A Deep Learning Approach to Eustress and Distress Detection Through Speech

Diego Pacheco, Juan Pedro Vásquez

December 3, 2023

Abstract

Este estudio se centra en desarrollar un modelo de detección de estrés en el habla mediante técnicas de Deep Learning, específicamente utilizando Redes Neuronales Convolucionales (CNN). Motivados por la relevancia del estrés en el ámbito psicológico y sus implicaciones, nuestro objetivo principal es superar el estado del arte en la clasificación de condiciones de estrés, diferenciando entre eustrés y distrés. Utilizamos diversas bases de datos open-source (RAVDESS, Crema-D, SAVEE y TESS), reetiquetando la data para abordar la falta de conjuntos específicos de entrenamiento. Destacamos la elección estratégica de combinar características como los coeficientes MFCC y el Mel Spectrogram. Nuestra CNN, entrenada con técnicas de Data Augmentation y regularización, logra un notable accuracy del 96.01% en el conjunto de validación, superando el rendimiento previo. La combinación de MFCC y Mel Spectrogram se revela crucial, con el Mel Spectrogram destacando como la característica más distintiva en la detección de patrones emocionales. Este estudio no solo avanza en la comprensión de la detección de estrés vocal, sino que también ofrece una herramienta efectiva con aplicaciones significativas en la salud mental y la optimización de entornos laborales.

1 Introduction

La creciente demanda de comunicación entre humanos y sistemas inteligentes ha impulsado la investigación en la detección automática de niveles de estrés (Choi et al., 2017). Una aproximación proactiva a esta detección posibilita intervenciones oportunas y apoyo personalizado, siendo la evaluación precisa a través del análisis vocal una herramienta esencial con implicaciones significativas en la vida diaria. La detección de estrés vocal no solo ofrece una ventana única al estado emocional, sino que también impacta en la mejora de la salud mental, la optimización de entornos laborales y la contribución a la investigación científica (Van Puyvelde et al., 2018).

En comparación con métodos tradicionales que incorporan sensores en el cuerpo, el enfoque vocal no intrusivo es preferido por su comodidad y la eliminación de la necesidad de dispositivos adicionales que puedan causar molestias. Esto facilita la implementación a gran escala, mejorando la eficiencia en la recopilación de datos sin generar incomodidades. Además, el análisis vocal captura respuestas emocionales de manera natural y en tiempo real, siendo una alternativa efectiva y prometedora para evaluar el estrés en diversas situaciones (Han et al., 2018; Gedam & Paul, 2021). Aunque se han destacado ventajas de arquitecturas como CNNs, RNNs y LSTMs para el reconocimiento de características vocales, hay espacio para mejorar la representación de estas junto con la selección de la arquitectura del modelo (Banerjee et al., 2021).

El objetivo principal de esta investigación es desarrollar y evaluar un modelo de detección de estrés en el habla utilizando técnicas de Deep Learning, específicamente mediante el empleo de Redes Neuronales Convolucionales (CNN). La motivación surge de la importancia del estrés en el ámbito psicológico y sus consecuencias en el rendimiento cognitivo y la toma de decisiones. El propósito fundamental es superar al state of art, que tiene un validation accuracy de 94.33% (Vaikole et al.,

2020) en la clasificación de condiciones de estrés, diferenciando entre eustrés (asociado a estados emocionales positivos) y distrés (vinculado a emociones negativas), mediante la extracción eficiente de características de señales de habla. La investigación se centra en la combinación estratégica de características, en particular los coeficientes MFCC y el Mel Spectrogram, para maximizar la precisión del modelo. Además, se aborda la falta de bases de datos específicas para entrenar modelos de clasificación de estrés, proponiendo una reetiquetación basada en la relación entre las emociones y el eustrés y distrés.

