Universidad Politécnica
de Madrid

Escuela Técnica Superior de Ingenieros Informáticos

Máster Universitario en Inteligencia Artificial

Trabajo Fin de Máster

Reconocimiento de Emociones a Partir de Voz mediante Shallow ANN

Autor(a): Javier Meitín Moreno
Tutor(a): Javier De Lope Asiaín

Madrid, Julio - 2024

Este Trabajo Fin de Máster se ha depositado en la ETSI Informáticos de la Universidad Politécnica de Madrid para su defensa.

Trabajo Fin de Máster
Máster Universitario en Inteligencia Artificial

Título: Reconocimiento de Emociones a Partir de Voz mediante Shallow ANN

Julio - 2024

Autor(a): Javier Meitín Moreno
Tutor(a): Javier De Lope Asiaín
Inteligencia Artificial
ETSI Informáticos
Universidad Politécnica de Madrid

Resumen

En este trabajo se propone el diseño y desarrollo de una **red neuronal de una sola capa** oculta capaz de reconocer emociones a partir de la voz. El clasificador es entrenado y evaluado con espectogramas mel de dimensiones 90x98 generados a partir de los archivos de audio de la base de datos pública RAVDESS. Dicho conjunto esta formado por 8 emociones (neutralidad, calma, felicidad, tristeza, enojo, miedo, disgusto y sorpresa). La red neuronal es capaz de clasificarlos con un **68,81 %** de precisión sin la necesidad utilizar otros algoritmos complementarios como aumento de datos. Además, se ha comprobado que seleccionando los 4000 pixels más importantes con **Gradient Boosting**, la precisión del modelo puede mejorar al **71,09 %**. Estos resultados, aunque no superiores a los de otros trabajos anteriores similares, permiten medir y ratificar el alcance y potencial de este tipo de redes neuronales en el ámbito del aprendizaje automático.

Abstract

In this work, the design and development of a **single hidden layer neural network** capable of recognizing emotions from speech is proposed. The classifier is trained and evaluated using Mel spectrograms of dimensions 90x98 generated from audio files from the public RAVDESS database. This dataset consists of 8 emotions (neutral, calm, happy, sad, angry, fearful, disgusted, and surprised). The neural network is able to classify them with a precision of **68,81 %** without the need to use other complementary algorithms such as data augmentation. Additionally, it has been found that by selecting the 4000 most important pixels with **Gradient Boosting**, the model's precision can improve up to **71,09 %**. These results, although not higher than those of some similar previous works, allow for measure and confirmation of the scope and potential of this type of neural network in the field of machine learning.

Tabla de contenidos

Índice de figuras

Índice de cuadros

Capítulo 1

Introducción

1.1. Redes neuronales poco profundas

Las *redes neuronales* (*Neural Networks*, NN) son modelos matemáticos que utilizan algoritmos de aprendizaje inspirados en el cerebro humano para almacenar información. Dado que las redes neuronales se forman y utilizan en ordenadores, también reciben la denominación *redes neuronales artificiales* (*Artificial Neural Network*, ANN). Estos modelos son frecuentemente utilizados en el campo de *aprendizaje automático* (*Machine Learning*, ML), una disciplina que, como su nombre indica, tiene como objetivo entrenar ordenadores para que aprendan a reconocer patrones complejos y tomar decisiones inteligentes basadas en datos. Debido a esto, el aprendizaje automático está estrechamente relacionado con campos como el reconocimiento de patrones y la inteligencia artificial.[1]

En 1960, el estándar de las redes neuronales era que tuvieran una sola capa oculta, actualmente conocidas como *redes neuronales superficiales o poco profundas* (*Shallow ANN*). No fue hasta 1980, con la evolución de las computadoras, que empezaron a tener entre dos o tres capas. Sin embargo, la auténtica revolución no ha surgido hasta la última década. El auge de los videojuegos ha supuesto un aumento en la complejidad de las imágenes y la velocidad de los procesadores, lo cual ha requerido un aumento de hardware. Esto se puede medir especialmente en la *unidad de procesamiento gráfico* (*Graphics Processing Unit*, GPU), la cual cuenta con miles de núcleos de procesamiento relativamente simples en un solo chip. El incremento de potencia de las GPU modernas promovió que las redes neuronales de más de una capa que se habían propuesto en 1980 se convirtieran en las redes de 10, 15, o 50 capas de hoy día. Las *redes neuronales profundas* (*Deep Neural Network*, DNN) reciben dicho adjetivo de la profundidad de sus capas ocultas.[2]

Las DNN presentan varias ventajas respecto a las de una sola capa. En primer lugar, son exponencialmente mejores en ciertas condiciones teóricas, como por ejemplo, al evitar la *maldición de la dimensionalidad* (*Curse of Dimensionality*). Dicho término hace referencia a un fenómeno en ML donde el rendimiento de los algoritmos empeora al aumentar la complejidad de los datos. Normalmente, a medida que la dimensionalidad de los datos aumenta, la cantidad de datos necesarios para describir de manera efectiva el espacio de características también aumenta exponencialmente. Sin embargo, al aumentar el número de capas de la red, esta se vuelve más maleable y

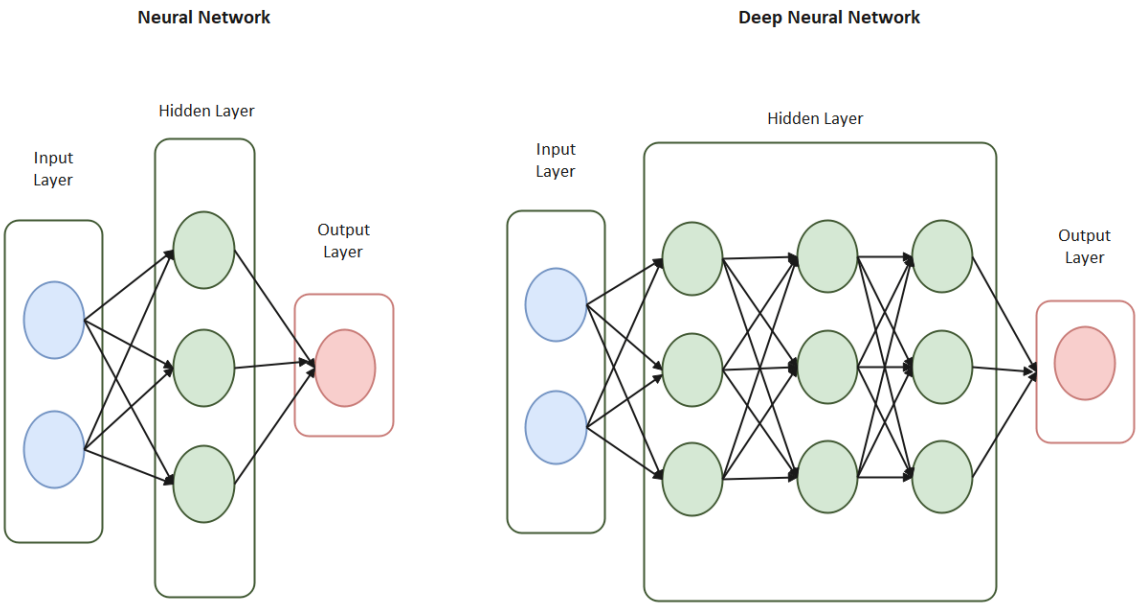

Figura 1.1: Diagrama comparativo de Deep y Shallow ANN, Neural *Network Diagram Complete Guide*, Edraw Content Team

adaptable.[3]

En segundo lugar, las redes profundas tienen una mayor capacidad para aproximar funciones compuestas y funciones de funciones sin sufrir problemas de sobreajuste, a diferencia de las redes superficiales.[3]

En tercer lugar, las redes neuronales profundas pueden ser eficientes en la cuantización vectorial jerárquica, lo que las convierte en potentes herramientas de memoria asociativa.[3]

Por todos estos motivos, las Shallow ANN actualmente se consideran una herramienta obsoleta.

