alto	80 mm
Ancho	29 mm

2.8. MediaPipe Face Landmarker

El identificador de rostros Face Landmarker de MediaPipe es un modelo de detección de rostros a partir de un modelo preentrenado, que permite identificar las regiones de la cara por medio de una malla de puntos (Figura 13). Este modelo se utiliza para modelos de aprendizaje automático.

El modelo genera 468 puntos de referencia en tres dimensiones a través de Machine Learning, utilizando como entrada una imagen. Por otra parte, la cadena de procesos o pipeline que permite al modelo de ML identificar los rostros en la imagen consiste en dos DNN, permitiendo así identificar tanto el rostro como la posición de las superficies, incluso cuando existen objetos frente a este (Mediapipe, 2023).

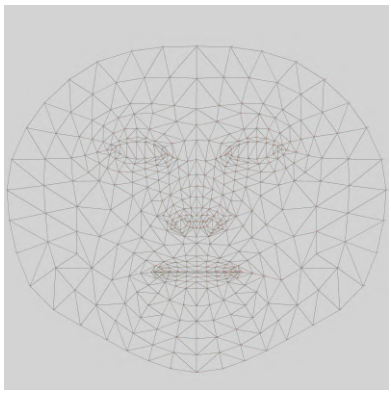

Figura 13: Malla de puntos de Media Pipe (Mediapipe, 2023).

2.9. Base de datos FER-2013

La base de datos de imágenes FER-2013 (*Facial Expression Recognition 2013 Dataset*), es una base de datos bajo la licencia *Creative Commons 4.0* que contiene siete categorías emocionales para el entrenamiento de modelos enfocados en el reconocimiento de expresiones faciales. Dentro de esta base de datos se encuentran imágenes para enojo, disgusto, miedo, felicidad, tristeza, sorpresa y neutral; teniendo una resolución de 48 x 48 píxeles en escala de gris. Esta base de datos se ha validado a través de diferentes estudios donde muestra precisiones de 74.11 %, 71.52 %, 78.90 %, a través de una *Deep Integrated CNN* (Shi et al. 2021), una *Multibranch Cross Conection CNN* (Saurav et al. 2021) y una CNN (Mohan et al. 2021), respectivamente.

2.10. Base de datos MESD

La base de datos Mexican Emotional Speech Database (MESD), contiene 864 grabaciones de voz, para enojo, desagrado, miedo, felicidad, tristeza y neutral. La información de esta base de datos se muestra a partir de cuatro hombres adultos con una edad promedio de 22.75 años y una desviación estándar de 2.06, cuatro mujeres adultas con una edad media de 22.25 y una desviación estándar de 2.50, y ocho niños con una edad media de 9.87 y una desviación estándar de 1.12. Estos participantes fueron voluntarios y no son actores profesionales; además, la base de datos fue generada a partir de una población desarrollada en un entorno sociocultural mexicano donde se excluyó al participante si presentaba alguna patología que afectara al comportamiento emocional o el habla. De esta forma se tienen tres categorías importantes: mujer adulta, hombre adulto y niños, y se validó por medio de una SVM, obteniendo resultados de 89.4 % para la categoría de mujer adulta, 93.9 % para la categoría de hombre adulto y 83.3 % para la categoría de niños (Duville et al. 2021). A partir de 48 palabras por emoción, por cada una de las categorías se tienen 864 grabaciones de voz capaces de otorgar la información necesaria para los sistemas de reconocimiento de emociones a través de la voz, enfocadas en la población mexicana.

3. Metodología

La metodología general para el desarrollo del presente trabajo de investigación se muestra en la Figura 14.

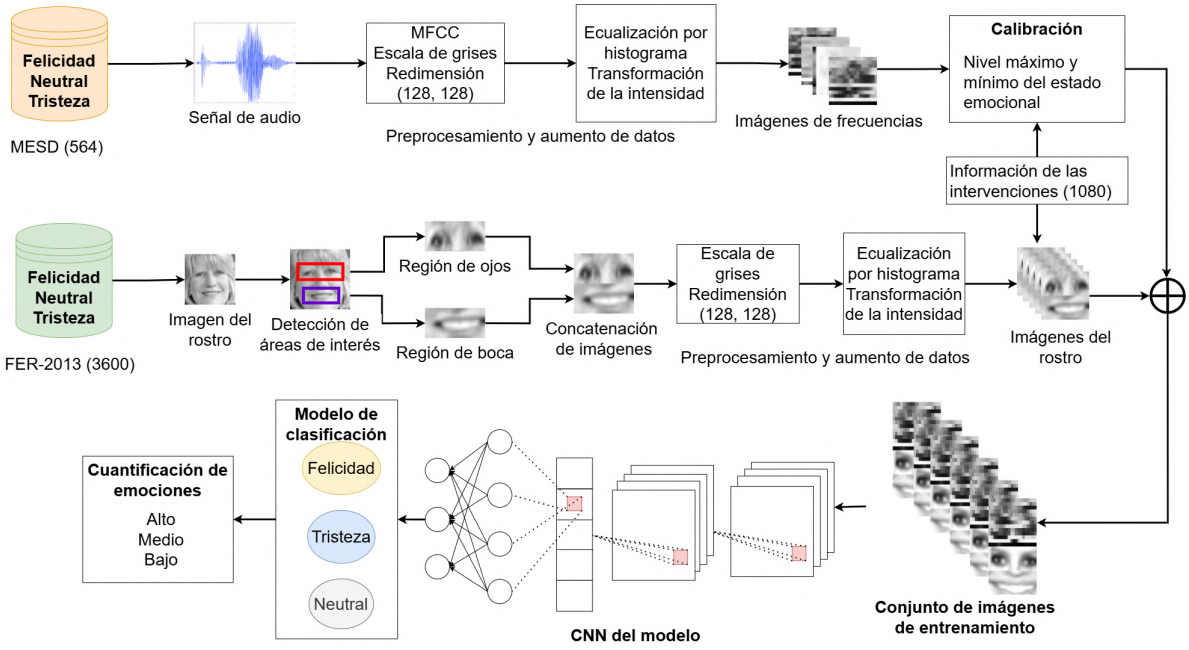

Figura 14: Metodología del proyecto (Autoría propia, 2024).

De manera general, el sistema permite contener los componentes necesarios para la detección, clasificación y cuantificación de emociones al adquirir el video y audio por medio de un micrófono y una cámara web, o de procesar información existente que ya se encuentre en estos formatos. Para el sistema SER se procesa la información de audio para generar imágenes de frecuencias y la calibración del sistema a través de la intensidad, mientras que para el sistema FER se consideran las regiones de ojos y boca para generar una nueva imagen con dicha información. Posteriormente las imágenes del sistema FER y SER se concatenan para obtener las imágenes de entrenamiento de la CNN. Después se utiliza información nueva proporcionada por profesionales de la salud mental, donde los protocolos para obtener la información son determinados, respetando las implicaciones éticas necesarias. Dicha información permite rentrenar el modelo base para considerar las características de los individuos. Una vez que el modelo de detección predice una emoción a partir de las nuevas entradas del sistema, este es capaz de entregar la cuantificación del nivel emocional. A continuación, se explican a mayor detalle cada uno de los módulos de esta metodología.

3.1. Preprocesamiento de datos

Para desarrollar los procesos de clasificación, calibración y cuantificación del sistema se realiza el entrenamiento y la calibración a partir de archivos de audio y video de sesiones de intervención psicológica virtuales, considerando las implicaciones éticas para el manejo de información y la confidencialidad de datos, y en colaboración con profesionales en el área de la salud mental para la identificación de las emociones durante las grabaciones. Las grabaciones corresponden a dos sujetos del sexo femenino con una media de edad de 33.5 y una desviación estándar de 4.5.

Los sujetos fueron evaluados por el profesional de la salud mental a partir la segunda edición del Inventario de Depresión de Beck (*Beck Depression Inventory II*, BDI-II, Anexo 7.4), la cual es una de las escalas más utilizadas a nivel internacional para detectar el nivel de depresión y cuenta con una traducción validada para población mexicana (González et al. 2015), además del Inventario de Ansiedad de Beck (Beck *Anxiety Inventory*, BAI, Anexo 7.5), la cual es una de las escalas más utilizadas a nivel mundial para detectar y evaluar la presencia de ansiedad de manera confiable y precisa (Padrós Blázquez et al. 2020), y la escala de Afecto Positivo/Afecto Negativo (PANAS, Anexo 7.6), la cual permite obtener una valoración del nivel de bienestar de una persona, donde se pueden identificar afectos positivos, como la motivación, el éxito, el nivel de energía, entre otros; y afectos negativos, como el miedo, la frustración, el fracaso, entre otros (Velasco Matus et al. 2021). A partir de esta información el profesional de la salud mental valida que los sujetos no presentan ninguna patología psicológica que pueda poner en riesgo su salud y afectar los resultados de las pruebas del sistema.

Los tiempos de inicio para cada emoción a lo largo de estos archivos son determinados por el profesional de la salud mental, por lo que únicamente se considera un periodo de 10 segundos a partir del tiempo de inicio para obtener la información referente a cada emoción. El procesamiento de esta información se muestra en la 15.

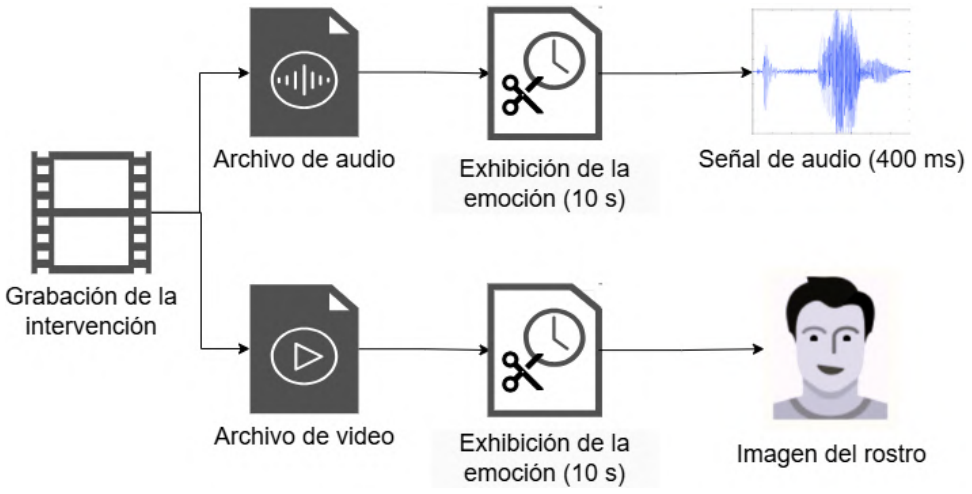

Figura 15: Procesamiento de los archivos de audio (Autoría propia, 2024).

Los segmentos de audio de 10 segundos son divididos en fracciones de 400ms para igualar la duración de los archivos de audio de la base de datos MESD. Esta nueva ventana de tiempo permite generar la información necesaria para obtener las imágenes de frecuencias a través de los MFCC.

En cuanto a la obtención de las imágenes para la detección de emociones a través de expresiones faciales, se guardan las imágenes al inicio de cada segmento de 10 segundos para su posterior segmentación en las regiones de interés.

De esta forma se tiene la información necesaria para realizar el entrenamiento, calibración y cuantificación del sistema a partir de la información de un solo usuario.

3.2. Sistema de reconocimiento de expresiones faciales (FER)

Para generar el sistema de reconocimiento de expresiones faciales se utiliza la información de la base de datos FER-2013, donde a través del identificador Face Mesh de Mediapipe se obtienen las regiones de interés que corresponden a las zonas de ojos y boca para las emociones de felicidad, tristeza y el estado neutral.

Estas zonas de interés facilitan la creación de dos imágenes con un área de selección determinada por rectángulos definidos por los puntos 206 y 431 para los ojos, y 70 y 346 para la boca. A partir de este par de imágenes se generó una nueva imagen con base en la concatenación de los datos de ojos y boca, como se muestra en la Figura 16.

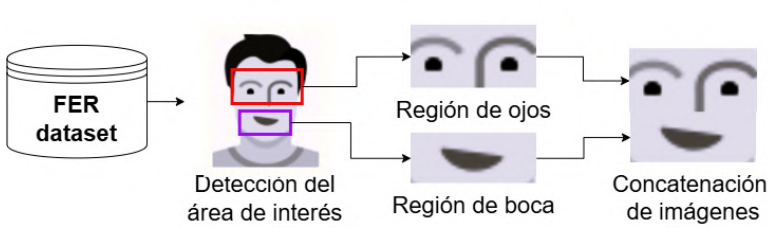

Figura 16: Concatenación de imágenes (Autoría propia, 2024).

Mediante la detección del rostro del modelo de Mediapipe se obtienen 600 imágenes frontales concatenadas para cada categoría de la base de datos FER-2013. Los datos (imágenes) se someten al siguiente preprocesamiento:

- Se aplica una transformación a escala de grises al conjunto de imágenes concatenadas.
- Las imágenes seccionadas se redimensionan a un tamaño de 64 x 128 pixeles para las zonas individuales de ojos y boca, con el fin de que puedan ser concatenadas sin problemas.
- Las imágenes resultantes de la concatenación de ojos y boca deben tener un tamaño de 128 x 128 pixeles.
- Se aplica una transformación de la intensidad con un valor alfa de 1.2 al conjunto de imágenes concatenadas.

- Dichas imágenes posteriormente pasan por una ecualización del histograma con el fin de mejorar la calidad de la imagen.
- Finalmente, se aplica un espejo horizontal a las imágenes, para un aumento de datos, obteniendo 1200 imágenes por cada categoría.

En cuanto a las imágenes para el sistema FER de las intervenciones se tienen inicialmente 180 imágenes de cada categoría para cada uno de los sujetos y una vez realizado el mismo preprocesamiento, se obtienen 360 imágenes para las emociones de tristeza, felicidad y el estado neutral.

El conjunto de imágenes procesadas se pondera para el desarrollo del sistema como se muestra en la Tabla 2.

Tabla 2: Ponderación de datos del sistema FER.

Proceso	**Ponderación (%)**
Entrenamiento	60
Validación	20
Pruebas	20

La configuración de las capas de la red neuronal para el sistema FER se define por medio de cuatro bloques iguales, los cuales contienen una capa convolucional de 32, 128, 32 y 64 filtros, una capa de normalización por lotes y una capa max pooling; el ultimo bloque de la configuración contiene una capa de aplanado, una capa de descarte y finalmente una capa fully-connected de 3 neuronas, como se muestra en la Figura 17.

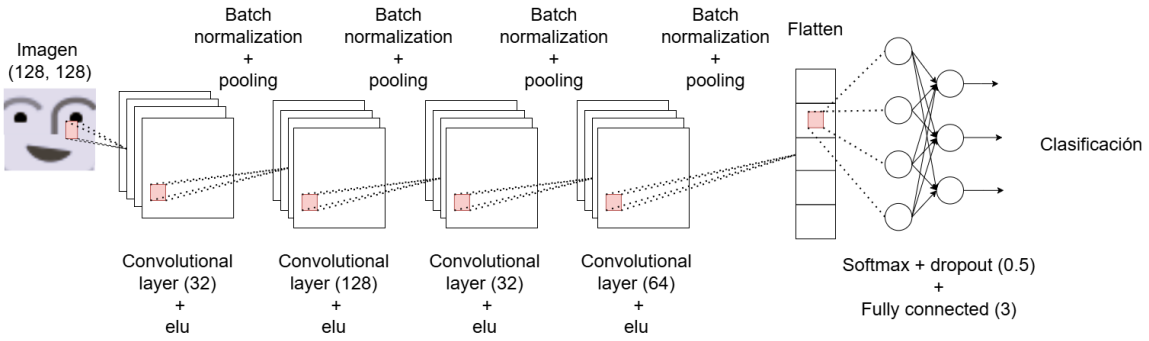

Figura 17: Configuración de la CNN del sistema FER (Autoría propia, 2024).