2 Related Work

La definición general de estrés destaca un cambio en las respuestas psicológicas de la calma a lo emocional. La evidencia científica sugiere que el estrés puede ser categorizado como eustrés, vinculado a estados emocionales positivos como la alegría y la emoción, y distrés, asociado a emociones negativas como la ira, ansiedad, tristeza, miedo, dolor y nerviosismo (Choi, Jeon, Wang, & Kim, 2017). Diversas investigaciones señalan una estrecha relación entre situaciones estresantes y la disminución en la eficacia de las habilidades de toma de decisiones (Wemm & Wulfert, 2017), así como una reducción en el rendimiento cognitivo y la motivación, junto con una disminución de la conciencia del entorno circundante (Morgado & Cerqueira, 2018; Kim, Měsíček, & Kim, 2021).

La extracción de características de una señal de habla desempeña un papel crucial al transformar la señal de audio en un formato comprensible para el modelo. En su estudio, Banerjee et al.(2021) llevaron a cabo una búsqueda minuciosa de distintas combinaciones de características de entrada no estructuradas y arquitecturas de modelos con el objetivo de identificar aquellas combinaciones de características y arquitecturas existentes que alcanzan las precisiones de validación más altas en el campo de la clasificación de emociones por medio del habla. Como resultado de su experimento, identificaron tres características de entrada no estructuradas que de manera constante superaron a todas las demás combinaciones en diversas network architectures: el Mel-spectrogram, los coeficientes MFCC y la Tonnetz Representation.

Un Mel-spectogram transforma de manera no lineal un espectrograma para reflejar cómo un ser humano percibe las frecuencias relativas, lo que es importante ya que refleja cómo se percibe una emoción, lo que lo convierte en el punto de partida adecuado para que un modelo aprenda representaciones emocionales latentes. El MFCC, por su parte, es esencialmente una versión comprimida y decorrelacionada del Mel-spectograma, utilizado principalmente para Modelos de Mezcla Gaussiana. A menudo, es suficiente entrenar clasificadores más fuertes como las CNN solo con el Mel-spectograma, ya que la CNN debería ser capaz de aprender las representaciones complejas dentro del propio espectrograma. Sin embargo, encontramos que con datos limitados, incluir MFCCs proporciona un impulso inicial para que el modelo aprenda representaciones latentes. (Jason, & Kumar, 2020; Madhavi, Chamishka, Nawaratne, Nanayakkara, Alahakoon, & De Silva, 2020).

Palanisamy et al. (2020) presenta un trabajo preliminar sobre el uso de modelos basados en CNN para la clasificación de audio. Encontraron que MECC, el centroide espectral, el contraste espectral y el croma STFT produjeron la mayor precisión del modelo, lo que nos ayudó a entender que la diversidad de estas características proporciona al modelo una información más completa y detallada sobre las propiedades del sonido, lo que puede ser especialmente útil en la identificación de patrones complejos y en la mejora del rendimiento general del modelo. Finalmente, los autores muestran que su modelo multimodal tuvo un rendimiento superior al modelo de vanguardia en ese momento (Poria et al., 2017) que solo utilizaba los coeficientes cepstrales en frecuencia mel (MFCC) como entrada. Esto respalda la idea de que un conjunto bien seleccionado y diverso de características, combinado con suficientes datos de entrenamiento, puede ser fundamental para lograr un mayor nivel de precisión en la clasificación de audio.

Finalmente, dos investigaciones respaldan que las Redes Neuronales Convolucionales (CNN) destacan por su eficacia identificando patrones en información no estructurada, como la proporcionada por las representaciones de Mel Frequency Cepstral Coefficients (MFCC), el Mel Spectrogram. Estos datos,

que incluyen información temporal y frecuencial crucial para las emociones expresadas en el habla, se benefician de la habilidad de las CNNs para aprender características jerárquicas y secuenciales. Las CNNs tienen la capacidad de reconocer patrones acústicos relevantes sin depender de la ubicación temporal precisa de esos patrones en los datos de entrada, lo que es particularmente útil en aplicaciones de procesamiento de señales, como el reconocimiento de voz, donde los eventos relevantes pueden ocurrir en diferentes momentos dentro de la secuencia temporal, y la red necesita ser capaz de identificarlos sin depender de la precisión temporal exacta (Lee, & Tashev, 2015; Abdel-Hamid, Mohamed, Jiang, Deng, Penn, & Yu, 2014).