1.2. Reconocimiento de emociones a partir de voz

La capacidad de hablar es una herramienta fundamental de la comunicación humana en la vida cotidiana. No solo permite transmitir información concreta como instrucciones u opiniones, sino también las emociones del emisor. Esto último es gracias a la variabilidad en la intensidad y entonación de la voz junto a la longitud de las pausas entre las palabras al pronunciar estas mismas. Estos atributos, unidos al mensaje y contexto de la conversación, permiten al receptor intuir cuáles son las emociones del hablante durante el momento de la interacción.

Debido a la gran trascendencia que desempeñan los sentimientos en las interacciones sociales, muchos han expresado la importancia de estudiarlos detenidamente.[4] En la última década, los sistemas de *reconocimiento de emociones en el habla* (*Speech Emotion Recognition*, SER) han tenido un auge impulsado por la nueva demanda

de aplicaciones digitales que faciliten las *interacciones humano-máquina* (*Human-Machine Interaction*, HMI) e investigación cerebral, entre otras.[5][6]

Debido a la gran complejidad intrínseca a una grabación de voz, estos modelos suelen preprocesar los datos a un formato más interpretable antes de clasificarlos mediante diversas herramientas como la *extracción de características* (*Feature Extraction*, FE) o los modelos de *aprendizaje profundo* (*Deep Learning*, DL).[4] Sin embargo, aun con un buen preprocesado de datos, la precisión del modelo puede verse afectada por otros factores, como el tamaño de la base de datos. Cuanto mayor es la complejidad de los datos, mayor dificultad tiene el modelo para extraer la importancia de las características y clasificarlos correctamente. En estos casos, ampliar el número de ejemplos en el entrenamiento puede resultar extremadamente beneficioso. No obstante, por este mismo motivo, utilizar un modelo intrincado como DL cuando el número de ejemplos es insuficiente puede desencadenar un rendimiento peor al que podría haber obtenido un algoritmo tradicional de ML.[7]

En el caso de los sistemas SER, la gran complejidad de los datos suele requerir un modelo de clasificación igualmente elaborado. Por este motivo, una de las mayores dificultades a las que se enfrentan los sistemas SER es la falta de datos. Aunque hay varios conjuntos públicos y gratuitos que son ampliamente utilizados, estos son relativamente pequeños.[7] Con el fin de poder valerse del potencial del DL, la mayoría de los investigadores han optado por utilizar técnicas de *aumento de datos* (*Data Augmentation*) para expandir el tamaño del conjunto de entrenamiento.[8]

1.3. Objetivos

El objetivo principal de este trabajo es el diseño y desarrollo de una Shallow ANN capaz de clasificar emociones a partir de la voz, obteniendo un rendimiento y precisión comparables, o superiores en el mejor de los casos, a otros sistemas SER anteriores. Esto, a su vez, tiene como objetivo reevaluar el potencial de las Shallow ANN, herramienta que actualmente se considera obsoleta en comparación con las DNN.

Además de los objetivos globales descritos anteriormente, este trabajo ha supuesto el cumplimiento de unos objetivos específicos complementarios, los cuales se listan a continuación:

- Revisión y estudio del estado del arte del reconocimiento de emociones a partir de la voz, en específico aquellos en los que se han empleado redes neuronales poco profundas.
- Estudio del potencial de modelos de selección de características según su importancia como herramienta complementaria a la red neuronal.

1.4. Metodología y plan de trabajo

En primer lugar, se investiga por qué las Shallow ANN han caído en desuso, al igual que el estado del arte en el reconocimiento de emociones a partir de voz. Tras haber determinado los enfoques principales que diversos autores han empleado en este problema, se decide ahondar en un artículo de 2020 de Anwer Slimi *et al.*[9], el único que ha empleado una red neuronal poco profunda en un sistema SER con anterioridad. Tras haber estudiado diversos artículos, se pasa a seleccionar la base de datos

sobre la que se desarrolla este trabajo. Una vez seleccionada, se realiza un estudio más profundo sobre la misma con el fin de entender cómo está estructurada, qué peculiaridades tiene, etc.

Una vez terminada la fase previa de investigación, se inicia la etapa de Desarrollo. Esta fase se centra en la implementación de un programa donde se transforman los archivos de audio a espectogramas y se alimenten a una red neuronal de una sola capa. Posteriormente, el sistema de software entrena la red neuronal y evalua su rendimiento. Dicho programa es escrito en el lenguaje *Python* en el entorno de *Google Colaboratory* con las siguientes especificaciones de hardware: Intel Xeon a 2.2GHz, 12GB de RAM y GPU Tesla K80 de 12GB. El software cuenta con el soporte de varias librerías, siendo las más destacadas TensorFlow, Keras, librosa y scikit-sklearn. Dicha implementación de código ha sido publicada en un el siguiente repositorio de Github: TFM-MUIA-2024 (https://github.com/javiermeitin/TFM-MUIA-2024).

Al finalizar la primera fase de desarrollo de una Shallow ANN, se decide realizar una optimización de hiperparámetros y un estudio de selección de las características más importantes. La etapa de optimización supone la más duradera de todo el desarrollo debido al gran número de pruebas que se realizan para garantizar resultados consistentes.

1.5. Estructura del documento

Este documento ha sido dividido en diversos capítulos a fin de transmitir una visión estructurada y ordenada de las diferentes etapas en las que se ha dividido el trabajo. En la primera, se realiza una revisión del estado del arte del reconocimiento de emociones a partir de la voz, al igual que los conjuntos de datos más utilizados en dicha área de investigación. En el siguiente capítulo, se explica el Desarrollo, dividido en dos partes: Implementación de técnicas anteriores y ajuste de parámetros. Este capítulo explica las características del modelo empleado y como se han procesado los datos que ha utilizado. En tercer lugar, se presentan los resultados experimentales de forma detallada. Finalmente, en el último capítulo, se redactan las conclusiones del proyecto y vías de trabajo a futuro.

Capítulo 2

Estado del arte

En este capítulo se realiza una revisión del estado del arte en lo relacionado a los sistemas SER, a fin de transmitir una visión general de las distintas bases de datos que se han desarrollado y los modolos que se han propuesto para procesarlas y clasificarlas.

2.1. Bases de datos utilizadas en sistemas SER

En la primera sección, se describe el estado del arte de las bases de datos más empleadas en sistemas de reconocimiento de emociones a partir de la voz. De esta manera, se busca explicar cuáles son las más usadas, qué caracteriza a cada una y en qué contextos se han empleado. El primer conjunto de todos, RAVDESS, es descrito en mayor detalle debido a que ha sido el seleccionado para entrenar y evaluar la red neuronal de una sola capa.

2.1.1. RAVDESS

RAVDESS (*Ryerson Audio-Visual Database of Emotional Speech and Song*) es un conjunto de datos de audio público y gratuito. Aunque no fue oficialmente publicado hasta 2018,[10] su creación data de varios años antes, siendo primeramente referenciado en 2012.[11] Está compuesto por 7356 muestras de audio y clips de video, de las cuales 1440 son grabaciones de habla. Estos archivos de audio están registrados por 24 actores profesionales (12 hombres y 12 mujeres) que leen dos frases en inglés: "*Kids are talking by the door*" y "*Dogs are sitting by the door*". Estas oraciones fueron seleccionadas por la similaridad de sus longitudes, fonemas y sílabas. Cada actor pronuncia cada oración un total de ocho veces, expresando una emoción distinta en cada una: **neutral, calma, felicidad, tristeza, enojo, miedo, disgusto y sorpresa**. Finalmente, todas las frases, a excepción de las del estado emocional "neutral", se pronuncian en dos niveles de intensidad distintos: **neutral y fuerte**. Las del estado emocional "neutral" solo aparecen en intensidad "neutral", habiendo así la mitad de oraciones de este tipo comparado a las otras siete emociones.[10]

Los archivos de audio son del formato wav. Su nomenclatura indica los atributos del mismo:

- **Modalidad** (01 = full-AV, 02 = solo vídeo, 03 = solo audio).