Para el entrenamiento de este modelo se utilizan los hiperparámetros que se muestran en la Tabla 3.

Tabla 3: Configuración de los hiperparámetros del modelo FER.

Hiperparámetro	**Valor**
Función de activación	ReLU
Épocas	50
Tamaño de lote	16
Clases	3
Dropout	0.5
Factor de descenso del LR	0.5
Tasa de aprendizaje final	1e-06
Tasa de aprendizaje inicial	0.001
Optimizador	Adam
Función de activación de la salida	Softmax
Altura de la imagen	128
Ancho de la imagen	128
Canales	1
Tamaño de kernel (conv)	3x3
Tamaño de kernel (padding)	2x2

3.3. Sistema de detección de emociones a través de la voz (SER)

Dentro de la base de datos MESD se encuentran señales de voz que corresponden a varios sujetos para cada una de las categorías emocionales, por lo que para realizar el sistema SER y la calibración del sistema únicamente se considera la información de uno de los sujetos, es decir, de los 144 archivos de audio dentro de cada emoción de la base de datos, únicamente se consideran 48.

Estas señales están clasificadas por la emoción que representan (felicidad, tristeza y el estado neutral).

De esta forma se cuenta con la información necesaria para extraer las características de las señales; las cuales se obtienen a partir de los MFCC y obtener los niveles de intensidad para generar el proceso de cuantificación.

Las señales de voz de un solo sujeto de la MESD se transforman a imágenes de frecuencias en escala de grises por medio de las librerías OpenCV, Matplotlib y Librosa, con el fin de poder entrenar la CNN (Figura 18).

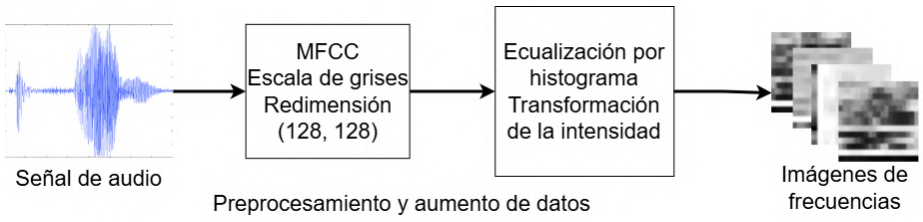

Figura 18: Obtención de imágenes de frecuencias (Autoría propia, 2024).

Estas imágenes de frecuencias se someten al preprocesamiento que se describe a continuación:

- Se seleccionan 48 imágenes de frecuencias de cada categoría y realiza una transformación a escala de grises.
- Las imágenes se redimensionan para tener un tamaño de 128x128 pixeles.
- Se hace una transformación de la intensidad a las imágenes con un valor alfa de 1.2, y se generan nuevos archivos obteniendo un total de 96 imágenes.
- Las imágenes se someten a una ecualización del histograma con el fin de mejorar la calidad de la imagen, obteniendo un total de 192 imágenes al generar nuevos archivos.

Para la información de las intervenciones se generan inicialmente 180 archivos de audio que después de realizarse la transformación a imagen, se someten al mismo preprocesamiento que las imágenes de la base de datos MESD, sin embargo, en este caso se sobrescribe la información al realizar la ecualización del histograma y se aumenta únicamente con la transformación de la intensidad, generando 360 imágenes para cada categoría de cada uno de los sujetos.

De esta forma se obtiene la información necesaria para el entrenamiento de la CNN utilizando la información de un solo sujeto. Esta información se pondera como muestra la Tabla 4 para el desarrollo del modelo del sistema SER.

Tabla 4: Ponderación de datos del sistema SER.

Proceso	**Ponderación (%)**
Entrenamiento	60
Validación	20
Pruebas	20

Para el sistema SER la configuración de las capas de la red neuronal se define por medio de cinco bloques, los cuales contienen una capa convolucional de 8, 16, 32, 64 y 64 filtros, una capa de normalización por lotes y una capa max pooling, el ultimo bloque de la configuración contiene la capa de aplanado, una capa fully-connected con 64 neuronas, una capa de descarte y finalmente una capa fully-connected con 3 neuronas, como se muestra en la Figura 19.

Por otra parte, los hiperparámetros definidos para el entrenamiento de este modelo se muestran en la Tabla 5.

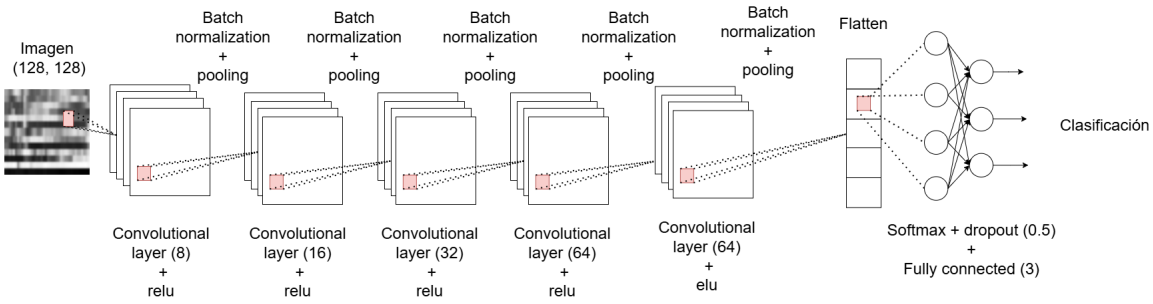

Figura 19: Configuración de la CNN del sistema SER (Autoría propia, 2024).

Tabla 5: Configuración de los hiperparámetros del modelo FER.

Hiperparámetro	**Valor**
Función de activación	ELU
Épocas	50
Tamaño de lote	32
Clases	3
Dropout	0.5
Factor de descenso del LR	0.5
Tasa de aprendizaje final	1e-06
Tasa de aprendizaje inicial	0.001
Optimizador	Adam
Función de activación de la salida	Softmax
Altura de la imagen	128
Ancho de la imagen	128
Canales	1
Tamaño de kernel (conv)	3x3
Tamaño de kernel (padding)	2x2

3.4. Calibración del sistema

Para la calibración del sistema se obtienen los valores máximos y mínimos de los coeficientes cepstrales de todo el conjunto de imágenes del sistema SER, de tal forma que se tenga el nivel de intensidad para el sujeto en cuestión (Ecuación 22). Este proceso permite tener la información necesaria para generar la calibración del sistema.

$$
\begin{aligned}
I_{max} &= \text{máx}\left(\text{máx}\left(I_1\right), \text{máx}\left(I_2\right), ..., \text{máx}\left(I_n\right)\right) \\
I_{min} &= \text{mín}\left(\text{mín}\left(I_1\right), \text{mín}\left(I_2\right), ..., \text{mín}\left(I_n\right)\right) \\
\delta_I &= I_{max} - I_{min}
\end{aligned} \tag{22}
$$

Donde $I_1, I_2, \ldots, I_n$ corresponde a la matriz de datos de los MFCC para el entrenamiento del modelo, I_{max} corresponde al valor máximo de intensidad de dicho conjunto de imágenes, mientras que I_{min} corresponde al valor mínimo y δ_I representa el rango emocional obtenido a partir de estos valores.

Posteriormente, los valores de I_{max} e I_{min} se almacenan en un archivo JSON, para poder acceder a los valores de δ_I al momento de procesar la información nueva y generar la cuantificación.

El proceso de calibración se realiza para cada uno de los sujetos a partir de la información de audio utilizada en el desarrollo del sistema SER.

3.5. Modelo de clasificación

Con la finalidad de combinar el sistema a través de la concatenación de la información de ambos sistemas se desarrolló un modelo, denominado como modelo de concatenación o CONCAT, a partir de las imágenes de los datos de voz generados con los MFCC del sistema SER y las imágenes faciales de ojos y boca del sistema FER, generando una nueva imagen como se muestra en la Figura 20.

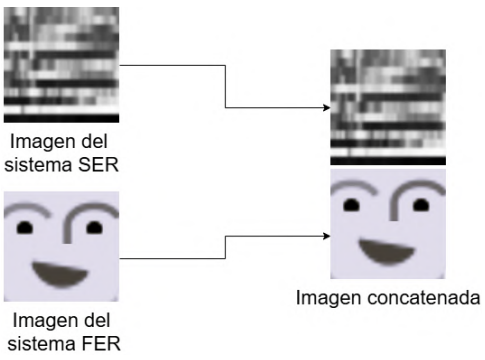

Figura 20: Concatenacíon de datos del sistema FER y SER (Autoría propia, 2024).

Realizando este proceso de concatenación se obtuvo una base de datos, denominada en este trabajo de investigación como CMF, con 1200 imágenes en escala de grises de 256x128 pixeles.

Para el entrenamiento de este modelo a partir de las imágenes de la base de datos CMF, se consideró la ponderación para el entrenamiento de la CNN que define la Tabla 6.

Tabla 6: Ponderación de datos del modelo concatenado.

Proceso	**Ponderación (%)**
Entrenamiento	60
Validación	20
Pruebas	20

Se configuró una CNN con dos capas convolucionales y pooling, para 16 y 8 filtros antes de entrar a la capa de aplanado, la capa de descarte y la capa fully-connected de 3 neuronas (Figura 21) y los hiperparámetros mostrados en la Tabla 7.

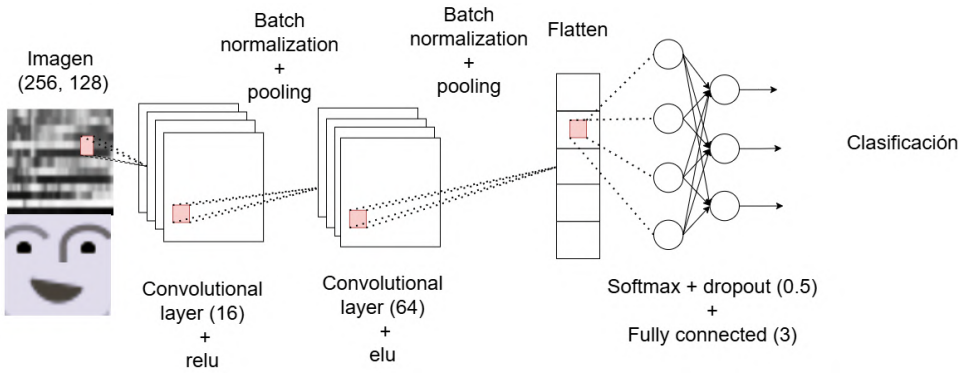

Figura 21: Configuración de la CNN del modelo CONCAT (Autoría propia, 2024).

A partir de la creación de la base de datos de imágenes concatenadas CMF y el sistema CONCAT se combinan los sistemas FER y SER. De igual forma se realiza la concatenación de datos de la información de las grabaciones de ambos sujetos.

Una vez entrenado este modelo, se realiza un reentrenamiento utilizando las imágenes concatenadas de cada uno de los sujetos, generando así dos modelos de concatenación calibrados a las características individuales de dichos sujetos.

3.6. Cuantificación del estado emocional

Para la cuantificación del estado emocional se considera la intensidad de cada emoción y la predicción del modelo CONCAT, donde se toman los valores de δ_I de la emoción predicha para generar el nivel emocional. Para esto se realiza un proceso similar al utilizado en la calibración donde se toma nueva información del conjunto de imágenes $J_1, J_2, \ldots, J_n$ para calcular sus valores máximos J_{max} y mínimos J_{min}. De esta manera se define la Ecuación 23.

$$n_e = \frac{J_{max} - J_{min}}{\delta_I} \tag{23}$$

Tabla 7: Configuración de los hiperparámetros del modelo CONCAT.

Hiperparámetro	**Valor**
Función de activación	ELU
Épocas	50
Tamaño de lote	32
Clases	3
Dropout	0.5
Factor de descenso del LR	0.5
Tasa de aprendizaje final	1e-06
Tasa de aprendizaje inicial	0.001
Optimizador	Adam
Función de activación de la salida	Softmax
Altura de la imagen	256
Ancho de la imagen	128
Canales	1
Tamaño de kernel (conv)	3x3
Tamaño de kernel (padding)	2x2

Donde n_e corresponde al nivel emocional de la nueva información con respecto a la escala de intensidad emocional de la calibración. Este nivel emocional se evalúa a través de la condición

$$q = \begin{cases} bajo & si \quad 0 < n_e \leq 33 \\ medio & si \quad 33 < n_e \leq 66 \\ alto & si \quad 66 < n_e \end{cases}$$

De tal forma que el valor de q representa el nivel emocional para la cuantificación.

Para validar la cuantificación del estado emocional se comparan los resultados del sistema respecto a la información de audio y video de dos sujetos, considerando los procesos de entrenamiento y calibración con su respectiva información.

3.7. Sistema embebido

La funcionalidad del sistema embebido se basa en el procesamiento a través de un módulo Jetson Nano Developer Kit de la marca NVIDIA, y por medio de una pantalla permite visualizar la información de cuantificación y clasificación. El sistema es capaz tanto de procesar la información existente, como de recolectar nueva información a través de una cámara web y un micrófono de solapa que se le puede colocar al individuo.

Los elementos internos que son parte del sistema embebido se enlistan a continuación:

- 1 Pantalla táctil.
- 1 Jetson Nano Developer Kit.

- 2 Ventiladores sin escobillas de 25x25x6.5 mm.
- 1 Cámara web.
- 1 Módulo de conexión inalámbrica.
- 1 Receptor del micrófono.

Para proteger el sistema embebido y poder contener todos sus elementos, es esencial contar con una carcasa, la cual debe ser capaz de proteger los elementos y a la vez ser relativamente pequeña para su fácil transporte y uso, por lo que se consideran las dimensiones que se muestran en la Figura 22.

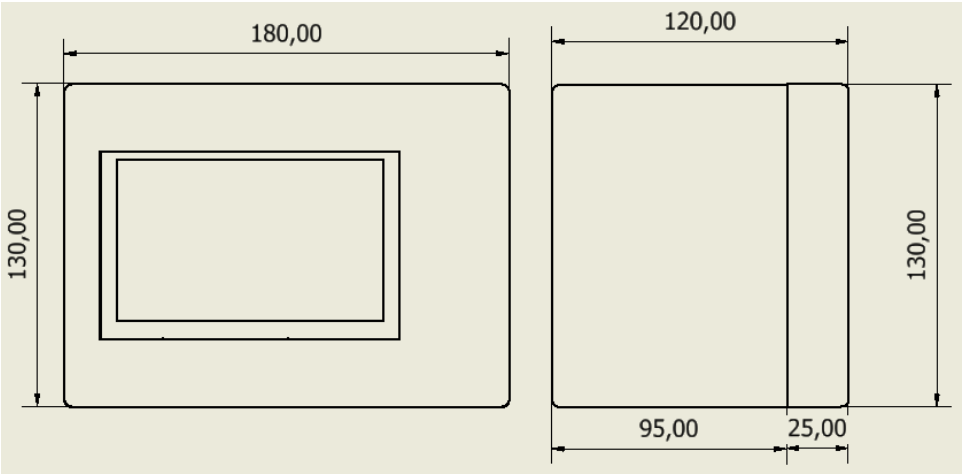

Figura 22: Plano del sistema embebido (Autoría propia, 2024).