3 Experimentation

3.1 Datasets

Por motivos de esta investigación hemos usado de 4 bases de datos open-source: Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), Crowd-Sourced Emotional Model Actor Dataset (Crema-D), Surrey Audio-Visual Expressed Emotion (SAVEE) y Toronto Emotional Speech Set (TESS), donde cada una contiene 1440, 7442, 480 y 2800 archivos de audio, respectivamente. Dichos audios fueron grabados por actores de doblaje, y están etiquetados cómo una de las siguientes emociones: angry, disgust, fearful, happy, neutral, sad, surprise y calm.

3.2 Data Preparation and Augmentation

Por desgracia, hasta el momento no existe una base de datos open-source de audios etiquetados cómo eustrés o distrés, que nos ayudarían a entrenar modelos para la clasificación de estrés en un hablante. A pesar de esto, los audios presentes en las bases de datos open-source: RAVDESS, TESS, Crema-D y SAVEE están etiquetados según la emoción presente en estos.

Anteriormente, fue mencionado que la evidencia científica sugiere que el estrés puede ser categorizado como eustrés, vinculado a estados emocionales positivos, y distrés, asociado a emociones negativas (Choi, Jeon, Wang, & Kim, 2017). Por ello, reetiquetaremos nuestra data en base a esta información, es decir, los audios que están etiquetados con emociones positivas cómo: happy, neutral, surprise (happy surprise), calm, serán reetiquetados cómo eustrés, y los audios etiquetados con emociones negativas cómo: angry, disgust, fearful, sad, serán reetiquetados cómo distrés. Luego de reetiquetar nuestra data, 4470 audios etiquetados como eustrés y 7692 audios etiquetados cómo distrés fueron obtenidos, lo que presenta un desbalanceo de clases "importante".

Se llevó a cabo un proceso de Data Augmentation con la intención de crear nuevas muestras de datos sintéticos y enriquecer el conjunto de datos mediante la adición de pequeñas perturbaciones al conjunto de entrenamiento inicial. Dicho proceso fue realizado solo a la clase minoritaria distrés, con la intención de equilibrar el conjunto de datos y mejorar el rendimiento del modelo al proporcionar más variedad y cantidad de ejemplos para dicha clase. Como resultado de dicho proceso tenemos un total de 15,384 archivos de audio.

3.3 Feature extraction

Existen muchas bibliotecas open-source para la extracción de características de audio, como Librosa, Essentia, aubio, Madmom y Marsyas. Diferentes bibliotecas pueden generar representaciones numéricas diferentes para la misma característica (por ejemplo, el espectro Mel). Para este estudio, se seleccionó Librosa como la biblioteca utilizada para la extracción de características de audio.

A través de las características extraídas del audio, se pueden obtener diversas descripciones de

propiedades, las cuales luego se introducen en el modelo. La característica útil dentro de la categoría del dominio de la señal se empleará en la detección de estrés debido a que engloba varias características esenciales vinculadas al audio en general. Las características en el dominio de la señal comprenden representaciones en el dominio temporal, en el dominio de frecuencia y representación cepstral. El dominio temporal implica una característica extraída directamente de la forma de onda del archivo de audio. El dominio de la frecuencia se centra en el componente de frecuencia de la señal de audio. Por lo general, la señal se convierte del dominio temporal al dominio de la frecuencia mediante una transformada de Fourier, y la representación cepstral es una característica que fusiona los componentes temporal y de frecuencia de la señal de audio, obtenida al aplicar una transformada de Fourier de corto tiempo a la forma de onda en el dominio temporal (Mcfee, Raffael, Liang, Ellis, Mcvicar, Battenberg, & Nieto, 2015).