- **Canal vocal** (01 = habla, 02 = canción).
- **Emoción** (01 = neutral, 02 = calma, 03 = felicidad, 04 = tristeza, 05 = enojo, 06 = miedo, 07 = disgusto, 08 = sorpresa).
- **Intensidad emocional** (01 = normal, 02 = fuerte). NOTA: No hay intensidad ”fuerte” para el estado emocional "neutral".
- **Oración** (01 = "Kids are talking by the door", 02 = "Dogs are sitting by the door").
- **Número de repeticiones** (01 = Primera repetición, 02 = Segunda repetición).
- **Actor** (01-24. Los pares son hombres y los impares mujeres).

De tal forma, el archivo “03-01-01-01-01-01-01.wav“ tiene los siguientes atributos:

ATRIBUTO	**VALOR**
Emoción	Neutral
Modalidad	Audio
Canal vocal	Habla
Intensidad emocional	Neutral
Oración	"Kids are talking by the door"
Número de repeticiones	1
Actor	1 (Mujer)

Cuadro 2.1: Archivo “03-01-01-01-01-01-01.wav“

2.1.2. Otras bases de datos

RAVDESS no es la única base de datos utilizada en problemas de reconocimiento de emociones a partir de voz. A continuación, se listan y describen algunas de las más notorias:

- **EmoDB**[12]: *Emotional Database* está formada por cuatro bases de datos alemanas que componen un total de 535 muestras de voz. Cada archivo de audio recoge la voz de uno de entre 10 actores alemanes nativos de los cuales la mitad son mujeres y la otra mitad hombres. Cada uno es encargado de interpretar 10 frases distintas (5 largas y 5 cortas) 6 veces cada una, simulando así una emoción distinta en cada una. Las emociones fundamentales son las siguientes: ira, felicidad, tristeza, miedo, asco y aburrimiento. Al haber sido publicada en 2005, ha sido más utilizada que RAVDESS. Algunos de los trabajos más destacados son Wang *et al.*[13], Gao et al.[14] y Zhao *et al.*[15]. Los clasificadores convencionales de ML mayormente empleados con este conjunto son asMFCC (*Mel Frequency Cepstral Coefficients*) y LPCC (*Linear Prediction Cepstral Coefficient*).[16]
- **MELD**[17]: Como su nombre sugiere, el *Multimodal EmotionLines Dataset* es una extensión de otra base de datos llamada *EmotionLines*[18] publicada también en el mismo año, 2018. Además, se trata de un conjunto multimodal, es decir, que incluye texto, audio, y video. Sin embargo, la principal diferencia respecto a los conjuntos mencionados anteriormente es que MELD está compuesto por diálogos extraídos de programas de televisión y películas, lo que aporta una

mayor variedad de contextos y situaciones emocionales. MELD está principalmente compuesta por 1300 audios de 1433 diálogos de la serie de televisión *Friends* (1994). Estos audios están etiquetados con ocho emociones principales, incluyendo felicidad, tristeza, ira, sorpresa, miedo, disgusto, neutralidad y desprecio.[17] Aunque contiene un mayor número de muestras que RAVDESS y EmoDB, no ha sido tan ampliamente utilizado.

- **IEMOCAP**[19]: A diferencia de las bases de datos anteriores que presentan expresiones emocionales individuales en un contexto más controlado, *Interactive Emotional Dyadic Motion Capture* se centra en interacciones emocionales entre dos personas en tiempo real, proporcionando así un contexto más dinámico y natural. Al tratarse de conversaciones interactivas, las emociones adquieren matices y mayor complejidad. Además, los audios están interpretados en distintos idiomas, incluyendo tanto archivos de vídeo como habla y captura de movimiento facial. En total, suponen más de 12 horas de contenido recogiendo 8 emociones distintas: neutralidad, felicidad, tristeza, ira, miedo, disgusto, sorpresa y excitación. Se trata de una base de datos muy completa y de gran calidad. Por estos motivos, ha sido muy utilizada en diversos trabajos. Por ejemplo, es referenciado en Yenigalla *et al.*,[20], Sarma *et al.*[21] y Zhiyun Lu *et al.*[22] Aunque ha sido empleado mediante diversas técnicas, actualmente es mayormente utilizado con *redes neuronales convoluconales* (*Convolutional Neural Networks*, CNN).[23][16]

- **Base de datos emocional de RML (Ryerson Multimedia Research Lab)**[24]: Incluye 720 muestras de expresiones emocionales audiovisuales recopiladas en el Laboratorio Multimedia de Ryerson en 7 idiomas diferentes. Consta de seis sentimientos humanos fundamentales: Ira, Asco, Miedo, Felicidad, Tristeza, Sorpresa.

- **TESS**[25]: El *Toronto emotional speech set* fue creado por la Universidad de Toronto en 2011. En él, dos actrices de diferente edad interpretran un conjunto de 200 palabras distintas introducidas por la frase portadora "*di la palabra...*"(*say the word...*). Este proceso se repite con siete emociones (ira, asco, miedo, felicidad, sorpresa agradable, tristeza y neutral), formando en total 2800 muestras de voz. Se utiliza principalmente para testear modelos de DL, aunque también se usa con métodos convencionales de ML.[16]

DATASET	**AUTOR**	**AÑO**	**EMPLEADO POR**
EmoDB [12]	Bukhardt	2005	Wang[13], Gao[14], Zhao[15].
IEMOCAP [19]	Busso	2008	Yenigalla,[20], Sarma[21], Zhiyun Lu[22].
TESS [25]	Dupuis	2011	Slimi[9]
RML [24]	Xie	2013	Slimi[9]
EmotionLines [18]	Chen	2018	Poria[17]
MELD [17]	Poria	2018	Gu[26], Ho[27].
RAVDESS [10]	Russo	2018	Popova[28], Jannat[29], Matin[30].

Cuadro 2.2: Estado del arte de las bases de datos.

2.2. Reconocimiento de emociones a partir de voz

Esta sección se centra en los distintos algoritmos que se han propuesto para clasificar datos en sistemas SER. En específico, aquellos que emplearon el conjunto de datos RAVDESS. Esta cribada de artículos se realiza con el fin de poder ahondar en mayor detalle en los que ya han utilizado la base de datos que se ha seleccionado para la Shallow NN. Además, se han dividido en dos subsecciones. Una para algoritmos convencionales de ML y otra para algoritmos de DL.

Los resultados y características de los algoritmos de estas dos subsecciones se pueden ver resumidos en los Cuadros 2.3 y 2.4 respectivamente. En ellos, se ordenan los modelos según el número de emociones que emplean para clasificar los datos (columna "Nº E" en los cuadros) y la precisión conseguida por cada uno ("%").

2.2.1. Algoritmos convencionales de aprendizaje automático

SVM (*Support Vector Machine*) es el enfoque de ML convencional más versátil y utilizado en el ámbito de los SER. Esto es debido al amplio número de variaciones que se han empleado en la última década.[16] En primer lugar, destaca un artículo de 2015 de Zhang *et al.*[31] Este trabajo se caracteriza por el uso de un grafo dirigido acíclico (*Directed Acyclic Graph*, DAG) que recibe como datos de entrada ondas Morlet. De las 8 emociones totales que componen el conjunto RAVDESS, Zhang y colaboradores únicamente utilizaron 6, alcanzando un 42,48 % de precisión.