A partir de esto se diseña una carcasa por medio de Autodesk Inventor capaz de contener todos los elementos electrónicos de manera que el sistema sea seguro de transportar y evite cualquier tipo de riesgo para el usuario. Para el diseño se consideró la ubicación de los elementos para no afectar la ventilación de la Jetson Nano y, a su vez, provocar un sobrecalentamiento del sistema.

La carcasa del sistema embebido se divide en dos partes, la primera (Figura 23) contiene los elementos como la Jetson Nano, los ventiladores, la cámara web y el módulo de conexión inalámbrica, además de orificios para el uso de los puertos USB, la ventilación y la alimentación.

La segunda (Figura 24) parte de la carcasa corresponde a la tapa y permite el montaje de la pantalla táctil.

Para validar las dimensiones de la carcasa se utilizaron los modelos de los elementos para poder ubicarlos de manera más clara (Figura 25).

Por lo que, la Figura 26 muestra el modelado final de la carcasa con los elementos ensamblados. Una vez diseñada, esta carcasa se imprime en 3D utilizando PLA como material de impresión y una impresora Flashforge Guider II.

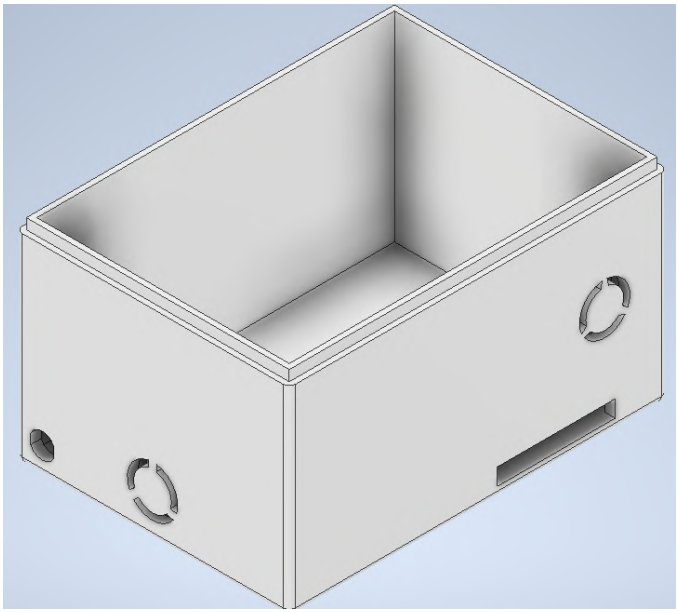

Figura 23: Carcasa del sistema embebido (Autoría propia, 2024).

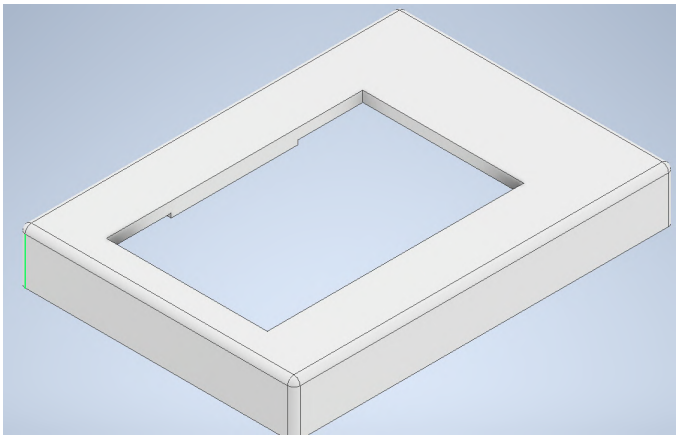

Figura 24: Tapa de la carcasa del sistema embebido (Autoría propia, 2024).

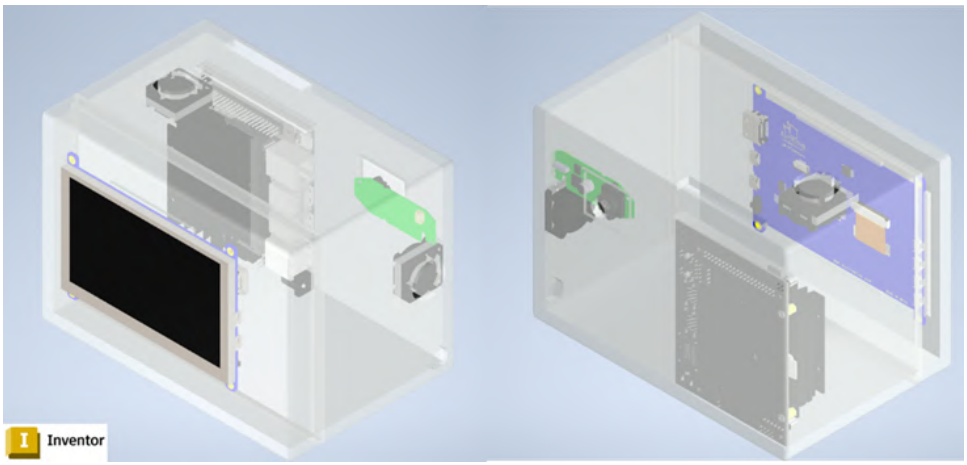

Figura 25: Modelado 3D del sistema embebido con vista a los componentes (Autoría propia, 2024).

Para la impresión de la carcasa se consideró el Reglamento de Laboratorio de la Universidad Autónoma de Querétaro (Anexo 7.3), para el uso de la impresora 3D, el uso del laboratorio de Visión Artificial y el manejo del desecho de filamento.

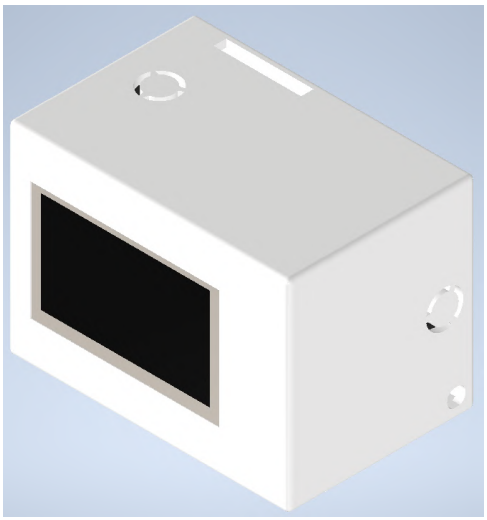

Figura 26: Modelado 3D del sistema embebido (Autoría propia, 2024).

En el diseño de la carcasa y la disposición de los elementos, se tomó en cuenta la Norma Oficial Mexicana NOM-241-SSA1-2021, la cual aborda las buenas prácticas relacionadas con la manufactura de dispositivos médicos. Esta normativa tiene como propósito establecer los requisitos mínimos que deben seguirse en las etapas de diseño, producción, almacenamiento y distribución de dispositivos médicos, teniendo en consideración su nivel de riesgo. Su objetivo principal es garantizar que estos dispositivos cumplan con los estándares de calidad, seguridad y funcionamiento, de modo que puedan ser utilizados de manera segura por los consumidores

finales o pacientes. Cabe destacar que su cumplimiento es obligatorio en todo el territorio nacional, aplicando a todos los establecimientos involucrados en la fabricación de dispositivos médicos, así como a los almacenes de acondicionamiento, depósito y distribución de estos dispositivos (D.O.F., 2021). En este caso, se desarrolla un sistema inocuo y no invasivo que aterriza en la Clase I de esta norma, por lo que cumple con sus indicaciones.

3.8. Pruebas y validación del sistema embebido

Para determinar la viabilidad del sistema embebido, se realiza la detección y cuantificación de emociones a partir de videos de sesiones de intervención psicológica, con la finalidad de visualizar el funcionamiento del sistema.

Se genera la detección y cuantificación de emociones a través de tres versiones del modelo de concatenación, el modelo CONCAT base, el modelo CONCAT reentrenado con la información del sujeto 1 y el modelo CONCAT reentrenado con la información del sujeto 2, con el fin de validar el proceso de calibración.

4. Resultados

Las métricas obtenidas por los modelos del sistema FER, el sistema SER y el modelo CONCAT se muestran en la Tabla 8. Estas métricas se obtienen por medio de una laptop con 16 GB de RAM DDR4, un procesador i7-13650HX, y una tarjeta gráfica NVIDIA GeForce RTX 4050.

Tabla 8: Métricas obtenidas por los modelos FER, SER y la concatenación.

Sistema	**FER**	**SER**	**CONCAT**
Base de datos	FER-2013	MESD	CMF
Precisión (%)	96.86	99.90	99.95
Pérdida	0.0822	0.004	0.002
Precisión val (%)	88.05	90.43	100
Pérdida val	0.3552	0.2856	1.40E-06
Exactitud (%)	89.71	92.87	100
Recall (%)	89.72	92.75	100
F1-Score (%)	89.71	92.71	100
Tiempo (min)	19.87	2.97	10.34

Se puede observar que el comportamiento del modelo CONCAT permite obtener mejores métricas que los modelos SER y FER de manera individual, además de reducir considerablemente el tiempo de entrenamiento con respecto al sistema FER.

A continuación, se presentan los resultados de cada modelo y se muestra su comportamiento a lo largo del entrenamiento.

4.1. Sistema FER

A partir del entrenamiento del sistema FER con la base de datos FER-2013 se obtuvo un comportamiento que indica que el sistema es capaz de reconocer las imágenes con una precisión del 96.86 % y una pérdida del 8.22 %, completando su entrenamiento en un tiempo de 19.87 minutos, obteniendo buenas métricas de rendimiento.

Se puede observar en la Figura 27 que el comportamiento de la precisión durante el entrenamiento y la validación es similar.

Sin embargo, el valor de la pérdida (loss) durante la validación es muy alta, en comparación con el entrenamiento (Figura 28), donde la diferencia es considerable y puede afectar el rendimiento del sistema ante las diferentes imágenes de entrada.

Mientras que la matriz de confusión (Figura 29) muestra que la emoción de felicidad es identificada más fácilmente por el modelo, a comparación de la tristeza y neutral, donde tiende a confundir una mayor cantidad de datos; sin embargo, en términos generales, las categorías son identificadas satisfactoriamente.

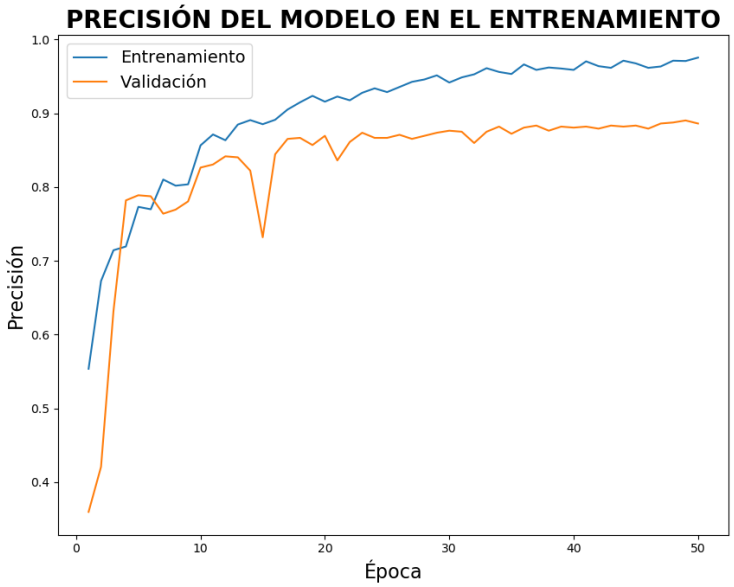

Figura 27: Evolución de la precisión del sistema FER (Autoría propia, 2024).

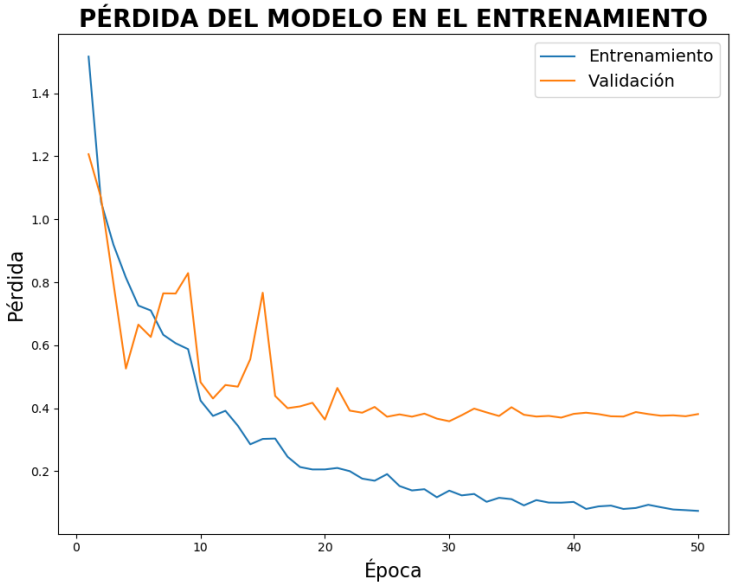

Figura 28: Evolución de la pérdida del sistema FER (Autoría propia, 2024).

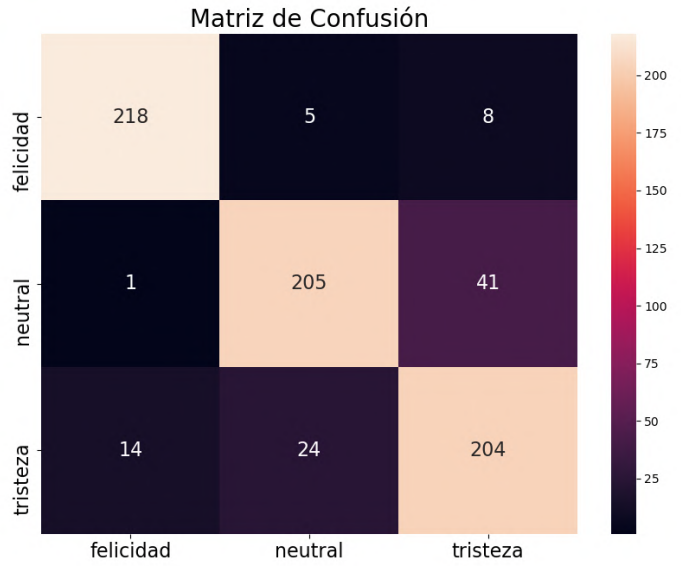

Figura 29: Matriz de confusión del sistema FER (Autoría propia, 2024).

4.2. Sistema SER

Por su parte, el sistema SER obtuvo una precisión del 99.90 % a partir de su entrenamiento con la base de datos MESD, con una pérdida del 0.04 %, permitiendo clasificar las emociones de felicidad, neutral y tristeza de manera correcta. El entrenamiento de este modelo fue considerablemente más rápido pese a contener más capas que el modelo FER, completando su entrenamiento en 2.97 minutos. La Figura 30 muestra la evolución de la precisión durante el entrenamiento y la validación.

A diferencia del sistema FER, los valores de la pérdida durante la validación se comportan de manera similar al entrenamiento (Figura 31), por lo que el modelo para este sistema presenta un mejor comportamiento.

Mientras que la matriz de confusión del modelo muestra que el modelo tiende a confundir los estados de tristeza y neutral sobre la felicidad (Figura 32).