En consonancia con la investigación de Banerjee, Lettiere y Huang (2021), en la que se realizó una cuidadosa exploración de distintas combinaciones de características de entrada no estructuradas y arquitecturas de modelos. Identificaron tres características de entrada no estructuradas que de manera constante superaron a otras combinaciones en diversas arquitecturas de redes. Estas características son el Mel-spectrogram, los coeficientes MFCC y la Tonnetz Representation.

En consecuencia, se ha tomado la decisión de utilizar tanto los coeficientes MFCC como el Mel Spectogram para la extracción de características en el modelo de detección de estrés. La elección de utilizar exclusivamente los coeficientes MFCC y el Mel-spectrogram para el entrenamiento del modelo CNN se fundamenta en su reconocida eficacia en tareas de procesamiento de señales de audio. Estas características ofrecen una representación rica y complementaria de la energía y distribución de frecuencias en el espectro de audio, siendo especialmente pertinentes para la detección de patrones relacionados con el estrés en el habla. La simplificación del modelo, al limitarse a dos conjuntos de características, contribuye a la eficiencia del entrenamiento y mejora la interpretación del modelo. La prescindencia de la representación de Tonnetz se justifica por su aplicabilidad principal en contextos musicales, mientras que la observación empírica respalda la suficiencia de los MFCC y el Mel-spectrogram para lograr altos niveles de precisión en la tarea de detección de estrés. Este enfoque también considera aspectos computacionales, favoreciendo un entrenamiento eficiente, especialmente en escenarios con recursos limitados.

Para formar cada una de estas features, utilizamos 512 FFTs, un hop length de 512, y nuestra máxima frecuencia fue de 22050 Hz. Y finalmente, repartimos nuestra data ya procesada en una partición 70-15-15.

3.4 CNN architecture

La arquitectura de nuestra CNN, mostrada en la imagen 1, se presenta en un modelo secuencial. Inicia con una capa convolucional (Conv2D) que tiene 32 filtros de tamaño 4 x 4, aplicados sin relleno, seguida de una capa de MaxPooling2D que reduce las dimensiones espaciales a la mitad mediante un paso de 2 x 2. Posteriormente, se incorpora normalización por lotes (BatchNormalization).

Este proceso se repite tres veces más, con cada iteración reduciendo aún más las dimensiones espaciales a través de convoluciones y pooling. Cada capa convolucional es seguida por una activación leaky ReLU y normalización por lotes.

Finalmente, la salida se aplana y se conecta a una capa densa (Dense) con 64 neuronas y activación ReLU. Se agrega una capa de Dropout para la regularización. La última capa densa tiene una sola neurona con una función de activación sigmoide, lo que hace que el modelo sea adecuado para problemas de clasificación binaria. La CNN ha sido entrenada durante 50 epochs con un batch size de 128 utilizando el optimizador Adam que ajustará dinámicamente la tasa de aprendizaje (que por defecto es 0.0001) si el proceso de aprendizaje se estanca.