Tan solo un año más tarde, Shegokar y Sircar[32] consiguieron mejorar sus resultados un 17,62 %. Su trabajo se centra en extraer las características relevantes mediante una transformación continua wavelet de Morlet. Utilizan como entrada *coeficientes cepstrales en las frecuencias de Mel* (*Mel-Frequency Cepstral Coefficients*, MFCC) y como clasificador SVM cuadrático. Además, fueron de los primeros en utilizar las 8 emociones del dataset. Sin embargo, aunque el modelo es capaz de clasificar la mayoría de las emociones con mucha precisión, obtiene peores resultados con las muestras de "Tristeza" y "Sorpresa".

En 2017, Gao *et al.*[14] propusieron como clasificador una combinación de SVM lineal junto con una selección de técnicas de extracción de patrones implícitos en los datos más amplia. Al igual que en los dos trabajos mencionados anteriormente, Gao y colaboradores utilizan coeficientes MFCC. Sin embargo, a diferencia del último, eliminan las muestras "Neutral" y "Sorpresa". Gracias a todos estos cambios, logran que la precisión alcance un 79,28 %.

Finalmente, se destaca el trabajo de 2020 de Matin y Valles[30] donde proponen el uso de SVM con con un kernel RBF (*Radial Basis Function*) y una combinación de MFCC y ZCR para clasificar emociones, aplicándolo en el contexto del tratamiento del autismo. El objetivo de su sistema es ayudar a los niños con TEA a reconocer emociones. El modelo se utiliza para entrenar a estos niños de manera que puedan identificar emociones humanas en conversaciones orales. A pesar de usar 7 emociones, el modelo alcanza una precisión del 77,38 %

Aunque SVM es el clasificador convencional más utilizado en sistemas SER, hay otros modelos que se han propuesto a lo largo de los años. Sin embargo, sus resultados no están al mismo nivel que los de SVM. Por ejemplo, en 2019, Iqbal y Barua[33] publicaron un artículo donde emplean *Gradient Boosting* para clasificar los datos de

RAVDESS. A pesar de utilizar solo 4 emociones, su precisión no consigue superar el 67,14 %, un resultado inferior al que Gao *et al.*[14] habían conseguido dos años antes con 6 emociones. Por otro lado, Zamil *et al.*[34] propone un *árbol de modelo logístico* (*Logistic Model Tree*, LMT) que emplea MFCC y 7 emociones. Este clasificador consiguió un 60,10 %.

AUTOR	AÑO	CLASIFICADOR	ATRIBUTOS	Nº E.	%
Iqbal	2019 [33]	GradientBoosting	MFCCs, energy, spectral	4	63.25
Zhang	2015 [31]	SVM(DAG)	MFCCs, energy, spectral, etc	6	42.48
Gao	2017 [14]	SVM(linear)	MFCCs, intensity, pitch, etc	6	79.28
Zamil	2019 [34]	LMT	MFCCs	7	67.14
Matin	2020 [30]	SVM(RBF)	MFCCs, ZCR	7	77.38
Shegokar	2016 [32]	SVM(quadratic)	Morlet wavelet	8	60.10

Cuadro 2.3: Estado del arte de los algoritmos de ML convencionales con el conjunto RAVDESS.

2.2.2. Algoritmos de aprendizaje profundo

A parte de los modelos que emplean algoritmos convencionales de ML, se han publicado muchos otros centrados en aprendizaje profundo. A diferencia de en la primera subsección que estaba liderada por variaciones de SVM, en las aplicaciones de DL se encuentra una mayor variedad. Además, los resultados de DL son notablemente mejores, sobretodo teniendo en cuenta que tienden a utilizar el conjunto RAVDESS entero y no solo algunas emociones para clasificar los datos. Sin embargo, este no ha sido siempre el caso. En 2018 Jannat *et al.*[29] propuso una arquitectura Inception-v3 con espectogramas básicos. A pesar de utilizar únicamente 2 emociones, sus resultados no superan el 66,41 %. Por otro lado, ese mismo año, Popova *et al.*[28] publicó un artículo donde empleaban una arquitectura VGG-16 y espectrogramas de Mel como entrada de la red convolucional. Este modelo es capaz de clasificar las 8 emociones con una precisión del 62.57 %.

Un año más tarde, Huang y Bao *et al.*[35] propusieron una arquitectura distinta. Una AlexNet que utiliza MFCCs como características, obteniendo un 52.72 %. Justifican la aplicación de capas convolucionales 2D con tales características porque combinan la información de los coeficientes en un eje y su valor en el otro eje. En 2022, José Antonio Nicolás también utilizó una variante de AlexNet en su Trabajo de Final de Máster para clasificar los datos de RAVDESS.[36] Su modelo híbrido sustituye las dos últimas capas densas por una de *memoria a corto y largo plazo* (*Long Short-Term Memory*, LSTM) para intentar reconocer patrones temporales ocultos en los espectrogramas de Mel. Este modelo consigue clasificar los datos con un 89,58 % de precisión. Por otro lado, Issa *et al.*[37] utiliza una CNN de una sola capa capaz de obtener hasta un 71,61 %. Unos años más tarde, De Lope *et al.*[8] crearon un modelo híbrido que combinaba una CNN de *tiempo distribuido* (*Time Distributed*, TD) con otra NN de LSTM. Esta variante de CNN consigue un 73,98 % de precisión.

Finalmente, se destaca el trabajo de 2019 de Zeng *et al.*[38] donde se utiliza una variante de ResNet llamada G-ResNet. A diferencia de los resultados que se obtendrían en años posteriores con otros algoritmos, Zeng *et al.* solo consiguen un 64.48 %.

AUTOR	**AÑO**	**CLASIFICADOR**	**ATRIBUTOS**	**Nº E.**	**%**
Jannat	2018 [29]	Inception-v3	Señales de audio sin procesar	2	66.41
Huang	2019 [35]	AlexNet	MFCCs	8	52.72
Popova	2018 [28]	VGG-16	Espectogramas Mel	8	62.57
Zeng	2019 [38]	G-ResNet	Espectogramas	8	64.48
Issa	2020 [37]	1D-CNN	MFCCs, Mel, etc	8	71.61
De Lope	2022 [8]	TD-CNN+LSTM	Espectogramas Mel	8	73.98
Slimi	2020 [9]	1-hidden-layer-NN	Espectogramas Mel	8	77.50
Nicolás	2022 [36]	AlexNet+LSTM	Espectogramas Mel	8	89.58

Cuadro 2.4: Estado del arte con el conjunto RAVDESS. DL.

Capítulo 3

Desarrollo

3.1. Implementación de técnicas anteriores

En este capítulo se describe el modelo propuesto, una red neuronal de una sala capa, al igual que cómo se procesan y filtran los datos empleados en su entrenamiento y evaluación. También se describen los diversos experimentos que se han realizado para seleccionar los hiperparámetros óptimos.

3.1.1. Generación de espectogramas

Al igual que en muchos otros trabajos,[20][9] se opta por realizar una conversión de los archivos de audio a espectogramas antes de utilizarlos para alimentar la red neuronal.[9] Como su nombre indica, un espectrograma es una representación visual del espectro de frecuencias de una señal a medida que varía con el tiempo. Es decir, una represantación visual bidimensional de un archivo de audio. Este formato es ampliamente utilizado en las investigaciones debido a la gran eficiencia que ha demostrado en los sistemas SER.[20]

El formato de la imagen depende del valor de varios parámetros como el tamaño de la ventana o el solapamiento. En el trabajo de Anwer Slimi, se decide re-muestrear los datos a 22050 Hz para después generar el espectograma utilizando los mismos parámetros que los de un trabajo anterior de Promod Yenigalla titulado *Speech Emotion Recognition Using Spectrogram & Phoneme Embedding*[20]. Estos cambios permiten crear espectogramas-mel, lo cual enfatiza las frecuencias bajas sobre las altas, similar a la capacidad perceptual del oído humano.[9] Los valores son los siguientes:

- **Frecuencia de muestreo**: 22050 Hz.
- **Longitud de ventana**: 2048.
- **Solapamiento o longitud de salto**: 512.
- **Escala de frecuencia**: mel.