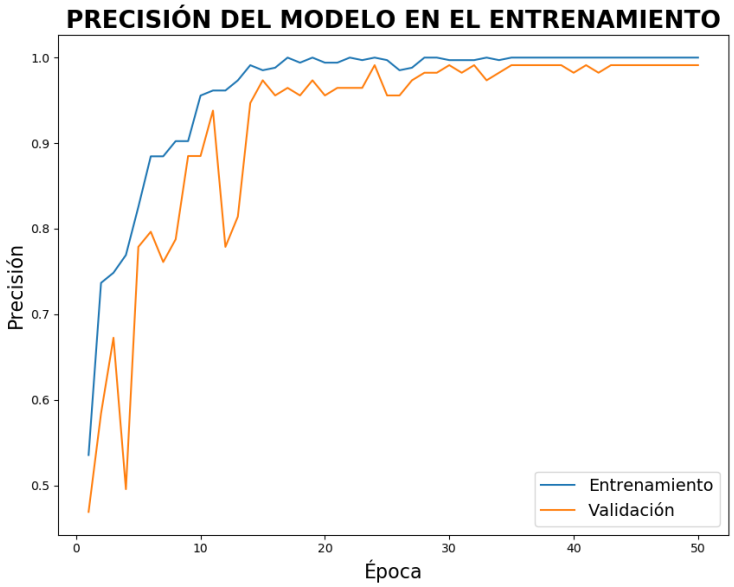

Figura 30: Evolución de la precisión del sistema SER (Autoría propia, 2024).

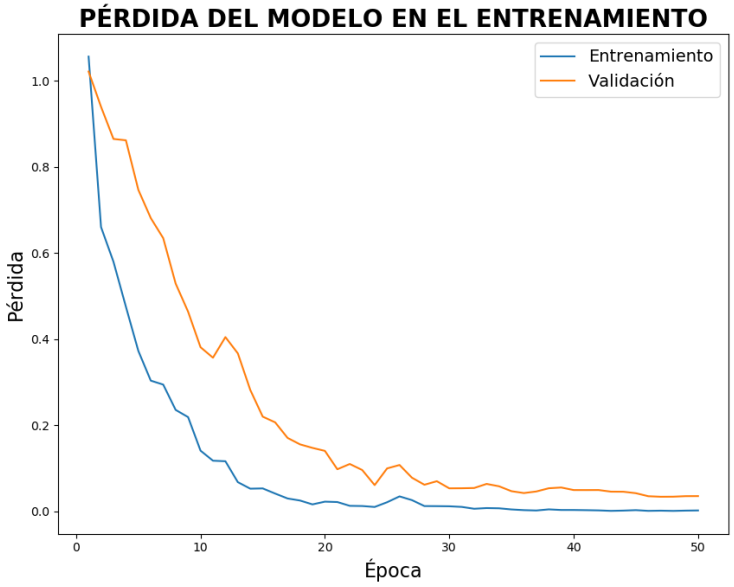

Figura 31: Evolución de la pérdida del sistema SER (Autoría propia, 2024).

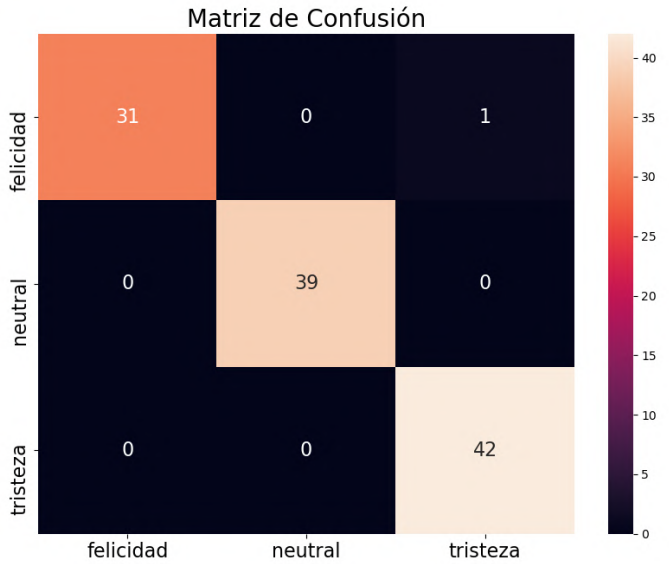

Figura 32: Matriz de confusión del sistema SER (Autoría propia, 2024).

4.3. Calibración

Para la calibración de los datos de la MESD, el sujeto 1 y el sujeto 2 se obtuvieron los valores máximos, mínimos y delta que se muestran en la Tabla 9.

Tabla 9: Datos de calibración.

Emoción	**Valor**	**MESD**	**Sujeto 1**	**Sujeto 2**
	I_{max}	210.5332	871.1627	965.2605
Felicidad	I_{min}	-543.6235	-168.5501	-195.1136
	δ_I	754.1567	1039.7129	1156.2680
	I_{max}	230.9116	789.4926	551.2440
Neutral	I_{min}	-542.8857	-17.0569	-21.3071
	δ_I	773.7974	806.5496	572.5515
	I_{max}	195.2069	888.9490	864.3171
Tristeza	I_{min}	-447.6268	-76.3115	-64.3744
	δ_I	642.5338	965.2605	928.6916

Se puede observar que la intensidad emocional de la felicidad y el estado neutral para la información de la MESD son similares, mientras que el rango de la tristeza es menor. Mientras que para los sujetos uno y dos se tiene que los rangos de tristeza, felicidad y neutral son diferenciables entre ellos.

4.4. Modelo de clasificación CONCAT

El modelo de concatenación de datos del sistema SER Y FER, o modelo CONCAT, pudo obtener una precisión del 100 % y una pérdida del 0.02 %, superando los resultados de los sistemas FER y SER de manera individual. El entrenamiento de este modelo tuvo una duración de 10.34 minutos, pero alcanzó las mejores métricas de rendimiento de los tres sistemas.

Como se puede observar en la Figura 33, el sistema fue capaz de alcanzar precisiones cercanas al 100 % en menos de 50 épocas, alcanzando la estabilidad antes de finalizar su entrenamiento.

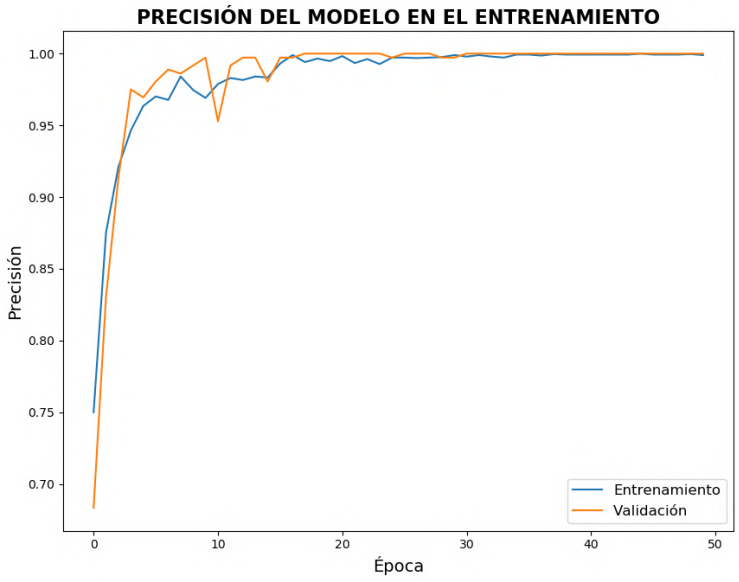

Figura 33: Evolución de la precisión del modelo de Concatenación (Autoría propia, 2024).

Mientras que en la Figura 34 se puede observar que la pérdida durante el entrenamiento y la validación disminuyó a valores cercanos a cero con un comportamiento similar en cuanto al número de épocas descrito para la precisión.

Y en la matriz de confusión (Figura 35) se puede observar que el sistema puede reconocer de manera precisa cada una de las categorías sin obtener falsos positivos.

Al realizar el reentrenamiento del modelo de concatenación como parte de la calibración del sistema, se obtienen las métricas de rendimiento que se muestran en la Tabla 10.

Se puede observar que la precisión del modelo disminuye con respecto al modelo de concatenación original; sin embargo, no es una diferencia considerable. Por otra parte, la pérdida en la validación aumenta de manera considerable a comparación del modelo original.

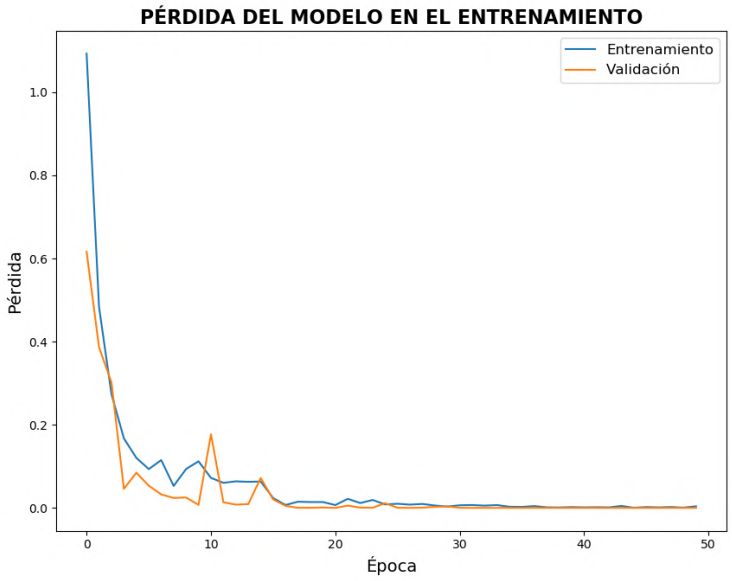

Figura 34: Evolución de la pérdida del modelo de Concatenación (Autoría propia, 2024).

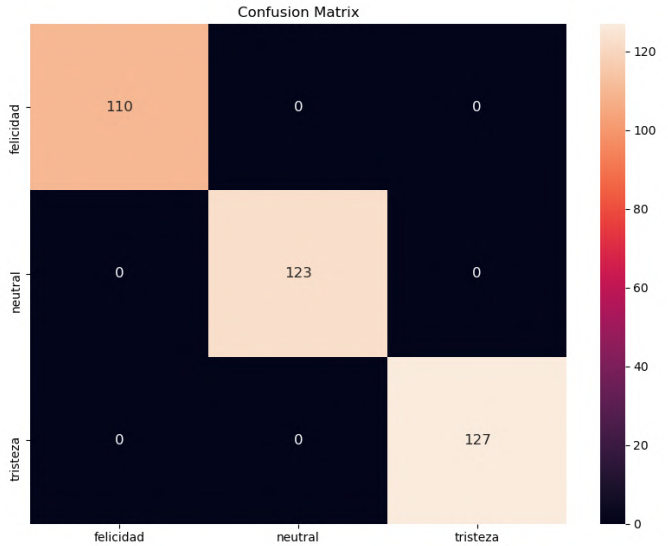

Figura 35: Matriz de confusión del modelo de Concatenación (Autoría propia, 2024).

Tabla 10: Métricas de rendimiento de los modelos reentrenados.

Base de datos	CMF	CMF + Sujeto 1	CMF + Sujeto 2
Precisión (%)	99.95	98.23	99.76
Pérdida (%)	0.02	5.64	1.77
Precisión val (%)	100	95.28	92.59
Pérdida val	1.40E-06	24.22	34.73
Exactitud (%)	100	93.27	90.83
Recall (%)	100	92.45	89.81
F1-Score (%)	100	92.23	89.99
Tiempo (min)	10.34	4.89	4.98

4.5. Cuantificación del estado emocional

Para la cuantificación del estado emocional se tiene que, dentro de la información utilizada de la base de datos MESD se excluyó un archivo de audio en el momento del entrenamiento, por lo que al realizar los cálculos de δ_J, n_e y q esta información se considera como información nueva. El proceso de selección se hace a través de una ventana generada por medio de Python y Tkinter únicamente para visualizar de mejor manera la información, donde se selecciona una imagen que corresponde a la felicidad y dicho audio que se omitió al realizar el entrenamiento y calibración del modelo (Figura 36).

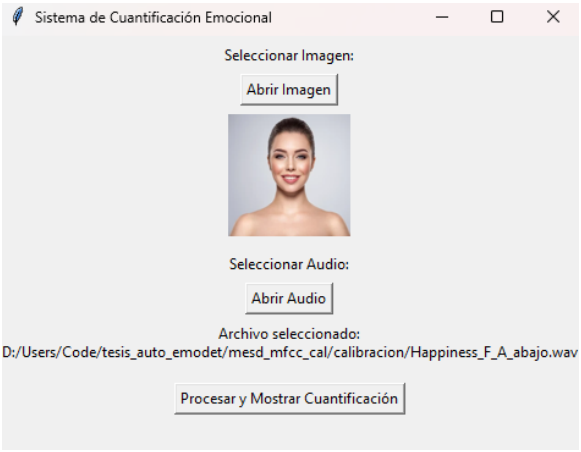

Figura 36: Ventana de selección de datos (Autoría propia, 2024).

Una vez seleccionados, se procesa la información, se obtiene la predicción del modelo de concatenación. Dando como resultado la emoción felicidad para la clasificación. Mientras que para la cuantificación se obtiene un valor alto para este estado emocional, esto se muestra con la información de la Tabla 11.

El proceso de cuantificación para las grabaciones de las sesiones de intervención se realiza analizando los archivos correspondientes, donde la Tabla 12 muestra el valor medio de δ_I,

Tabla 11: Resultados de la cuantificación para la base de datos MESD.

Parámetro	**Valor**
Emoción	Felicidad
n_e	73
q	alto
d_I	754.157
d_J	555.872
J_{max}	175.567
J_{min}	-380.305

δ_J, así como de J_{max} y J_{min} calculado a partir de todas las cuantificaciones realizadas por el modelo para el sujeto 1, para cada una de las categorías emocionales.

Tabla 12: Datos de cuantificación del sujeto 1.

Parámetro (μ)	**Felicidad**	**Neutral**	**Tristeza**
n_e	73.71	94.44	94.44
d_J	771.6663	765.6566	768.2747
J_{max}	178.2863	165.9940	181.2672
J_{min}	-593.3800	-599.6626	-587.0076

Mientras que valores medios para la cuantificación del estado emocional del sujeto 2 se observan en la Tabla 13. Para el sujeto 2 no se detectó la emoción de tristeza, por lo que no se cuenta con información para la cuantificación de dicha emoción.

Tabla 13: Datos de cuantificación del sujeto 2.

Parámetro (μ)	**Felicidad**	**Neutral**	**Tristeza**
n_e	72.80	100.00	0.00
d_J	848.7888	765.6566	0.0000
J_{max}	129.7702	74.2333	0.0000
J_{min}	-719.0186	-565.6804	0.0000

4.6. Sistema embebido

La carcasa del sistema embebido se generó a partir de impresión 3D, utilizando filamento de PLA como material de impresión en una Flashforge Guider II (Figura 37).

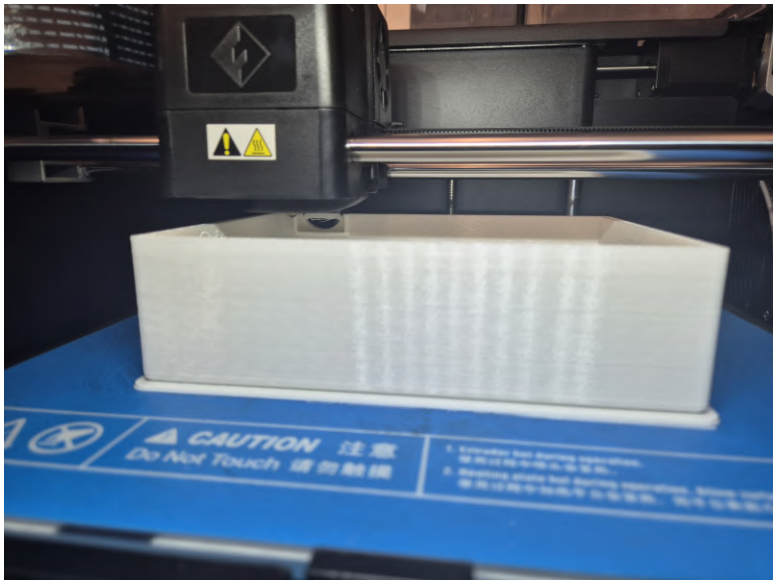

Figura 37: Impresión de la carcasa del sistema embebido (Autoría propia, 2024).