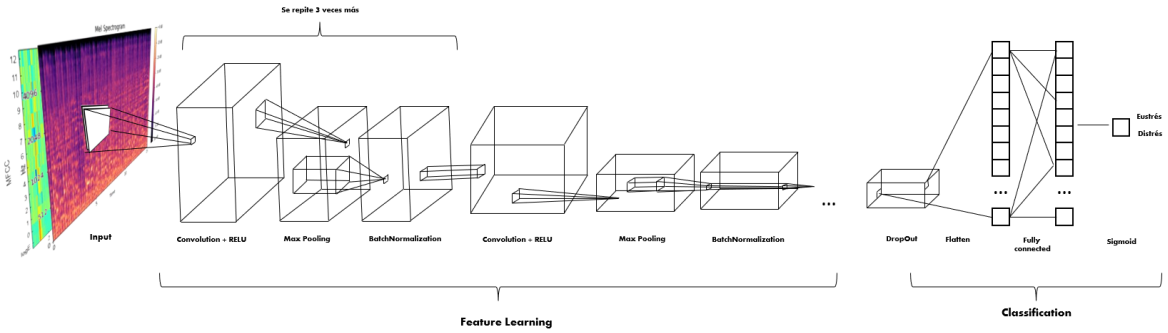

Figure 1: Convolutional Neural Network Model Architecture

4 Results and Analysis

4.1 Desempeño del Modelo y Análisis de Características

En el proceso de evaluación de nuestro modelo de detección de estrés basado en CNN, se llevaron a cabo pruebas exhaustivas utilizando diversas características del dominio cepstral, como el Mel Spectrogram y los coeficientes MFCC. El objetivo principal fue alcanzar una precisión superior al estado del arte (Vaikole et al., 2020) en la clasificación de estrés en el habla.

La elección estratégica de combinar características como los coeficientes MFCC y el Mel Spectrogram demostró ser crucial para el éxito del modelo. La complementariedad de la información entre estas dos representaciones permitió al modelo capturar patrones más complejos y mejorar su capacidad de generalización. Este enfoque, respaldado por la investigación de Banerjee et al. (2021), resultó en un rendimiento superior en comparación con el uso individual de estas características.

4.2 Precisión del Modelo y Superación del Estado del Arte

El éxito de nuestro modelo se evaluó principalmente a través del validation accuracy, una métrica comúnmente utilizada en la literatura para comparar el rendimiento de modelos y características. Como se muestra en las Figuras del Anexo, las validation y test curves, y las matrices de confusión detalladas respaldan la eficacia de nuestro enfoque.

Notablemente, logramos superar el estado del arte con un validation accuracy impresionante del 96.01%, cómo puede ser observado en la Figura 2. Este resultado excepcional destaca la robustez y la capacidad predictiva de nuestro modelo en la tarea de detección de estrés.

De hecho el uso individual del MFCC como característica de entrada fue el que menos rindió en comparación al Mel Spectogram y al uso de ambas características, lo que sorprende ya que el state of art utilizó solo esta característica para obtener sus resultados.

El análisis detallado revela que el Mel Spectrogram emergió como la característica más distintiva para identificar condiciones de estrés en el habla. Su representación rica y directa de la distribución de energía en frecuencias específicas resultó ser crucial para la detección de patrones emocionales. La capacidad inherente de las CNN para procesar patrones complejos en información no estructurada, como el Mel Spectrogram, contribuyó significativamente al rendimiento superior del modelo.

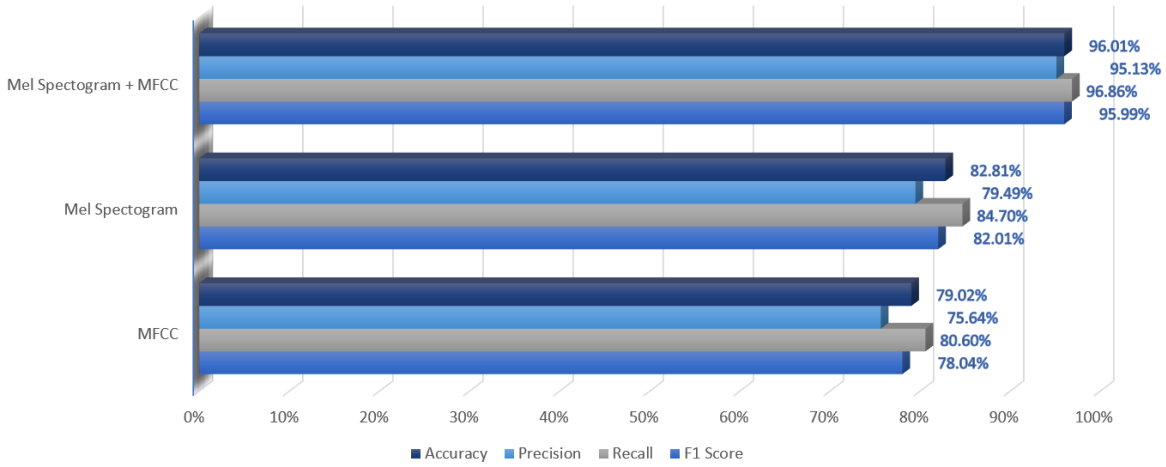

Figure 2: Comparison of model performance using MFCC and Mel Spectrogram features