3.1.2. Redimensión de espectogramas

De acuerdo a Slimi *et al.* (2020), los espectogramas generados por su sistema SER tienen las dimensiones 3600x2400. Esto supone un total de 8.640.000 pixels por

cada imagen, lo cual es mucho para una red neuronal, en especial una de una sola capa. Por este motivo, la imagen es reducida a una de 150x66 pixels antes de ser alimentada a la red.[9]

Sin embargo, estas dimensiones son incongruentes con lo descrito en el apartado anterior. Si se utilizan únicamente los parámetros especificados para la longitud de la ventana y salto, las imágenes generadas tienen 128 pixels de alto y un promedio de 140 de ancho (dependiendo de la longitud del audio). Una opción es redimensionar las imáganes de 128x140 a 3600x2400, pero esto es perjudicial para el rendimiento, dado que al hacer la imagen más grande, esta está añadiendo nuevos pixels que no existían originalmente, lo cual solo genera más ruido para el clasificador.

En este trabajo, se han planteado tres posibles opciones:

- **Primer método**: La imagen de 128x140 se transforma a una de 3600x2400 y luego a una de 150x66.
- **Segundo método**: La imagen de 128x140 se transforma a una de 150x66 directamente.
- **Tercer método**: El espectograma 128x140 es redimensionado a uno de 90x98, respectando así su aspect-ratio.

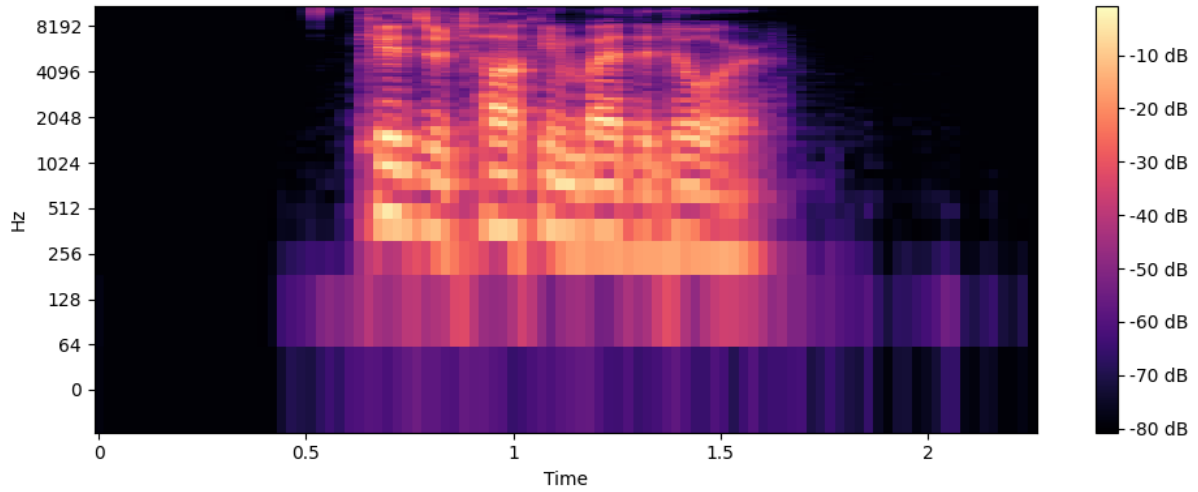

Figura 3.1: Espectograma con dimensiones 90x98 correspondiente al archivo "03-01-01-01-01-01-01.wav".

Otro problema que plantea el uso de RAVDESS es la longitud de los archivos de audio. Hay diferencias de varios segundos entre los audios más largos y los más cortos, generados principalmente por la inclusión de intervalos de silencio antes y después de recitar la oración. Esto ocasiona dos problemas. En primer lugar, el ancho de la imagen depende de la longitud del audio. Cuanto más dure la grabación, más ancha será la imagen, independientemente del valor que se haya usado en el muestreo, solapamiento, etc. El que no todas las imágenes tengan las mismas dimensiones impide poder usar una red neuronal para el entrenamiento, pues el número de neuronas en la capa de entrada debe ser igual al número de características, en este caso, pixels, que se le suministran al intertar clasificar cada imagen. En segundo lugar, incluso

si todos los archivos durasen lo mismo, la inclusión de silencios seguiría suponiendo un problema, dado que no hace nada más que añadir “ruido” y pixels extras que no aportan ninguna información sobre la emoción de dicha frase. Para solventar estos problemas, en este trabajo se opta por eliminar los intervalos de silencio después de generar el espectrograma y luego redimensionarlo a 128x140, ya que 140 representa el ancho promedio de las imágenes una vez eliminado el silencio.

3.1.3. Red neuronal de una capa oculta

Una red neuronal de una sola capa oculta es una red compuesta únicamente por una capa de entrada, una capa oculta y una capa de salida. La capa de entrada toma los valores de todos los píxeles del espectrograma, lo que significa que el número de nodos de esta capa es igual a 9900 en el ”Primer método” y ”Segundo método” dado que las dimensiones son 150x66, mientras que en el ”Tercer método” vale 8820, pues las imágenes son de 90x98.

Slimi *et al.* emplean los hiperparámetros listados en el Cuadro 3.1 para entrenar la red neuronal.[9] En este trabajo se han utilizado los mismos valores a excepción del número de neuronas y epochs, los cuales han sido optimizados.

HIPERPARÁMETROS	**VALOR**
Número de capas ocultas	1
Número de nodos	2000
Learning rate	0.0001
Optimizador	Adam (beta1=0.9, beta2=0.999)
Batch size	128
Epochs	2600
Dropout	0.5
Función de activación en capa oculta	ReLu
Función de activación en capa salida	Softmax

Cuadro 3.1: Hiperparámetros usados por Anwer Slimi (2020).

3.1.4. Tamaño del conjunto de entrenamiento

La base de datos RAVDESS cuenta con 1447 muestras, de las cuales únicamente 96 pertenecen al subconjunto ”Neutral”. Dado que las otras siete emociones tienen en torno a 192-197 muestras cada una, no se cumple una distribución uniforme en el espacio muestral. La falta de ejemplos ”neutrales” es debida a que dicha emoción solo aparece con el nivel de intensidad "neutral", a diferencia de las otras, las cuales también fueron grabadas con intensidad ”fuerte”.[10]

Lo razonable es duplicar de algún modo los ejemplos ”neutrales” para evitar sesgar el clasificador en su contra. Sin embargo, dado que el propósito principal de este trabajo es medir el alcance de Shallow ANN, se decide no implementar técnicas complementarias como Data Augmentation. Para lidiar con el problema de la distribución de las clases, se ha planteado realizar el experimento con tres conjuntos distintos:

- **Original**: 1447 muestras sin modificar el conjunto (96 son ”neutrales”).
- **Neutral duplicado**: 1543 muestras. (192 son ”neutrales”).