La Tabla 14 muestra las características finales del sistema embebido.

Tabla 14: Características del sistema embebido.

Característica	**Valor**	**Unidad**
Alto	130	mm
Ancho	120	mm
Largo	180	mm
Masa	500	g

El sistema embebido (Figura 38) contiene todos los elementos necesarios para hacer la detección y cuantificación del estado emocional a través de las grabaciones de las sesiones de intervención psicológica.

Además, permite visualizar la emoción detectada, así como la cuantificación de su nivel emocional a lo largo del proceso (Figura 39).

Una vez finalizado el análisis del nivel emocional, el sistema es capaz de proveer gráficas de información, donde se muestra la recurrencia emocional a lo largo de la grabación, así como la distribución de los niveles emocionales para cada categoría.

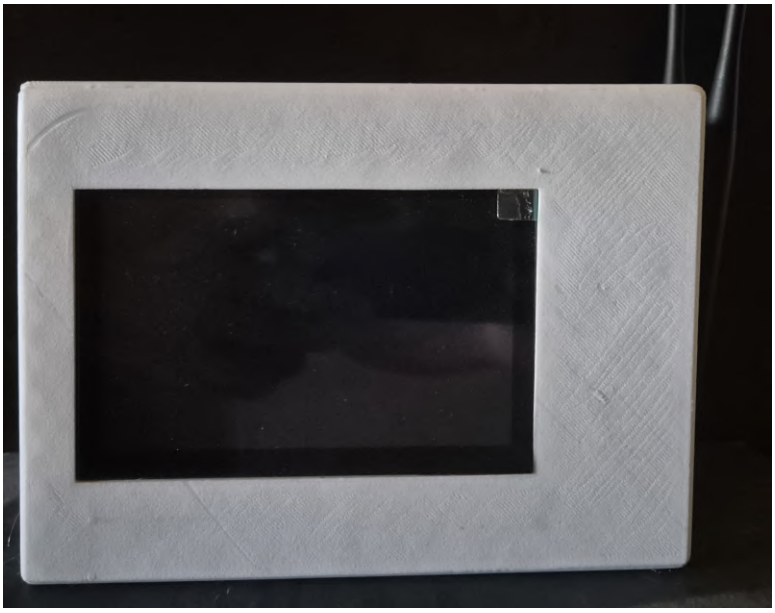

Figura 38: Sistema embebido (Autoría propia, 2024).

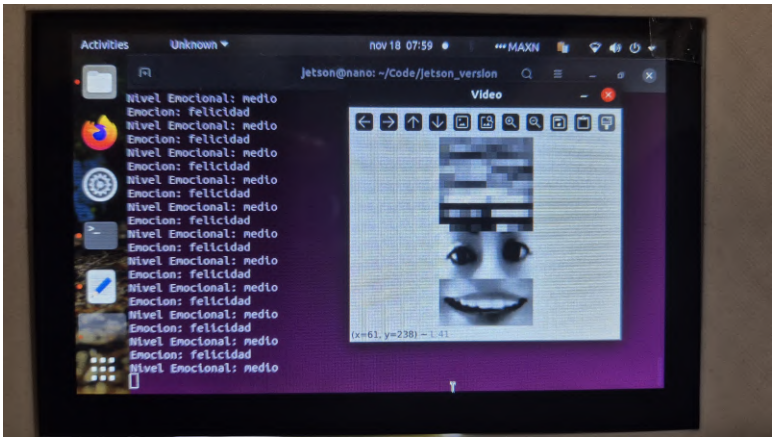

Figura 39: Proceso de detección y cuantificación en el sistema embebido (Autoría propia, 2024).

4.7. Validación del sistema embebido

El sistema embebido es capaz de generar las imágenes requeridas por el modelo CONCAT, a partir del análisis de las grabaciones de las sesiones de intervención psicológica para los sujetos 1 y 2. Esto permite obtener los valores necesarios para la cuantificación del estado emocional al integrar nueva información. Durante el análisis, el sistema muestra el estado emocional en una interfaz de video que muestra la información procesada.

La Figura 40 muestra la detección del estado emocional de felicidad, con un nivel emocional alto, para el sujeto 1, a través de la ventana de la interfaz.

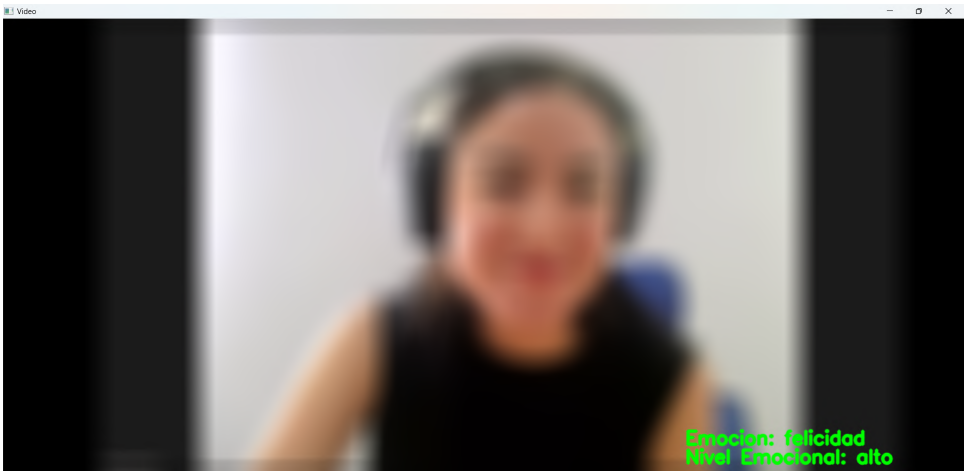

Figura 40: Identificación de felicidad en nivel alto para el sujeto 1 (Autoría propia, 2024).

Por otra parte, dentro de la misma sesión analizada, el modelo detectó la emoción tristeza con un nivel medio (Figura 41). También se detectó el estado neutral con un nivel alto (Figura 42).

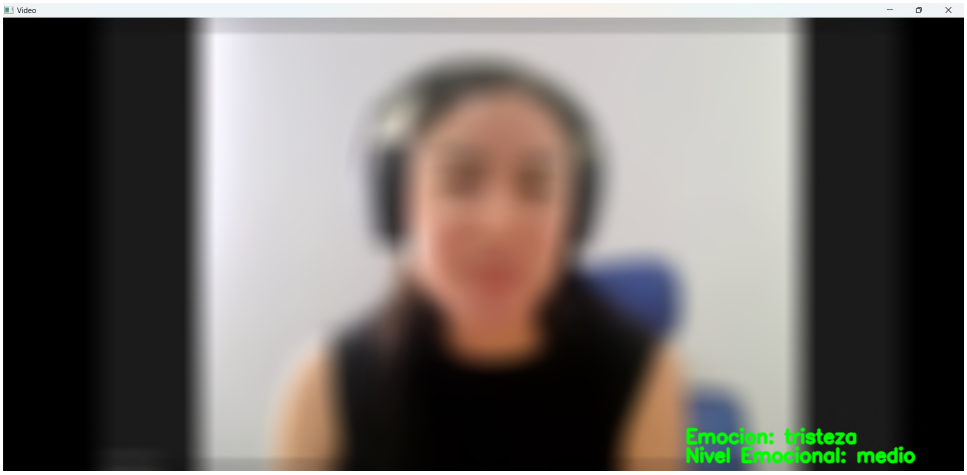

Figura 41: Identificación de tristeza en nivel medio para el sujeto 1 (Autoría propia, 2024).

Las imágenes de la interfaz dentro de este documento fueron modificadas para salvaguardar la privacidad del sujeto; sin embargo, el sistema muestra las imágenes reales al momento del análisis.

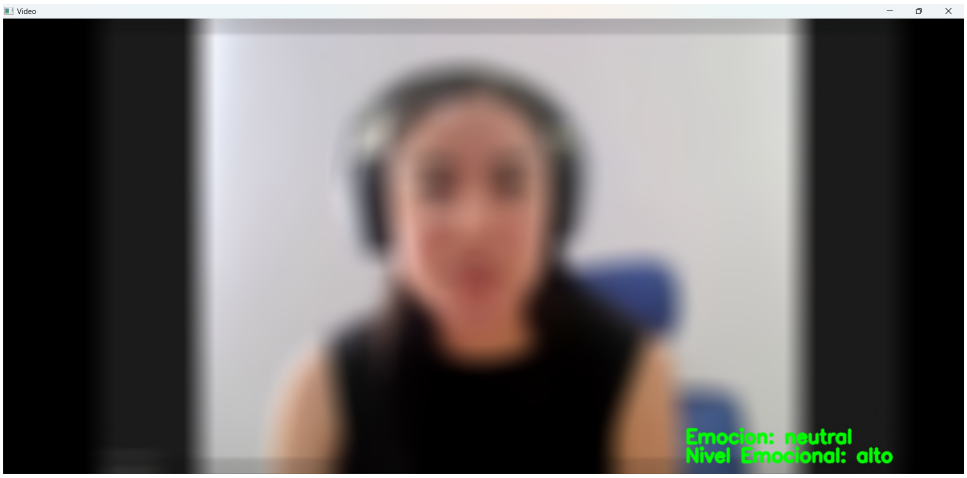

Figura 42: Identificación del estado neutral en nivel alto para el sujeto 1 (Autoría propia, 2024).

La Figura 43 muestra la imagen de entrada al modelo CONCAT, que incluye la transformación de un segmento de audio de 400 ms en una imagen de frecuencias, junto con la imagen de las zonas de interés del rostro utilizadas para la detección de emociones.

Figura 43: Imagen de entrada del modelo CONCAT (Autoría propia, 2024).

Además, se comparan los resultados de la detección y cuantificación de las emociones felicidad y tristeza y el estado neutral a partir de una versión del modelo CONCAT sin el reentrenamiento con la información de los sujetos. La Tabla 15 muestra los resultados del porcentaje de la recurrencia de las emociones para los modelos reentrenados para ambos sujetos y el modelo CONCAT base.

Tabla 15: Porcentaje de emociones detectadas durante el análisis.

Emoción	**CONCAT**	**CONCAT + Sujeto 1**	**CONCAT + Sujeto 2**
felicidad	99.43	70.37	28.72
neutral	0.07	11.22	0.00
tristeza	0.49	18.40	71.27

La Figura 44 muestra la recurrencia emocional detectada por el modelo CONCAT base.

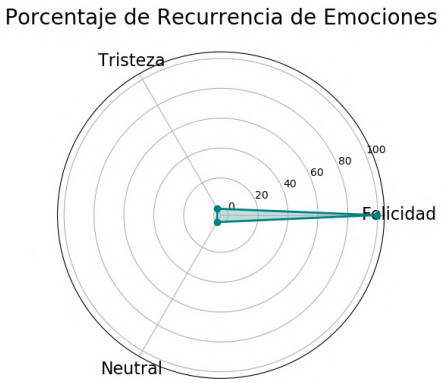

Figura 44: Emociones detectadas por el modelo base (Autoría propia, 2024).

Mientras que la Figura 45 muestra los niveles emocionales cuantificados por el modelo base.

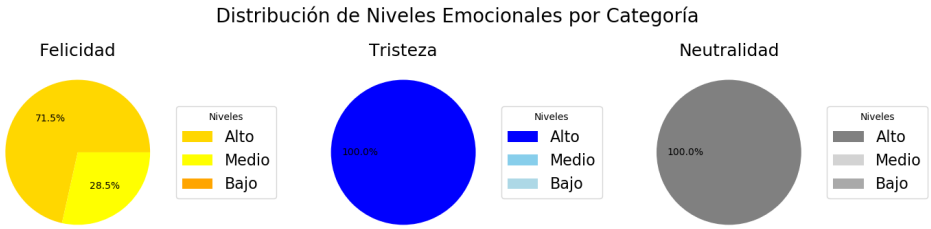

Figura 45: Distribución de niveles emocionales del modelo base (Autoría propia, 2024).

En contraste, el modelo CONCAT reentrenado con la información del sujeto 1 (Figura 46) obtiene una mayor variación en la detección emocional a lo largo de las grabaciones.

Mientras que la Figura 47 muestra la distribución de los niveles emocionales para cada categoría.

A partir de esto se puede observar que para el sujeto 1 la emoción predominante durante el análisis es la felicidad, y el nivel emocional para esta emoción es mayormente alto. Por

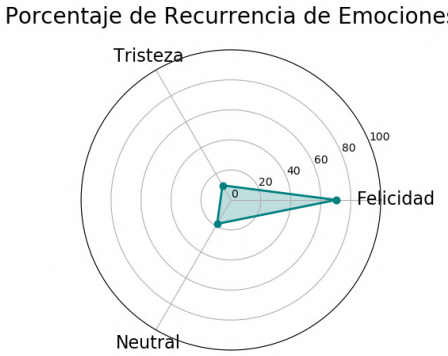

Figura 46: Emociones detectadas para el sujeto 1 (Autoría propia, 2024).

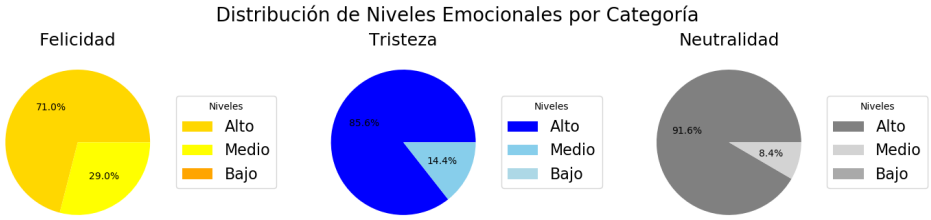

Figura 47: Distribución de niveles emocionales del sujeto 1 (Autoría propia, 2024).

otra parte, la tristeza se presenta en menor medida, pero también en niveles altos, al igual que el estado neutral.

Las emociones exhibidas por el sujeto 2 demuestran que presenta un estado neutral durante la mayor parte de la intervención, mientras que la emoción de tristeza no se presenta durante el análisis, al utilizar el modelo CONCAT reentrenado con la información de dicho sujeto (Figura 48).

Mientras que, para la emoción de felicidad, el nivel predominante fue el alto durante la intervención, pese a que dicha emoción se presentó en menor medida que el estado neutral (Figura 49).

Es importante mencionar que el sistema puede ser interrumpido al cerrar el video. Sin embargo, la información de detección y cuantificación de emociones que se procesó hasta ese momento permite generar las gráficas de recurrencia y distribución, asegurando que el análisis no se vea afectado.

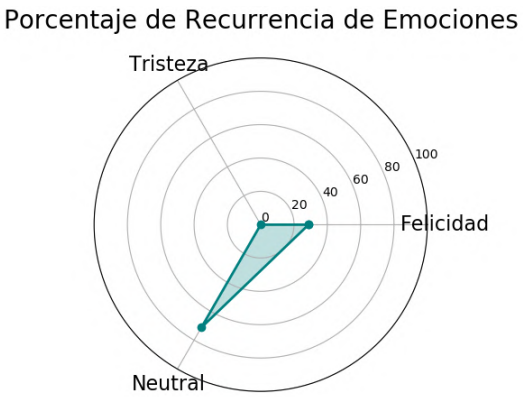

Figura 48: Emociones detectadas para el sujeto 2 (Autoría propia, 2024).