4.3 Razones del Alto Validation Accuracy

El rendimiento destacado del modelo en el conjunto de validación se debe a una cuidadosa integración de diversos elementos en su diseño y entrenamiento. La técnica de Data Augmentation desempeñó un papel crucial al enriquecer el conjunto de datos, proporcionando variabilidad y robustez, y mitigando el riesgo de sobreajuste, común en clasificación con conjuntos limitados.

La arquitectura de la red neuronal convolucional (CNN) se destacó por la inclusión de técnicas avanzadas, como la normalización por lotes y la regularización, probadas en la extracción efectiva de características durante el entrenamiento. En comparación con estudios previos sin estas mejoras, nuestra elección permitió una mejor adaptación del modelo a datos no vistos.La normalización por lotes garantizó una distribución estable de activaciones en cada capa, agilizando el entrenamiento de la red. La regularización, al introducir penalizaciones para evitar complejidad excesiva, previno el sobreajuste y mejoró la capacidad de generalización. La combinación de Data Augmentation y una arquitectura de CNN mejorada con normalización por lotes y regularización demostró ser sinérgica, mejorando el rendimiento del modelo.

5 References

- Dillon, R., & Ni Teoh, A. (2021). Real-time Stress Detection Model and Voice Analysis: An Integrated VR-based Game for Training Public Speaking Skills. In IEEE Conf. Games (pp. 1–4).
- Morgado, P., & Cerqueira, J. (2018). The Impact of Stress on Cognition and Motivation. Frontiers in Behavioral Neuroscience.
- Gedam, S., & Paul, S. (2021). A Review on Mental Stress Detection Using Wearable Sensors and Machine Learning Techniques. IEEE Access, 9, 84045–84066.
- König, A., et al. (2021). Measuring stress in health professionals over the phone using automatic speech analysis during the COVID-19 pandemic: Observational Pilot study. Journal of Medical Internet Research, 23(4), 1–14.
- Mcfee, B., Raffel, C., Liang, D., Ellis, D., Mcvicar, M., Battenberg, E., & Nieto, O. (2015). Librosa: Audio and music signal analysis in Python. Proceedings of the 14th Python in Science Conference.

- Kim, T. Y., Měsíček, L., & Kim, S. H. (2021). Modeling of Child Stress-State Identification Based on Biometric Information in Mobile Environment. Mobile Information Systems, 2021.
- Banerjee, G., Lettiere, A., & Huang, E. (2021). Understanding Emotion Classification In Audio Data.
- Poria, S., Cambria, E., Hazarika, D., Majumder, N., Zadeh, A., & Morency, L.-P. (2017). Context-dependent sentiment analysis in user-generated videos. In Proceedings of the 55th annual meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 873-883).
- Lee, J., & Tashev, I. (2015). High-level Feature Representation using Recurrent Neural Network for Speech Emotion Recognition. DOI: 10.21437/Interspeech.2015-336.
- Jason, C. A., & Kumar, S. (2020). An Appraisal on Speech and Emotion Recognition Technologies based on Machine Learning. International Journal of Recent Technology and Engineering, 8(5), 2266–2276.
- Palanisamy, K., Singhania, D., & Yao, A. (2020). Rethinking CNN Models for Audio Classification. arXiv preprint arXiv:2007.11154.
- Van Puyvelde, M., Neyt, X., McGlone, F., & Pattyn, N. (2018). Voice stress analysis: A new framework for voice and effort in human performance. Frontiers in Psychology, 9(NOV), 1–25.
- Vaikole, S., Mulajkar, S., More, A., Jayaswal, P., & Dhas, S. (2020). Stress Detection through Speech Analysis using Machine Learning. International Journal of Scientific Research in Science and Technology (IJSRST), 8(5), May 2020, Online ISSN: 2320-28820.
- Madhavi, I., Chamishka, S., Nawaratne, R., Nanayakkara, V., Alahakoon, D., & De Silva, D. (2020). A Deep Learning Approach for Work Related Stress Detection from Audio Streams in Cyber Physical Environments. In IEEE Symposium on Emerging Technology and Factory Automation (ETFA) (pp. 929–936). (2020-Septe).
- Han, H., Byun, K., & Kang, H.-G. (2018). A Deep Learning-based Stress Detection Algorithm with Speech Signal. In Proceedings of the 2018 Workshop on Audio-Visual Scene Understanding for Immersive Multimedia (AVSU'18). Association for Computing Machinery.
- Vamsinath, J., Varshini, B., Sandeep, T., Meghana, V., & Latha, B. (2022). A Survey on Stress Detection Through Speech Analysis Using Machine Learning. International Journal of Scientific Research in Science and Technology (IJSRST), 9(4), 326-333.
- Abdel-Hamid, O., Mohamed, A., Jiang, H., Deng, L., Penn, G., & Yu, D. (2014). Convolutional Neural Networks for Speech Recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 22(10), 1533–1545.