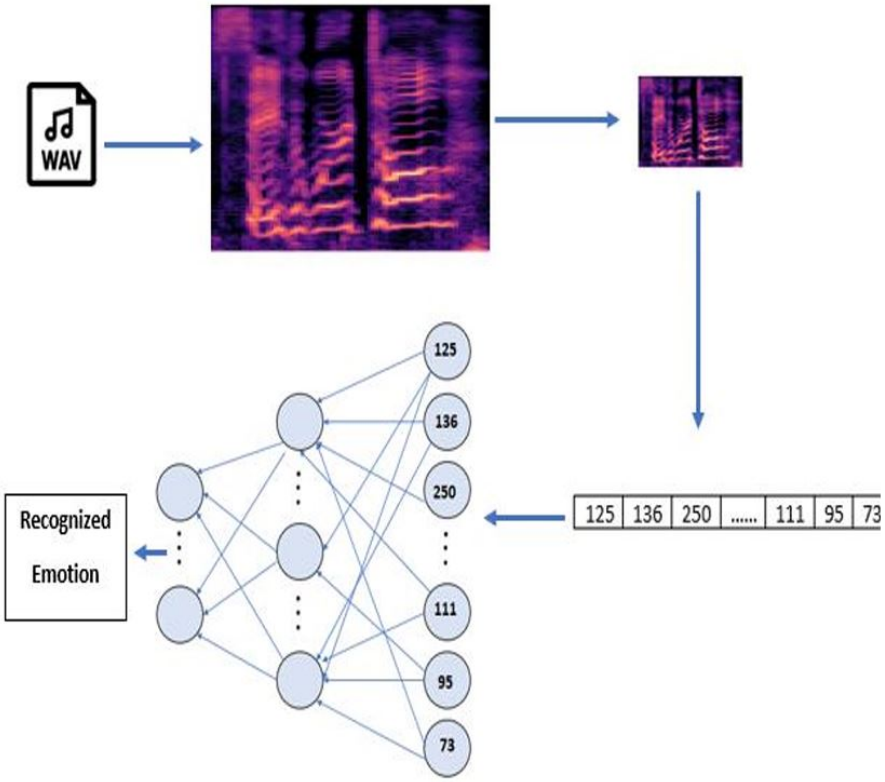

Figura 3.2: Modelo propuesto. Imagen tomada de Slimi (2020)

- **Sin neutral**: 1351 muestras. (Se eliminan las emociones "neutrales", por lo tanto, la red neuronal pasa a tener 7 neuronas en la capa de salida en vez de 8).

EMOCIÓN	NÚMERO DE MUESTRAS
Neutral	96 (192 cuando se duplica)
Calma	197
Felicidad	194
Tristeza	192
Enojo	192
Miedo	192
Disgusto	192
Sorpresa	192

Cuadro 3.2: Distribucón de los datos RAVDESS

Independientemente del método seleccionado, el conjunto de datos se divide en tres subconjuntos: train, test, y validation, siguiendo una proporción del 80-10-10 % respectivamente. El 80 % de los datos se usan en el entrenamiento y el 10 % al final para evaluar la precisión del modelo una vez este ha sido entrenado.

3.2. Ajuste de parámetros

A diferencia de la sección anterior que estaba centrada en aplicar técnicas anteriores, en específico, generar los espectogramas, lidiar con los problemas que plantea la distribución de las clases de RAVDESS, y desarrollar una red neuronal básica similar a la de Slimi *et al.*[9], en esta sección se realiza un trabajo específico. Esta sección está dividida en dos partes. En la primera, se realiza una optimización de hiperparámetros, mientra que en la segunda, se lleva a cabo un filtrado de características según su importancia para determinar si realmente se necesita el espectrograma completo para clasificar las emociones o solo algunos píxeles.

3.2.1. Optimización de hiperparámetros

Tras haber desarrollado una primera versión de la red neuronal en la sección anterior, se decide realizar un estudio de sus hiperparámetros para optimizarlos. Tras realizar un pequeño estudio preliminar, se determina que los hiperparámetros más significativos son "Neuronas" Y "Epochs". Por tanto, el estudio se centra únicamente en esos dos, y cómo sus valores mejoran o empeoran la precisión del modelo. Los valores de las "Neuronas" que se prueban son {800, 1000, 1200, 1500, 1600}. Por otro lado, los de "Epochs" son {40, 50, 80, 100, 150, 200}. Para garantizar que los resultados son consistentes, se repite el entrenamiento y evaluación de la red un total de 10 veces por cada combinación posible de dichos parámetros. Similarmente, se seleccionan unos conjuntos train, validation y test distintos para cada prueba.

En esta parte del trabajo se utilizan únicamente imágenes de tamaño 90x98, las cuales, a diferencia de las del 'Tercer método', han sido generadas fijando los parámetros **n_mels=128** y **fmax=8000** en el método **librosa.feature.melspectrogram(...)**. Esta decisión se toma debido a que los estudios preliminares determinaron que la precisión obtenida con imágenes de 90x98 no varía mucho en comparación con las de 150x66. Además, al tener que realizar un estudio tan exhaustivo tanto en esta como en la siguiente sección, resulta más cómodo utilizar imágenes más pequeñas que consumen menos recursos.

Finalmente, el conjunto de datos que se ha empleado es "Neutral Duplicado".

Los resultados experimentales se pueden ver resumidos en el Cuadro 4.1.

3.2.2. Importancia de las características

Además de intentar optimizar los valores de los hiperparámetros, también se opta por estudiar si una selección previa de las características, es decir, píxeles, más importantes puede mejorar la precisión del modelo. Para lograr este objetivo, se utilizan dos métodos de selección distintos: **Gradient Boosting** y **Random Forest**. También se repite el experimento dos veces. En la primera, se utilizan los 4000 píxeles más importantes seleccionados por cada modelo. En la segunda, únicamente 1300.

De manera similar a la sección de optimización de hiperparámetros, se entrena la red un total de 10 veces por cada tupla de valores {Neuronas, Epochs} posible, utilizando unos conjuntos de entrenamiento, validación y prueba distintos cada vez.

Como se puede observar en las Figuras 3.3 y 3.4, Gradient Boosting y Random Forest asignan distintas importancias a cada pixel en la imagen 90x98. Ni siquiera concuer-

dan los diez más importantes en cada una, como se puede ver en los Cuadros 3.3 y 3.4.

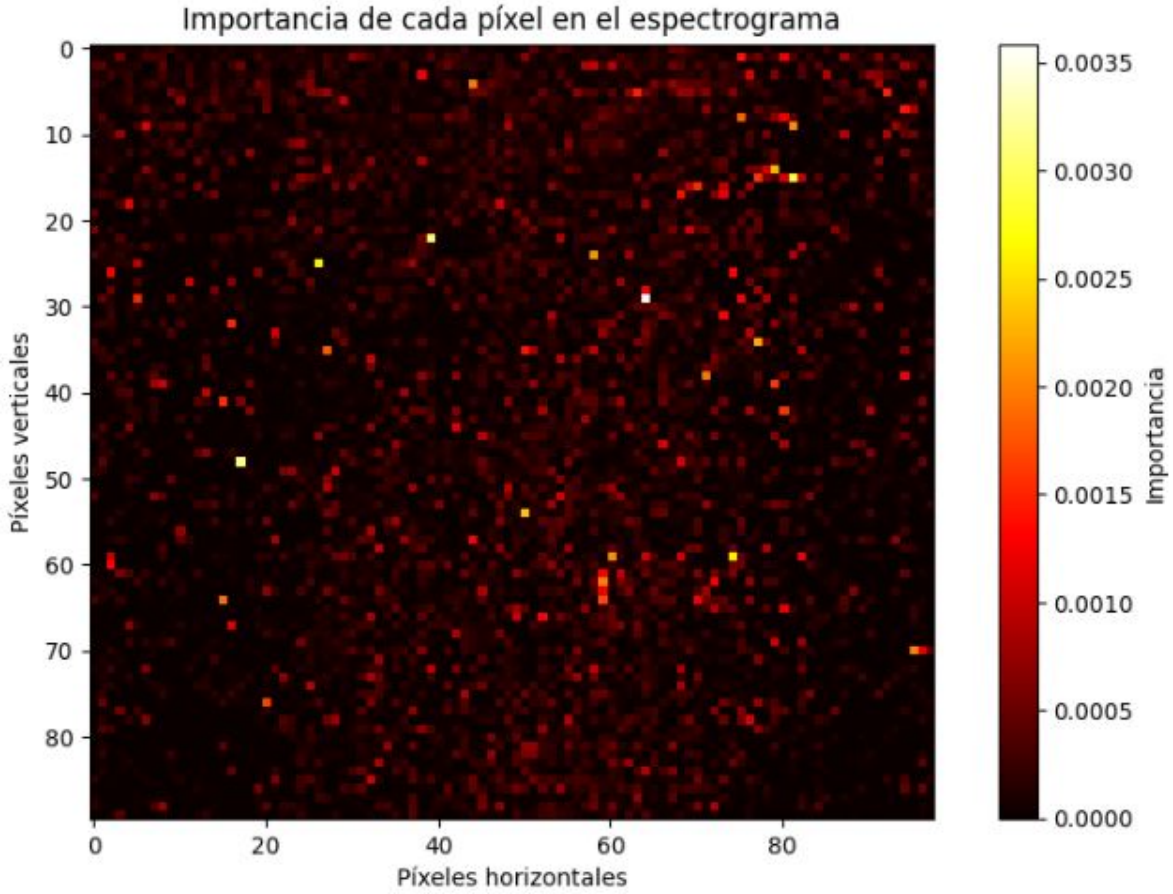

Figura 3.3: Importancia de los pixels de acuerdo a Gradient Boosting.