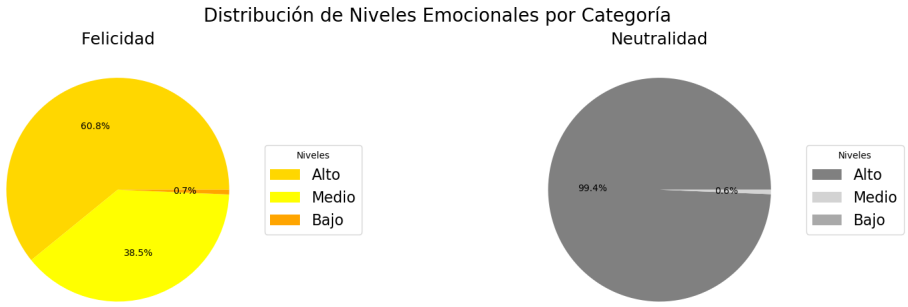

Figura 49: Distribución de niveles emocionales del sujeto 2 (Autoría propia, 2024).

5. Conclusiones

A partir de la combinación de los sistemas de reconocimiento de expresiones faciales y de reconocimiento de emociones a través de la voz se logró desarrollar un sistema embebido que permite detección y cuantificación de las emociones de felicidad, tristeza y el estado neutral basado en el uso de redes neuronales convolucionales.

A través del uso del identificador de rostros de Mediapipe para generar las regiones de interés, las herramientas de visión por computadora para el preprocesamiento de la información y el entrenamiento por medio de una CNN se logró desarrollar un modelo del sistema FER. Sin embargo, durante el entrenamiento se encontraron múltiples problemas debido a las características de las imágenes de la base de datos FER-2013, por ello, únicamente se consideraron las imágenes frontales para el desarrollo de esta investigación. Por otra parte, el entrenamiento del sistema FER de manera individual permitió obtener valores aceptables en la precisión, sin embargo, la diferencia entre la pérdida del entrenamiento y la pérdida de la valoración es alta, con una diferencia del 27.2 %.

Utilizando los coeficientes cepstrales en la escala de Mel y las herramientas de visión por computadora se desarrolló un sistema SER capaz de identificar emociones basado en una arquitectura de CNN. De manera similar al comportamiento del sistema FER, durante el entrenamiento de forma individual el sistema SER fue capaz de obtener valores de precisión adecuados, sin embargo, la diferencia entre pruebas y validación de la pérdida es considerable, por lo que el modelo tiende a confundir los datos.

Sin embargo, al combinar los sistemas a partir de la concatenación de las imágenes de ambos modelos se obtuvo un sistema más robusto, con una arquitectura más simple a comparación de los modelos individuales, y que permite que el sistema converja obteniendo una precisión del 99.95 %, una pérdida de 0.02 % y en un tiempo de entrenamiento de 10.34 minutos. Este sistema presenta un mejor comportamiento tanto en el entrenamiento como en la validación para la precisión y la pérdida.

Las métricas obtenidas a través de esta metodología basada en la concatenación de los datos demuestran que es funcional para la detección de emociones a través de las expresiones faciales y la voz. Esta metodología de concatenación de datos fue base para la obtención del producto *Emotion recognition based on facial gestures and Convolutional Neural Networks*, presentado en el congreso internacional *The 2024 IEEE 3rd Conference on Information Technology and Data Science* (Anexo 7.7). Además, es importante mencionar que la arquitectura simple de la CNN utilizada en el desarrollo de esta investigación permite obtener mejores resultados a comparación de modelos utilizados dentro de la literatura donde si bien alcanzan métricas representativas, algunos de estos modelos utilizan en adición a su arquitectura propia, modelos preentrenados para la detección de emociones, haciendo su configuración más compleja. Mientras que la calibración del modelo a partir de los datos de intensidad generados por el sistema SER permiten dar un rango de intensidad valido para la posterior cuantificación del nivel emocional. El uso de múltiples archivos de audio para encontrar los niveles máximos y mínimos permite encontrar un rango de intensidad con la información suficiente.

El reentrenamiento del modelo de clasificación a partir de la información de cada uno de

los individuos permite mejorar la detección y cuantificación de los estados emocionales ya que considera las características de éstos.

Por otra parte, la cuantificación del estado emocional, a partir de los cálculos de los valores máximos y mínimos de la intensidad, permitió obtener un rango emocional para cada uno de los sujetos evaluados, tanto de la información de la base de datos MESD, como de las sesiones de intervención psicológica. El proceso de cuantificación está completamente ligado a la calibración del sistema a partir de los datos de la frecuencia porque si bien, no es posible dar un nivel exacto del estado emocional, el sistema puede informar un nivel emocional relativo a la información de máximos y mínimos que se tenían previamente.

Por otra parte, se generó un diseño capaz de proteger los elementos internos del sistema y la visualización de la información considerando las Normas Oficiales Mexicanas mencionadas anteriormente, así como las consideraciones del manejo de desecho de filamento y el uso de las instalaciones del Laboratorio de Visión Artificial de la Universidad Autónoma de Querétaro.

El sistema embebido fue capaz de realizar la detección y cuantificación de los tres estados emocionales, sin embargo, al requerir un costo computacional alto, no es posible generar el procesamiento en línea de información adquirida por el sistema. A través de la implementación del sistema FER y SER dentro de un sistema embebido se pudo desarrollar un modelo de clasificación basado en una CNN capaz de clasificar emociones de manera correcta, este modelo puede ser calibrado a partir de la información de un sujeto de modo que se ajuste a sus niveles de intensidad para obtener una cuantificación del nivel emocional.

5.1. Prospectivas

Aunque el sistema es capaz de identificar y cuantificar de forma correcta dos estados emocionales y el estado neutral, se trabajará en un futuro para que sea capaz de identificar las seis emociones básicas y el estado neutral. Además, realizar el entrenamiento y validación de los sistemas FER, SER y de Concatenación con este aumento de estados emocionales.

Enfocar el procesamiento del sistema embebido a la nube, con la finalidad de poder realizar el entrenamiento de los modelos en un menor tiempo debido al costo computacional y que permita una detección en tiempo real.

Generar la calibración y cuantificación para los seis estados emocionales básicos y el estado neutral a partir de los niveles de intensidad. Adicionar el uso de una batería de tal modo que el dispositivo pueda ser portátil y así, pueda ser utilizado en cualquier entorno, sin la necesidad de una conexión eléctrica. Además, mejorar el diseño de la carcasa del sistema embebido de tal forma que se consideren aspectos de la ergonomía para su uso y la sujeción a un tripoide.

Realizar la validación del sistema en un entorno real, con la finalidad de adaptar el sistema a diferentes condiciones de ruido e iluminación para aumentar la robustez del sistema.

6. Referencias

Abbaschian, B. J., Sierra-Sosa, D. & Elmaghraby, A. (2021). Deep Learning Techniques for Speech Emotion Recognition, from Databases to Models. *Sensors*, *21*(4), 1249. https://doi.org/10.3390/s21041249

Abdul, Z. K. & Al-Talabani, A. K. (2022). Mel Frequency Cepstral Coefficient and its Applications: A Review. *IEEE Access*, *10*, 122136-122158. https://doi.org/10.1109/access.2022.3223444

Albekairi, M., Kaaniche, K., Abbas, G., Mercorelli, P., Alanazi, M. D. & Almadhor, A. (2024). Advanced Neural Classifier-Based Effective Human Assistance Robots Using Comparable Interactive Input Assessment Technique. *Mathematics*, *12*(16), 2500. https://doi.org/10.3390/math12162500

Apatean, A., Emerich, S. & Lupu, E. (2009). Emotions Recognition By Speech And Facial Expressions Analysis. *17th European Signal Processing Conference (EUSIPCO 2009)*. https://doi.org/10.5281/ZENODO.41698

Begazo, R., Aguilera, A., Dongo, I. & Cardinale, Y. (2024). A Combined CNN Architecture for Speech Emotion Recognition. *Sensors*, *24*(17), 5797. https://doi.org/10.3390/s24175797

Bhagat, D., Vakil, A., Gupta, R. K. & Kumar, A. (2024). Facial Emotion Recognition (FER) using Convolutional Neural Network (CNN). *Procedia Computer Science*, *235*, 2079-2089. https://doi.org/10.1016/j.procs.2024.04.197

Brener, S. A., Frankenhuis, W. E., Young, E. S. & Ellis, B. J. (2023). Social Class, Sex, and the Ability to Recognize Emotions: The Main Effect is in the Interaction. *Personality and Social Psychology Bulletin*, *50*(8), 1197-1210. https://doi.org/10.1177/01461672231159775

Chai, J., Zeng, H., Li, A. & Ngai, E. W. (2021). Deep learning in computer vision: A critical review of emerging techniques and application scenarios. *Machine Learning with Applications*, *6*, 100134. https://doi.org/10.1016/j.mlwa.2021.100134

Coto-Jiménez, M., Martínez-Licona, F. M. & Goddard-Close, J. (2014). Acoustic Vowel Analysis in a Mexican Spanish HMM-based Speech Synthesis. *Research in Computing Science*, *86*(1), 53-62. https://doi.org/10.13053/rcs-86-1-4

Datta Rakshith, K., Rudresh, M. & Shashibhushsan, G. (2021). Comparative performance analysis for speech digit recognition based on MFCC and vector quantization. *Global Transitions Proceedings*, *2*(2), 513-519. https://doi.org/10.1016/j.gltp.2021.08.013

D.O.F. (2010). *Ley Federal de Protección de Datos Personales en Posesión de los Particulares*. Diario Oficial de la Federación. https://www.diputados.gob.mx/LeyesBiblio/pdf/LFPDPPP.pdf

D.O.F. (2014). *Reglamento de la Ley General de Salud en Materia de Investigación para la Salud*. Diario Oficial de la Federación. https://www.diputados.gob.mx/LeyesBiblio/regley/Reg_LGS_MIS.pdf

D.O.F. (2021). *NORMA Oficial Mexicana NOM-241-SSA1-2021, Buenas prácticas de fabricación de dispositivos médicos. (NOM-241-SSA1-2021)*. Diario Oficial de la Federa-

ción. https://dof.gob.mx/nota_detalle.php?codigo=5638793&fecha=20/12/2021#gsc.tab=0

Duville, M. M., Alonso-Valerdi, L. M. & Ibarra-Zarate, D. I. (2021). The Mexican Emotional Speech Database (MESD): elaboration and assessment based on machine learning. *2021 43rd Annual International Conference of the IEEE Engineering in Medicine; Biology Society (EMBC)*. https://doi.org/10.1109/embc46164.2021.9629934

Ekman, P. & Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, *17*(2), 124-129. https://doi.org/10.1037/h0030377

El Ayadi, M., Kamel, M. S. & Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, *44*(3), 572-587. https://doi.org/10.1016/j.patcog.2010.09.020

Eleyan, A. & Demirel, H. (2007). PCA and LDA Based Neural Networks for Human Face Recognition. *Face Recognition*. I-Tech Education; Publishing. https://doi.org/10.5772/4833

Fernandes, J. V. M. R., Alexandria, A. R. d., Marques, J. A. L., Assis, D. F. d., Motta, P. C. & Silva, B. R. d. S. (2024). Emotion Detection from EEG Signals Using Machine Deep Learning Models. *Bioengineering*, *11*(8), 782. https://doi.org/10.3390/bioengineering11080782

Fernandes, V., Mascarehnas, L., Mendonca, C., Johnson, A. & Mishra, R. (2018). Speech Emotion Recognition using Mel Frequency Cepstral Coefficient and SVM Classifier. *2018 International Conference on System Modeling &; Advancement in Research Trends (SMART)*, 200-204. https://doi.org/10.1109/sysmart.2018.8746939

Ferreiros, J., Macias-Guarasa, J., Pardo, J. & Villarrubia, L. (1998). *Introducing multiple pronunciations in Spanish speech recognition systems*. Proc. Modeling Pronunciation Variation.

García Guajardo, J. G. (2011). *Sistema de reconocimiento de voz usando perceptrón multicapa y Coeficientes Cepstrales de Mel* (Tesis de maestría). Universidad Autónoma de Querétaro.

González, D. A., Reséndiz, A. & Reyes, I. (2015). Adaptation of the BDI–II in Mexico. *Salud mental*, *38*(4), 237-244. https://doi.org/10.17711/sm.0185-3325.2015.033

Goodfellow, I., Bengio, Y. & Courville, A. (2016). *Deep Learning*. MIT Press. http://www.deeplearningbook.org

Guo, Y., Jia, X., Zhao, B., Chai, H. & Huang, Y. (2020). Multifeature extracting CNN with concatenation for image denoising. *Signal Processing: Image Communication*, *81*, 115690. https://doi.org/10.1016/j.image.2019.115690

Gursesli, M. C., Lombardi, S., Duradoni, M., Bocchi, L., Guazzini, A. & Lanata, A. (2024). Facial Emotion Recognition (FER) Through Custom Lightweight CNN Model: Performance Evaluation in Public Datasets. *IEEE Access*, *12*, 45543-45559. https://doi.org/10.1109/access.2024.3380847

Hashim, S. & Mccullagh, P. (2023). Face detection by using Haar Cascade Classifier. *Wasit Journal of Computer and Mathematics Science*, *2*(1), 1-5. https://doi.org/10.31185/wjcm.109

Hiroyuki, K. & Qiangfu, Z. (2007). Face detection with clustering, lda and NN. *2007 IEEE International Conference on Systems, Man and Cybernetics*. https://doi.org/10.1109/icsmc.2007.4413760

INEGI. (2017). *Porcentaje de los integrantes del hogar de 12 años y más que se ha sentido deprimido*. Instituto Nacional de Estadística y Geografía. https://www.inegi.org.mx/app/indicadores/?t=148&ag=00

Ioffe, S. & Szegedy, C. (2015). Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. https://doi.org/10.48550/ARXIV.1502.03167

Jeremías-Ambrogio, E. (2020). Reconocimiento de objetos a través de la metodología Haar Cascades. Revista Argentina de Ingeniería.