6 Anexos

6.1 Test and Validation curves

Para evaluar el desempeño del modelo, se emplea una métrica de validation y test accuracy, común en la clasificación, que determina la proporción de salidas correctamente clasificadas con respecto al resultado general de la clasificación obtenido.

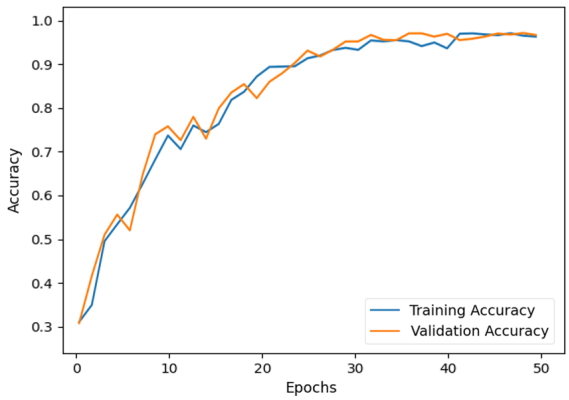

Figure 3: Mel Spectogram + MFCC - Test and Validation Curves

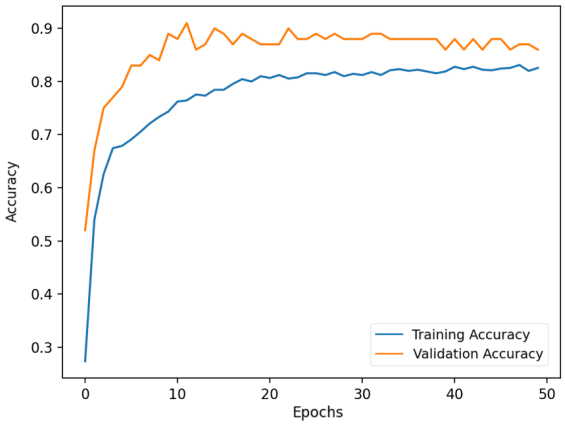

Figure 4: MFCC - Test and Validation Curves

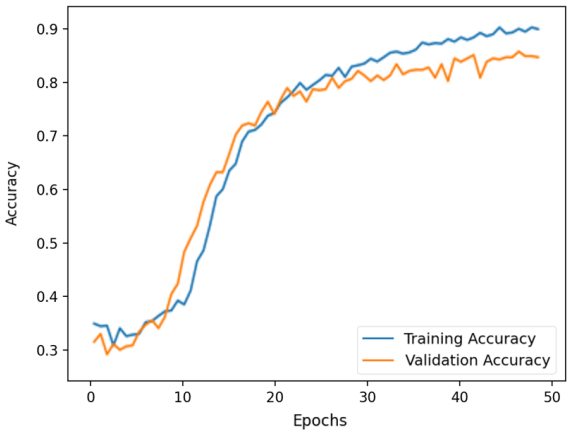

Figure 5: Mel Spectogram - Test and Validation Curves

6.2 Confusion Matrix

Los resultados de la predicción pueden presentar diferentes situaciones, siendo Verdadero Positivo (TP), Falso Positivo (FP), Falso Negativo (FN) y Verdadero Negativo (TN) los posibles estados. TP señala que tanto los resultados de la