PIXEL	IMPORTANCIA
2906	0.003586303908377886
4721	0.0031632352620363235
2195	0.003060118993744254
1551	0.0028829402290284634
2476	0.0026183081790804863
5856	0.0026070845779031515
5342	0.0023866998963057995
1451	0.002293357625603676
3409	0.0022261361591517925
2410	0.0021061731968075037

Cuadro 3.3: Los diez pixels más importantes de acuerdo a Gradient Boosting.

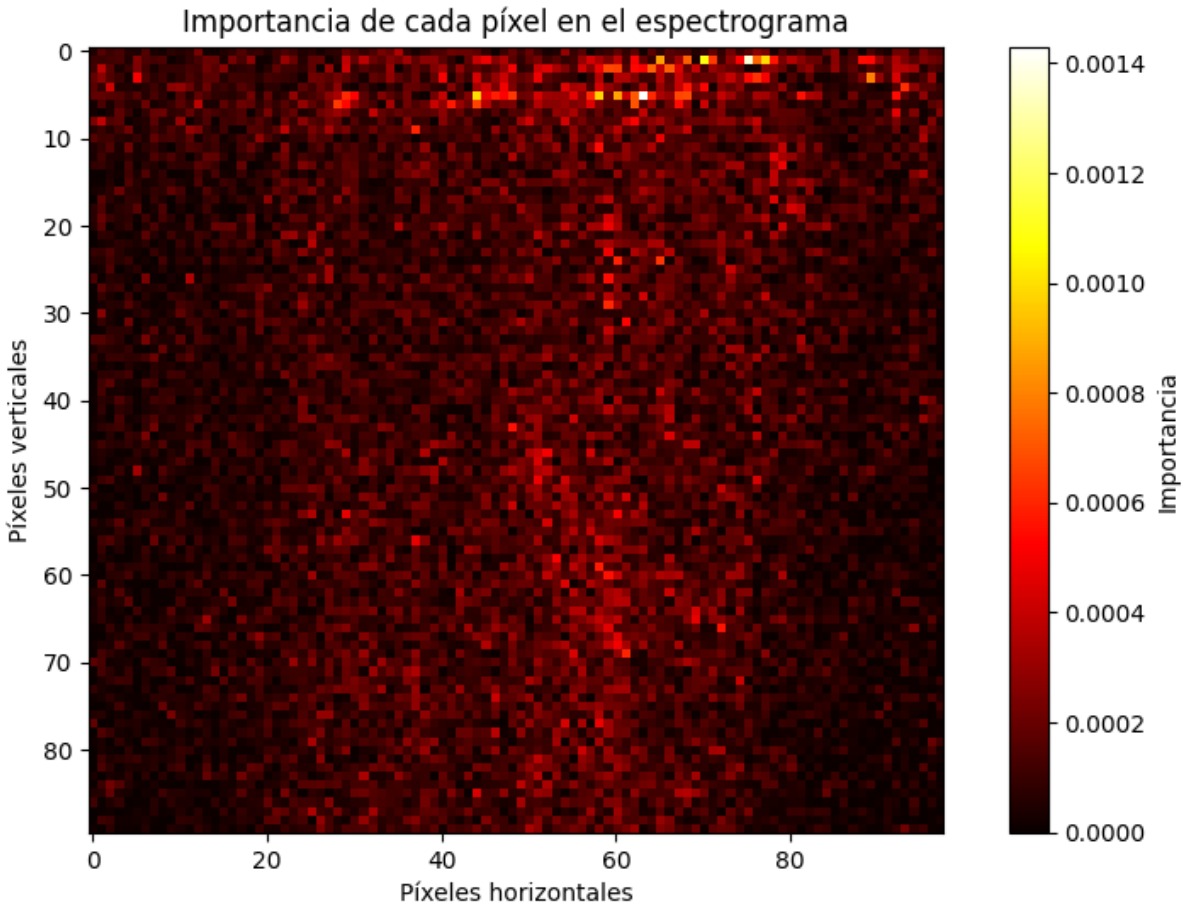

Figura 3.4: Importancia de los pixels de acuerdo a Random Forest.

PIXEL	**IMPORTANCIA**
553	0.0014299496656879414
173	0.0013706049448599029
168	0.0010852468087251153
534	0.0010026489406149426
175	0.0009891245228386735
548	0.0009754510826263263
550	0.0008742662328082703
163	0.0008663621449017998
383	0.000797924087741721
174	0.00073264667508195

Cuadro 3.4: Los diez pixels más importantes de acuerdo a Random Forest.

Capítulo 4

Resultados experimentales

En este capítulo se narran los resultados de los experimentos descritos en el capítulo anterior.

4.1. Precisión del modelo

En primer lugar, se describe la precisión de los distintos modelos que se han planteado, al igual que como estos mejoran o empeoran al añadir ciertas mejoras o cambiar el valor de los hiperparámetros.

4.1.1. Red neuronal de una capa oculta

Como se puede observar en el Cuadro 4.1, una red neuronal de una sola capa oculta es capaz de alcanzar una precisión máxima de **68,81 %** en el conjunto de prueba. Estos resultados se obtienen con con imágenes **90x98** y el conjunto **"Neutral" duplicado**. Además, solo requieren de **1000 neuronas y 80 epochs**, lo cual supone un modelo relativamente sencillo y compacto. Esto permite entrenar el modelo en tan solo unos pocos minutos. Además, se observa que la precisión del modelo solo varía un 0,43 % al utilizar 800 neuronas y 50 epochs, por lo que se podría considerar utilizar este segundo conjunto de valores en el caso de que se desee primar la velocidad de entrenamiento y consumo de recursos.

También se puede observar que la pérdida de datos en el conjunto de prueba es pequeña, siendo esta siempre menor a 1,71. Sin embargo, en las Figuras 4.1 y 4.2 se puede medir que el modelo tiene dificultades para generalizar. Mientras que la precisión en el conjunto de entrenamiento supera con creces el 90 %, el de validación nunca supera el 70 %. Es decir, el modelo presenta una alta varianza.

En el el Cuadro 4.1, se puede observar remarcado en amarillo la mejor combinación de hiperparámetros para cada valor de "neuronas" dado. En verde, aparece la tupla que alcanza el mejor resultado.

NEURONAS	EPOCHS	PRECISIÓN (%)	PÉRDIDA
800	40	67,52	1,06
800	50	68,38	0,97
800	80	67,74	1,09
800	100	68,17	1,06
800	150	67,74	1,11
800	200	67,09	1,16
1000	40	65,16	1,08
1000	50	65,8	1,04
1000	80	68,81	1,08
1000	100	68,38	1,09
1000	150	68,17	1,1
1000	200	66,02	1,54
1200	40	66,23	1,07
1200	50	68,38	1,07
1200	80	66,02	1,16
1200	100	67,52	1,12
1200	150	66,02	1,32
1200	200	66,66	1,35
1500	40	67,95	1,09
1500	50	66,88	1,06
1500	80	68,38	1,07
1500	100	67,95	1,13
1500	150	67,52	1,25
1500	200	66,32	1,54
1600	40	66,88	1,12
1600	50	66,45	1,12
1600	80	67,31	1,1
1600	100	64,73	1,31
1600	150	64,73	1,71
1600	200	67,31	1,43

Cuadro 4.1: Estudio de hiperparámetros óptimos con imágenes 90x98 (8200 pixels). En amarillo aparece remarcada la tupla {Neuronas, Epochs} óptima para cada valor de "neuronas" evaluado. En verde, la tupla con el mejor resultado (68,81 %).