Lecun, Y., Bottou, L., Bengio, Y. & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278-2324. https://doi.org/10.1109/5.726791

Li, S., Guo, L. & Liu, J. (2022). Towards East Asian Facial Expression Recognition in the Real World: A New Database and Deep Recognition Baseline. *Sensors*, *22*(21), 8089. https://doi.org/10.3390/s22218089

Lih-Heng, C., Sh-Hussain, S. & Chee-Ming, T. (2009). PCA, LDA and neural network for face identification. *2009 4th IEEE Conference on Industrial Electronics and Applications*. https://doi.org/10.1109/iciea.2009.5138403

Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z. & Matthews, I. (2010). The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*. https://doi.org/10.1109/cvprw.2010.5543262

Lyons, M., Akamatsu, S., Kamachi, M. & Gyoba, J. (1998). Coding facial expressions with Gabor wavelets. *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*. https://doi.org/10.1109/afgr.1998.670949

Martínez-Mascorro, G. A. & Aguilar-Torres, G. (2013). Reconocimiento de voz basado en MFCC, SBC y Espectrogramas. *Ingenius*, (10). https://doi.org/10.17163/ings.n10.2013.02

Matveev, Y., Matveev, A., Frolova, O., Lyakso, E. & Ruban, N. (2022). Automatic Speech Emotion Recognition of Younger School Age Children. *Mathematics*, *10*(14), 2373. https://doi.org/10.3390/math10142373

Mediapipe. (2023). *MediaPipe Face Mesh*. Google. https://github.com/google-ai-edge/mediapipe/blob/master/docs/solutions/face_mesh.md

Mellouk, W. & Handouzi, W. (2020). Facial emotion recognition using deep learning: review and insights. *Procedia Computer Science*, *175*, 689-694. https://doi.org/10.1016/j.procs.2020.07.101

Mohan, K., Seal, A., Krejcar, O. & Yazidi, A. (2021). FER-net: facial expression recognition using deep neural net. *Neural Computing and Applications*, *33*(15), 9125-9136. https://doi.org/10.1007/s00521-020-05676-y

NVIDIA. (2019). *Developer Kit Setup and Hardware*. NVIDIA Developer. https://developer.nvidia.com/embedded/jetson-nano-developer-kit

Padrós Blázquez, F., Montoya Pérez, K. S., Bravo Calderón, M. A. & Martínez Medina, M. P. (2020). Propiedades psicométricas del Inventario de Ansiedad de Beck (BAI, Beck Anxiety Inventory) en población general de México. *Ansiedad y Estrés*, *26*(2–3), 181-187. https://doi.org/10.1016/j.anyes.2020.08.002

Pan, R., García-Díaz, J. A., Rodríguez-García, M. Á. & Valencia-García, R. (2024). Spanish MEACorpus 2023: A multimodal speech–text corpus for emotion analysis in Spanish from natural environments. *Computer Standards and; Interfaces*, *90*, 103856. https://doi.org/10.1016/j.csi.2024.103856

Qamar, R. & Ali Zardari, B. (2023). Artificial Neural Networks: An Overview. *Mesopotamian Journal of Computer Science*, 130-139. https://doi.org/10.58496/mjcsc/2023/015

Ramos, O. L., Rojas, D. A. & Góngora, L. A. (2016). Reconocimiento de patrones de habla usando MFCC y RNA. *Visión electrónica*, *10*(1), 5-11. https://doi.org/10.14483/22484728.11712

Revina, I. & Emmanuel, W. S. (2021). A Survey on Human Face Expression Recognition Techniques. *Journal of King Saud University - Computer and Information Sciences*, *33*(6), 619-628. https://doi.org/10.1016/j.jksuci.2018.09.002

Ristea, N.-C., Dutu, L. C. & Radoi, A. (2019). Emotion Recognition System from Speech and Visual Information based on Convolutional Neural Networks. *2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*. https://doi.org/10.1109/sped.2019.8906538

Samane Sharifi, M. & Rada, L. (2024). Understanding Artificial Neural Networks: A Comprehensive Review. *Unpublished*. https://doi.org/10.13140/RG.2.2.34372.85121

Saurav, S., Gidde, P., Saini, R. & Singh, S. (2021). Dual integrated convolutional neural network for real-time facial expression recognition in the wild. *The Visual Computer*, *38*(3), 1083-1096. https://doi.org/10.1007/s00371-021-02069-7

Saxena, A. (2022). An Introduction to Convolutional Neural Networks. *International Journal for Research in Applied Science and Engineering Technology*, *10*(12), 943-947. https://doi.org/10.22214/ijraset.2022.47789

Shi, C., Tan, C. & Wang, L. (2021). A Facial Expression Recognition Method Based on a Multibranch Cross-Connection Convolutional Neural Network. *IEEE Access*, *9*, 39255-39274. https://doi.org/10.1109/access.2021.3063493

Shreya, M. S., Indrayani, S. P., Aniket, P. S., Dipali, V. L., Kalyani, A. S. & Harsha, R. V. (2023). A Review Paper on Computer Vision. *International Journal of Advanced Research in Science, Communication and Technology*, 673-677. https://doi.org/10.48175/ijarsct-8901

Szeliski, R. (2011). *Computer Vision: Algorithms and Applications*. Springer London. https://doi.org/10.1007/978-1-84882-935-0

Takahashi, K. (2004). Remarks on emotion recognition from multi-modal bio-potential signals. *2004 IEEE International Conference on Industrial Technology, 2004. IEEE ICIT '04.* https://doi.org/10.1109/icit.2004.1490720

Ullah, R., Asif, M., Shah, W. A., Anjam, F., Ullah, I., Khurshaid, T., Wuttisittikulkij, L., Shah, S., Ali, S. M. & Alibakhshikenari, M. (2023). Speech Emotion Recognition Using

Convolution Neural Networks and Multi-Head Convolutional Transformer. *Sensors*, *23*(13), 6212. https://doi.org/10.3390/s23136212

Velasco Matus, P. W., Rivera Aragón, S., Domínguez Espinosa, A. d. C., Méndez Rangel, F. & Díaz Loving, R. (2021). Positive Affect/Negative Affect Scale for Mexicans (PANA-M): Evidences of Validity and Reliability. *Acta de Investigación Psicológica*, *11*(1), 95-113. https://doi.org/10.22201/fpsi.20074719e.2021.1.377

Venkatesan, R., Shirly, S., Selvarathi, M. & Jebaseeli, T. J. (2023). Human Emotion Detection Using DeepFace and Artificial Intelligence. *RAiSE-2023*, 37. https://doi.org/10.3390/engproc2023059037

Wang, Y., Yang, X. & Zou, J. (2013). Research of Emotion Recognition Based on Speech and Facial Expression. *TELKOMNIKA Indonesian Journal of Electrical Engineering*, *11*(1). https://doi.org/10.11591/telkomnika.v11i1.1873

Wood, L., Tan, Z., Stenbit, I., Bischof, J., Zhu, S., Chollet, F., Sreepathihalli, D., Sampath, R. et al. (2022). KerasCV.

Yurtay, Y., Demirci, H., Tiryaki, H. & Altun, T. (2024). Emotion Recognition on Call Center Voice Data. *Applied Sciences*, *14*(20), 9458. https://doi.org/10.3390/app14209458

Yustiawati, R., Husni, N. L., Evelina, E., Rasyad, S., Lutfi, I., Silvia, A., Alfarizal, N. & Rialita, A. (2018). Analyzing Of Different Features Using Haar Cascade Classifier. *2018 International Conference on Electrical Engineering and Computer Science (ICECOS)*. https://doi.org/10.1109/icecos.2018.8605266

Zhao, X., Wang, L., Zhang, Y., Han, X., Deveci, M. & Parmar, M. (2024). A review of convolutional neural networks in computer vision. *Artificial Intelligence Review*, *57*(4). https://doi.org/10.1007/s10462-024-10721-6

Zhao, X. & Zhang, S. (2016). A Review on Facial Expression Recognition: Feature Extraction and Classification. *IETE Technical Review*, *33*(5), 505-517. https://doi.org/10.1080/02564602.2015.1117403

Zhou, K., Sisman, B., Liu, R. & Li, H. (2022). Emotional voice conversion: Theory, databases and ESD. *Speech Communication*, *137*, 1-18. https://doi.org/10.1016/j.specom.2021.11.006

Zhou, K., Sisman, B., Rana, R., Schuller, B. W. & Li, H. (2023). Emotion Intensity and its Control for Emotional Voice Conversion. *IEEE Transactions on Affective Computing*, *14*(1), 31-48. https://doi.org/10.1109/taffc.2022.3175578

7. Anexos

7.1. Carta de Consentimiento Informado

Universidad Autónoma de Querétaro
Facultad de Ingeniería

CARTA DE CONSENTIMIENTO INFORMADO

San Juan del Río, Querétaro, a _____ de _____ del año 20_____

Antes de expresar su consentimiento y, en su caso, aceptar participar, le instamos a leer detenidamente este documento. Por favor, no dude en plantear cualquier pregunta o inquietud que surja, a fin de garantizar una comprensión completa de los objetivos, procedimientos y resultados del estudio, incluyendo posibles riesgos y beneficios.

Propósito Principal del estudio: Desarrollar un sistema de detección y cuantificación de emociones en población mexicana adulta mediante el análisis de expresiones faciales y reconocimiento de voz.

Beneficios: Es importante destacar que esta investigación no persigue ningún beneficio económico, ni para el investigador ni para los colaboradores del estudio. Si decide participar, tendrá el derecho de solicitar toda la información relacionada con los resultados y los análisis derivados de los datos registrados a lo largo del proceso.

Participantes: La participación en este estudio es completamente voluntaria, y puede decidir mantenerla o abandonarla en cualquier momento. Los datos e información que proporcione, así como los obtenidos a través de las mediciones, seguirán siendo de su propiedad. Al firmar el consentimiento informado, autoriza el uso de estos datos en beneficio de esta investigación.

Duración Estimada: De 8 a 13 sesiones virtuales de intervención psicológica con una duración de 40 a 50 minutos, las cuales recibirá como beneficio del estudio.

Derecho a Retirarse del Estudio de Investigación: Tiene el derecho de retirarse del estudio en cualquier momento. No obstante, es importante mencionar que los datos recopilados hasta ese momento seguirán siendo parte del estudio, a menos que solicite expresamente que se elimine su identificación y la información asociada de la base de datos.

Procedimiento: Se realizarán grabaciones de audio y video de las sesiones virtuales de intervención psicológica para generar la información necesaria para las pruebas del dispositivo de detección y cuantificación de emociones. Las sesiones de intervención se llevarán a cabo por un terapeuta en formación bajo la supervisión de un terapeuta experto con número de cédula 08777824, mientras que únicamente el procesamiento de la información será llevado a cabo por los investigadores principales del estudio. El video y el audio serán utilizados única y exclusivamente para fines de investigación, asegurando la confidencialidad y anonimato de estos.

Las pruebas para la valoración de salud mental y estado emocional que se le aplicarán de obtener su consentimiento se describen a continuación:

- **Escala BAI:** Permite detectar y evaluar el nivel de ansiedad.
- **Escala BDI-II:** Permite detectar nivel de depresión.
- **Escala PANAS:** Permite detectar emociones positivas y negativas a partir del nivel de satisfacción.

Estos documentos serán aplicados y evaluados por un profesional de salud mental, el cual ha firmado una carta de confidencialidad, con el fin de que su información sea resguardada de manera segura y respetando su privacidad.

Este consentimiento informado cumple con los estándares establecidos en el Reglamento de la Ley General de Salud en Materia de Investigación para la Salud, la Ley Federal de Protección de Datos Personales en Posesión de Particulares, la Declaración de Helsinki y las Buenas Prácticas Clínicas emitidas por la Comisión Nacional de Bioética.

Investigadores principales durante la aplicación del protocolo:

- Ing. Francisco Emiliano Sánchez Callejas, Facultad de Ingeniería, UAQ
- Dr. Irving Armando Cruz Albarrán, Facultad de Ingeniería, UAQ
- Lic. Li Erandi Tepepa Flores, CAEPSI "Dr. Benjamín Domínguez"
- Carlos Demian Betanzos de la Vega, CAEPSI "Dr. Benjamín Domínguez"
- Dr. Luis Alberto Morales Hernández, Facultad de Ingeniería, UAQ

Yo, _____ he leído el procedimiento descrito en el presente documento. El equipo de investigadores responsables me ha explicado el estudio y han contestado mis preguntas. Voluntariamente doy mi consentimiento para participar en el proyecto denominado "Detección y cuantificación de emociones en población mexicana mediante el análisis de expresiones faciales y reconocimiento de voz".

_____

Firma del participante

Preguntas o dudas sobre los derechos como participante en este proyecto, pueden ser dirigidas a: Ing. Francisco Emiliano Sánchez Callejas, (emiliano.sanchez@.uaq.mx).

Facultad de Ingeniería, UAQ.

7.2. Carta de Confidencialidad de Datos

Universidad Autónoma de Querétaro
Facultad de Ingeniería

CARTA DE CONFIDENCIALIDAD DE DATOS

San Juan del Río, Querétaro, a _____ de _____ del año 20_____

Yo, _____ me comprometo a aceptar de manera íntegra el presente acuerdo de confidencialidad de datos. La información recopilada se obtendrá en el marco del proyecto titulado "Detección y cuantificación de emociones en población mexicana mediante el análisis de expresiones faciales y reconocimiento de voz". Los compromisos de confidencialidad a los que me obligo son los siguientes:

- Salvaguardar la información obtenida en el marco de este proyecto de investigación.
- No divulgar ninguna información de carácter confidencial.
- Utilizar la información obtenida exclusivamente con fines académicos e investigativos.

Además, acepto los compromisos, requisitos y posibles sanciones que conlleva el presente acuerdo de confidencialidad de datos.

_____
Firma del responsable

Preguntas o dudas sobre los derechos como participante en este proyecto, pueden ser dirigidas a: Ing. Francisco Emiliano Sánchez Callejas, (emiliano.sanchez@.uaq.mx).

Facultad de Ingeniería, UAQ.

7.3. Reglamento de Laboratorio de la Universidad Autónoma de Querétaro

Reglamento General del Laboratorio de Ingeniería Electromecánica y Automotriz

ACCESO

1. Los usuarios podrán usar solamente la maquinaria siempre bajo la vigilancia de los laboratoristas y/o profesores. Otro tipo de uso requerirá de previa autorización del jefe de laboratorios siempre y cuando se haya recibido cursos de capacitación específicos.

2. Como requisitos de acceso en las áreas de Máquinas – Herramienta, CNC y Robótica, Metalografía, Soldadura, Mecánica Automotriz y Carpintería, se deberán observar las siguientes medidas:

   - Sujetar el cabello largo.
   - Vestir camisola, bata u overol, preferentemente en mezclilla, dependiendo la actividad.
   - Vestir pantalón largo, preferentemente mezclilla.
   - Usar exclusivamente zapatos cerrados de piel.
   - Guardar pulseras, anillos, collares, aretes largos, y demás objetos que puedan causar riesgo de lesiones.
   - No jugar ni correr por los pasillos de los laboratorios.

3. El acceso a los laboratorios será exclusivamente para realizar actividades académicas, por lo que está prohibido realizar cualquier otra actividad.

4. Está prohibido introducir y/o consumir alimentos y bebidas en los talleres de Cómputo, Química, Eléctrica, Física y Electrónica, Metalografía, Mecánica Automotriz y Carpintería.

5. Está prohibido fumar dentro de los laboratorios.

6. El acceso a las áreas de Cómputo, Metalografía, Física y Electrónica, Soldadura, Mecánica Automotriz, CNC y Robótica, Soldadura y Máquinas-Herramientas, será exclusivamente para clases y/o actividades previamente autorizadas.

MAQUINARIA Y EQUIPO

1. El uso de la maquinaria y equipo es exclusivo para actividades académicas; es indispensable acatar las indicaciones y restricciones de uso y seguridad correspondientes.

2. Para hacer uso de la maquinaria y equipo, cada usuario deberá usar equipo de seguridad personal obligatoria, el cual se compone de guantes, peto, gafas y/o mascarilla dependiendo la actividad a realizar. Además, en el caso de los alumnos, es obligatorio tener vigente su seguro social facultativo.

3. Cada usuario será responsable del uso de la maquinaria y equipo que está utilizando en su funcionamiento y limpieza. En caso de detectar anomalías en el funcionamiento de estas, se deberá dar aviso inmediatamente al laboratorista en turno.

4. En caso de que la descompostura de la maquinaria sea causada por imprudencia o mal uso, el usuario deberá asumir el costo por el mantenimiento correctivo necesario.

5. La aplicación de pinturas y barnices, en aerosol o espray, será exclusivamente fuera de los laboratorios en el área de maniobras.

6. Todo usuario que sea sorprendido pintando o barnizando en la maquinaria, los muros o el piso, se les negará el acceso a los laboratorios por el resto del semestre.