predicción como la etiqueta real indican que el habla está experimentando dístrés. FP se refiere a una situación en la cual los resultados de la predicción indican que el habla está en un estado de distrés, pero la etiqueta real indica una condición eustrés. FN indica una situación en la cual los resultados de la predicción indican que el habla está en un estado eustrés, pero en realidad, se encuentra en un estado de distrés. Finalmente, TN indica una situación en la cual tanto los resultados predichos como la etiqueta real indican que el habla está experimentando eustrés.

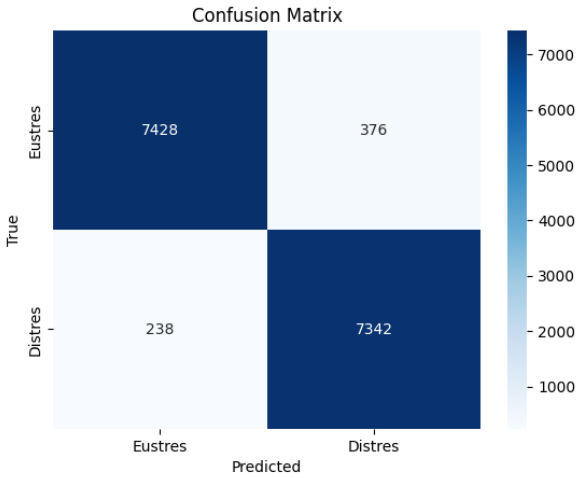

Figure 6: Mel Spectogram + MFCC - Confusion Matrix

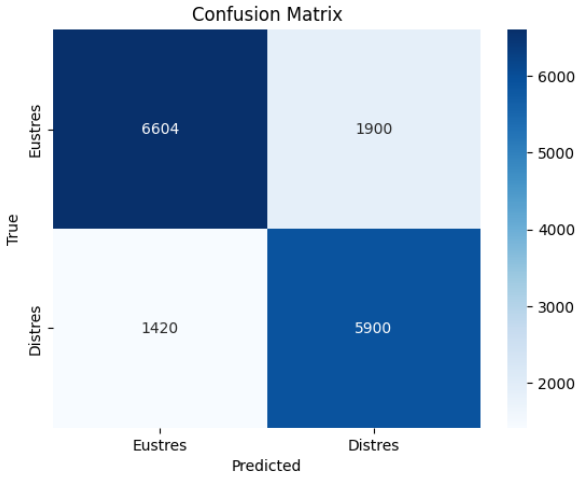

Figure 7: MFCC - Confusion Matrix

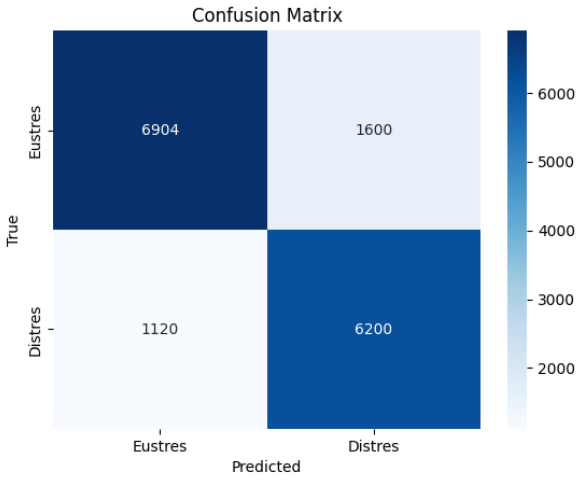

Figure 8: Mel Spectogram - Confusion Matrix