7. Toda persona que haga uso de los laboratorios deberá depositar la basura y materiales de desperdicio, tanto de las mesas de trabajo como de las máquinas, en los botes de basura colocados dentro de los laboratorios.

8. Se prohíbe el apartado de las mesas, maquinaria y equipo.

9. Las mochilas y materiales que no se estén usando, podrán colocarse en las gavetas, siendo el único responsable el dueño de estos. La UAQ no se hace responsable por el daño o robo de objetos depositados en dichos espacios.

10. Si se requiere hacer uso de los laboratorios fuera del horario establecido, deberá hacerse la solicitud al jefe de laboratorios de los talleres por lo menos con dos días de anticipación.

11. Al término de cada día todo el material y trabajos que se encuentren olvidados dentro de las mesas de trabajo o del área de maquinaria, serán depositado en el estante de Materia Prima, y se podrá disponer de ellos como material de reciclaje. Si se desea guardar algún trabajo durante el semestre, se deberá solicitar la autorización correspondiente con el jefe de laboratorios.

12. Al finalizar el semestre no podrá ser guardado ningún trabajo por ningún motivo. Aquellos trabajos olvidados serán desechados o destruidos para utilizarse como material de reciclaje. La fecha límite para recoger trabajos será el correspondiente al siguiente día de la exposición final del semestre en curso.

13. La UAQ no se hace responsable por los daños ocasionados al material almacenado en el estante de Materia Prima.

14. A todo aquel que sea sorprendido robando material, herramienta u objetos personales, se le aplicará el reglamento vigente de la UAQ.

PRÉSTAMO DE HERRAMIENTA Y EQUIPO

1. El préstamo de herramienta y equipo se hará en el almacén durante los días hábiles marcados en el calendario escolar de la UAQ en horario de 08:00 a 22:00 hrs.

2. El préstamo de herramienta y equipo será personal, mediante el llenado de la Solicitud de Préstamo (vale) y entrega de la credencial vigente de la UAQ, de alumno o profesor, en su defecto, la credencial de elector; ésta última, a reserva de que el solicitante sea conocido. La credencial permanecerá en garantía hasta la devolución de toda la herramienta y equipos amparados.

3. El préstamo es personal y solo es válido para el día solicitado. Toda herramienta y equipo deberán ser devueltos el mismo día, de lo contrario, el usuario se hará acreedor a una multa. Otro tipo de préstamos especiales requerirá de previa autorización del jefe de laboratorios.

4. Toda herramienta y equipo no podrán ser extraídos de los laboratorios por ninguna circunstancia, de lo contrario, se le aplicará una sanción al usuario y podría ser tipificado como robo.

5. Es responsabilidad del usuario revisar que la herramienta y equipo solicitados se encuentren en buen estado, completos y funcionando; no podrá hacer reclamaciones al momento de la devolución de los mismos.

6. Cada usuario será responsable de la herramienta y equipo que aparezcan en la solicitud de préstamo responsabilizándose en caso de descompostura o pérdida de éstas.

7. En caso de pérdida o descompostura de alguna de las herramientas o equipo que se hayan solicitado, el usuario deberá reportarla inmediatamente al laboratorista y/o profesor y contará con cinco días hábiles para la reposición de la misma o el pago correspondiente a la reparación de las mismas, mientras tanto la credencial quedará en garantía.

8. Toda herramienta y equipo deberán ser devueltos al almacén, limpia y en buen estado, de no ser así, se considerará pérdida.

9. La herramienta y equipo que se reponga por pérdida deberá ser nueva, de la misma marca, modelo, medidas y especificaciones que la solicitada en el préstamo originalmente.

10. En caso de que la herramienta y equipo reportados como pérdida aparezcan en poder de otra persona se le considerará como robo.

11. Toda situación no contemplada en este reglamento quedará bajo el criterio del jefe de laboratorio, del coordinador de Ingeniería Electromecánica y de los reglamentos vigentes de la UAQ.

7.4. Inventario de depresión de Beck (BDI-II)

Nombre:_____ Fecha:_____

Sección II

Lea cuidadosamente cada grupo de enunciados, después **escoja el que mejor describe la forma en que se ha estado sintiendo durante las últimas 2 semanas incluyendo el día de hoy.** Subraye el enunciado que escogió. Por favor no deje ningún grupo en blanco. Si varios enunciados dentro de un grupo le parecen adecuados para su situación, simplemente elija el enunciado que tenga el número más alto. ****Asegúrese de solo marcar un enunciado.**

1. **Tristeza**
    0) No me siento triste
    1) Me siento triste la mayoría del tiempo
    2) Me siento triste todo el tiempo
    3) Me siento tan triste o infeliz que no lo puedo soportar
2. **Pesimismo**
    0) No me siento desalentado sobre mi futuro
    1) Me siento más desalentado de mi futuro de lo que solía estar
    2) No espero que las cosas me salgan bien
    3) Siento que mi futuro no tiene esperanza y que únicamente empeorará
3. **Errores Pasados**
    0) No me siento como un fracasado
    1) He fallado más de lo que debería
    2) Cuando reflexiono en mi pasado, veo demasiados fracasos
    3) Siento que soy un completo fracaso como persona
4. **Pérdida de Placer**
    0) Me causan tanto placer las cosas que me gustan como antes lo hacían
    1) No disfruto tanto de las cosas como antes

2) Obtengo muy poco placer de las cosas que antes disfrutaba
3) No puedo obtener placer de las cosas que antes disfrutaba

5. **Sentimientos de Culpa**

0) No me siento particularmente culpable
1) Me siento culpable sobre cosas que hice o debí haber hecho
2) Me siento culpable la mayor parte del tiempo
3) Me siento culpable todo el tiempo

6. **Sentimiento de Castigado**

0) No siento que me estén castigando
1) Siento que puedo estar siendo castigado
2) Siento que me van a castigar
3) Siento que estoy siendo castigado

7. **Desagrado con uno Mismo**

0) Me siento igual conmigo mismo que siempre
1) He perdido confianza en mí mismo
2) Estoy decepcionado de mí mismo
3) Me desagrado

8. **Autocrítica**

0) No me critico ni me culpo más de lo usual
1) Soy más crítico conmigo mismo de lo usual
2) Me critico por todos mis errores o faltas
3) Me culpo de todo lo malo que pasa

9. **Pensamientos Suicidas y Muerte**

0) No tengo ningún pensamiento sobre suicidarme
1) He tenido pensamientos suicidas, pero no los llevaría a cabo
2) Quisiera suicidarme
3) Si tuviese la oportunidad, me suicidaría

10. **Llorar**

0) No lloro más de lo usual

1) Lloro más de lo que solía
2) Lloro por cosas insignificantes
3) Me siento como si quisiera llorar, pero no puedo

11. **Agitación**

0) No me encuentro más agitado de lo usual
1) Me siento más agitado de lo usual
2) Me siento tan agitado que me es difícil mantenerme quieto
3) Me siento tan agitado que tengo que estar moviéndome o haciendo algo

12. **Pérdida de Interés**

0) No he perdido el interés en otras personas o actividades
1) Estoy menos interesado en otras personas o actividades que antes
2) He perdido la mayoría del interés en otras personas o actividades
3) Es difícil el interesarme en algo

13. **Indecisión**

0) Tomo decisiones tan bien como siempre
1) Encuentro más difícil tomar decisiones que antes
2) Tengo mucha mayor dificultad en tomar decisiones que antes
3) Tengo problemas en tomar cualquier decisión

14. **Sin Valía**

0) No me siento como una persona sin valía
1) No me considero tan útil o con tanta valía como solía hacerlo
2) Siento que valgo menos comparado con otras personas
3) Me siento completamente devaluado, sin valía alguna como persona

15. **Pérdida de Energía**

0) Tengo tanta energía como siempre
1) Tengo menos energía de lo usual
2) No tengo suficiente energía para hacer demasiado
3) No tengo suficiente energía para hacer cualquier cosa

16. **Cambios en Patrones de Sueño**

0) No he experimentado ningún cambio en mis patrones de sueño
1) Duermo algo más de lo usual
2) Duermo algo menos de lo usual
3) Duermo mucho más de lo usual
4) Duermo mucho menos de lo usual
5) Duermo la mayor parte del día
6) Me despierto 1-2 horas más temprano de lo usual y ya no puedo volver a dormir

17. **Irritabilidad**

0) No me encuentro más irritable de lo usual
1) Estoy más irritable de lo usual
2) Estoy mucho más irritable de lo usual
3) Estoy irritable todo el tiempo

18. **Cambios en el Apetito**

0) No he tenido cambios en mi apetito
1) Mi apetito es algo menos de lo usual
2) Mi apetito es algo más de lo usual
3) Mi apetito es mucho menos que antes
4) Mi apetito es mucho más que antes
5) No tengo apetito
6) Pienso en comida todo el tiempo

19. **Dificultad para Concentrarse**

0) Me puedo concentrar tan bien como siempre
1) No me puedo concentrar tan bien como antes
2) Me es difícil mantener mi atención en algo por mucho tiempo
3) No me puedo concentrar en nada

20. **Cansancio o Fatiga**

0) No me siento más cansado o fatigado de lo normal
1) Me he cansado o fatigado más fácilmente de lo usual
2) Estoy muy cansado o fatigado como para hacer muchas de las cosas que antes hacía
3) Estoy muy cansado o fatigado como para hacer todo lo que hacía antes

21. **Pérdida de Interés en el Sexo**

0) No he notado cambios recientes en mi interés por el sexo
1) Estoy menos interesado en el sexo de lo que solía estar
2) Estoy mucho menos interesado en el sexo ahora
3) He perdido el interés en el sexo por completo

7.5. Inventario de ansiedad de Beck (BAI)

Nombre:_____ Fecha:_____

Edad:_____ Sexo:_____

A continuación, encontrarás síntomas que usualmente se experimentan con la ansiedad. Indica el grado de intensidad o malestar que los siguientes síntomas te han provocado en las últimas **2 SEMANAS.**

#	**Síntoma**	**Nada**	**Leve**	**Moderado**	**Severo**
1	Hormigueo o entumecimiento				
2	Sensación de calor				
3	Debilidad en las piernas				
4	Incapacidad para relajarme				
5	Miedo a que suceda lo peor				
6	Mareo o vértigo				
7	Palpitaciones o taquicardia				
8	Intranquilo o inestable				
9	Asustado o atemorizado				
10	Nerviosismo				
11	Sensación de ahogarme				
12	Temblor de manos				
13	Temblor generalizado				
14	Miedo a perder el control				
15	Dificultad para respirar				
16	Miedo a morir				
17	Asustado				
18	Indigestión o molestia abdominal				
19	Sensación de desmayarse				
20	Rubor o enrojecimiento facial				
21	Sudoración				

7.6. Escala de afecto positivo/afecto negativo (PANAS)

Nombre:_____ Fecha:_____

Edad:_____ Sexo:_____

	Nunca	**Casi nunca**	**A veces**	**Frecuentemente**	**Siempre**
Felicidad					
Alegría					
Plenitud					
Satisfacción					
Calma					
Bienestar					
Paz					
Dicha					
Tranquilidad					
Placer					
Sufrimiento					
Dolor					
Tristeza					
Desilusión					
Desdicha					
Melancolía					
Soledad					
Miedo					
Incertidumbre					
Irritación					

7.7. Productos obtenidos

2024 IEEE 3rd Conference on Information Technology and Data Science (CITDS) • August 26–27, 2024 • Debrecen, Hungary

Emotion recognition based on facial gestures and Convolutional Neural Networks

Francisco Emiliano Sanchez-Callejas
Autonomous University of Queretaro
Faculty of Engineering
Queretaro, Mexico
emiliano.sanchez@uaq.mx

Irving A. Cruz-Albarran
Autonomous University of Queretaro
Faculty of Engineering
Queretaro, Mexico
irving.cruz@uaq.mx

Luis A. Morales-Hernandez
Autonomous University of Queretaro
Faculty of Engineering
Queretaro, Mexico
luis.morales@uaq.mx

***Abstract*—Humans express emotions verbally and non-verbally through their voice, facial expressions, and body language. Facial expression recognition systems can identify the emotional state of any person by using different intelligent algorithms, such as Support Vector Machines, Hidden Markov Models, and Convolutional Neural Networks, among others. This study focuses on facial expression recognition using eye and mouth regions of images from the FER-2013 dataset by training convolutional neural network (CNN) models. Seven emotional states - happy, sad, fear, anger, disgust, surprise and neutral - were identified. The methodology included segmenting and concatenating the images to form three CNN models. The best-performing model, a four-layer CNN with 8, 16, 32, and 64 filters, achieved remarkable results: 99.05% accuracy, 100.00% precision, 93.75% recall, 96.77% F1-score, 95.95% validation accuracy, and a 0.15 validation loss with a processing time of 3.03 minutes. It was possible to develop a CNN model capable of identifying seven emotional states from only the data of the eye and mouth region using concatenated images.**

***Index Terms*—Facial gestures, emotion recognition, data concatenation, FER, CNN, image segmentation.**

I. Introduction

Humans can communicate verbally and non-verbally using voice, facial expressions, and body language to express emotions [1]. Emotions are usually classified as primary and secondary and play a fundamental role in human interaction. Primary emotions, such as anger, fear, happiness, sadness, disgust, and surprise, are inherent to human beings, so they are present from birth [2].

There are several methods to understand facial expressions, such as Facial Expression Recognition (FER), which uses a variety of techniques, including Multi-Layer Perceptron (MLP), which is a type of neural network that allows solving non-linearly separable problems, Support Vector Machines (SVM), which is a supervised learning algorithm used in classification and regression problems, Convolutional Neural Networks (CNN), Artificial Neural Networks (ANN), which are computational models based on the behavior of biological neurons where the model can learn through the inputs and outputs of the system, and Hidden Markov Models (HMM), which is a statistical model that can describe the evolution of observable events that depend on internal factors that are not directly observable, among others [3][4][5][6]. These systems are able to detect and classify emotions from images, although they often have to deal with data sets that may contain irrelevant information.

Research has shown that critical regions of the face, such as the eyes and mouth, known as fast signals, are particularly influential in emotional expression. In contrast, elements such as skin tone, head shape, eye and mouth position and size, among others, are known as slow signals that do not contribute to the generation of expression [4].

Studies show that algorithms such as SVM obtain results between 85.2% and 90.3% accuracy. In contrast, methods such as the K-Nearest Neighbors (KNN) algorithm have an accuracy of around 84% [7]. In a separate study, various methods were evaluated, including ANN, which achieved an accuracy of 60.09%; HMM, which attained an accuracy of 78.64%; and SVM, which reached an accuracy of 79.88%, among others, when analysing the accuracy for different databases [8].

Similarly, the removal of noise from images enhances the precision of convolutional neural network (CNN) models. However, this approach necessitates a greater computational load, due to the processing required to retain pertinent information while reducing noise. In contrast, object detection algorithms such as Haar Cascade models are frequently employed, wherein the positive images contain the object to be detected, while the negative images display irrelevant content. Furthermore, integrating these algorithms with Deep Neural Networks (DNN), or ANNs, can improve model accuracy [9][10]. Although intelligent algorithms are effective in object detection and pattern recognition, most studies focus on processing whole images without focusing on specific regions of interest. FER systems are employed in a number of fields, including psychology, where they are used as indicators of emotional and mental state. They are also utilized in security, as lie detectors, in the detection of operator fatigue, and in autonomous counseling systems, among other applications [11].

This study proposes a methodology for emotion recognition based on the fusion of facial expression data. The accuracy and processing time of this methodology are compared with the results of studies used to validate the dataset, regardless of factors such as age, gender, and ethnicity. This is to evaluate its impact on the accuracy of facial expression recognition.