Figura 4.1: Evolución de la precisión (Accuracy) de los conjuntos train y validation con hiperparametros óptimos e imágenes 90x98.

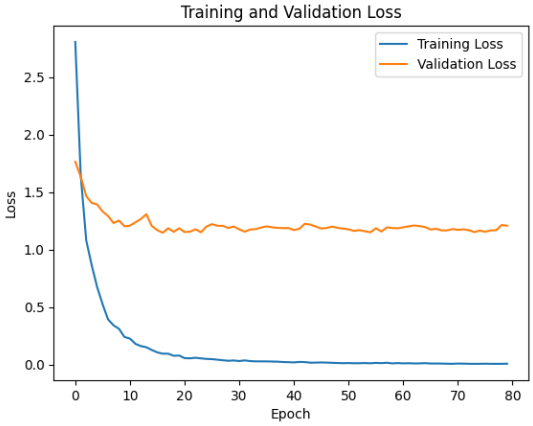

Figura 4.2: Evolución de la pérdida (Loss) de los conjuntos train y validation con hiperparámetros óptimos e imágenes 90x98.

4.1.2. Importancia de las características

Como se puede observar remarcado en verde el Cuadro 4.2, la red neuronal ha logrado un **71,09 %** de precisón al emplear únicamente las características más importantes en vez de las imágenes 90x98 completas durante el entrenamiento. Estos resultados se obtienen seleccionando los **4000 pixels** más importantes siguiendo el criterio de ordenación de **Gradient Boosting**. Sin embargo, es importante destacar que estos resultados requieren aumentar las **"neuronas" y "epochs" a 1500 y 200** respectivamente, a diferencia de en la subsección anterior, donde el resultado óptimo (68,81 %) se conseguía con tan solo 1000 neuronas y 80 epochs.

También es importante recalcar que la precisión de los modelos que utilizan 4000 pixels no solo tienden a mejorar para cada combinación de valores de la tupla {neuronas, epochs} dada, sino que muchas de ellas superan el 68,81 %. Estos casos aparecen remarcados en amarillo. Tambíen se puede observar que los resultados de Random Forest son ligeramente peores que los de Gradient Boosting.

NEURONAS	EPOCHS	IMG. COMPLETA	4000-GB	4000-RF
800	40	67,52	66,83	66,83
800	50	68,38	67,16	67,67
800	80	67,74	68,58	69,16
800	100	68,17	69,61	68,70
800	150	67,74	68,9	68,83
800	200	67,09	70,06	68,51
1000	40	65,16	68,83	66,83
1000	50	65,8	68,77	67,80
1000	80	68,81	68,51	69,09
1000	100	68,38	69,41	69,16
1000	150	68,17	70,32	68,70
1000	200	66,02	70,06	69,54
1200	40	66,23	68,51	67,80
1200	50	68,38	68,51	68,00
1200	80	66,02	68,83	68,19
1200	100	67,52	70,06	67,87
1200	150	66,02	70,38	69,80
1200	200	66,66	69,8	69,35
1500	40	67,95	68,32	69,16
1500	50	66,88	69,35	68,19
1500	80	68,38	69,48	68,77
1500	100	67,95	66,45	69,03
1500	150	67,52	70,25	69,87
1500	200	66,32	71,09	69,54

Cuadro 4.2: Estudio de hiperparámetros óptimos con 4000 pixels. En amarillo aparecen remarcados los casos que superan el 68,81 % (mejor resultado obtenido con la imagen completa). En verde, el mejor resultado posible.

En el Cuadro 4.3, aparecen los resultados obtenidos con cuando se filtran 1300 pixels en vez de 4000. Aquí se puede observar la superioridad de Gradient Boosting con mayor facilidad, dado que es el único que consigue superar el 68,81 % de la imagen completa (celdas amarillas). Random Forest, por otro lado, no solo no lo supera, si no que optiene resultados peores que la imagen completa en la mayoría de tuplas {neuronas, epochs} dada. El máximo en este caso es 70,90 %, y se consigue alcanzar con 1500 neuronas y 150 epochs.

NEURONAS	EPOCHS	IMG. COMPLETA	1300-GB	1300-RF
800	40	67,52	63,93	63,35
800	50	68,38	65,61	63,99
800	80	67,74	67,03	65,16
800	100	68,17	68,12	67,41
800	150	67,74	68,19	67,48
800	200	67,09	70,32	68,06
1000	40	65,16	65,67	64,7
1000	50	65,8	65,93	65,61
1000	80	68,81	68,19	66,51
1000	100	68,38	69,22	66,51
1000	150	68,17	69,74	67,61
1000	200	66,02	68,9	67,41
1200	40	66,23	67,22	63,93
1200	50	68,38	67,61	64,7
1200	80	66,02	69,48	67,29
1200	100	67,52	68,06	66,7
1200	150	66,02	69,29	67,54
1200	200	66,66	70,38	66,77
1500	40	67,95	67,41	65,29
1500	50	66,88	67,80	65,87
1500	80	68,38	68,90	67,16
1500	100	67,95	70,06	67,87
1500	150	67,52	70,90	67,61
1500	200	66,32	70,70	68,32

Cuadro 4.3: Estudio de hiperparámetros óptimos con 1300 pixels. En amarillo aparecen remarcados los casos que superan el 68,81 % (mejor resultado obtenido con la imagen completa). En verde, el mejor resultado posible.

En resumen, estos resultados indican que la selección previa de características importantes puede tener un impacto significativo en la mejora de los resultados. Sin embargo, también se observa que no todos los clasificadores responden de la misma manera a la selección de pixels importantes, lo cual indica la relevancia de elegir el método de selección de características más adecuado para el problema en cuestión.

4.1.3. Comparación de resultados con el estado del arte

Se destaca que los resultados obtenidos con la Shallow ANN en las subsecciones anteriores apenas superan el 71 % incluso aun habiendo implementado un método de selección de características como mejora. Esto es llamativo dado que Slimi *et al.*[9], el único trabajo anterior en el estado del arte donde se ha empleado una red neuronal de una capa, afirma haber conseguido un **77,5 %**. Por este motivo, se decide repetir el experimento empleando los mismos hiperprámetros que ellos (ver Cuadro 3.1).

En su artículo no se menciona como lidian con la falta de ejemplos "neutrales", así que, como se mencionó en el capítulo de Desarrollo, se repite el experimento tres veces: sin neutral, con neutral, y con neutral duplicado. También se han probado distintos tipos de espectogramas. Sin embargo, en ninguno de los casos se consigue reproducir sus resultados. Todo lo contrario, la precisión baja al **64,30 %** y la pérdida

aumenta a **12,73** en el mejor de los casos. Como se puede observar en las Figuras 4.3 y 4.4, su modelo presenta una varianza extrema. A partir de las primeras 300 epochs, la pérdida en el conjunto de validación no hace nada más que aumentar, lo cual plantea la duda de porqué consideraron necesario utilizar 2600 en vez de las 1000-1500 que se proponen en este trabajo.

CONJUNTO	PERDIDA	PRECISIÓN (%)
Normal	14.88	54,48
Neutral duplicado	15.44	59,35
Sin Neutral	15.05	58,82

Cuadro 4.4: Valores de pérdida y precisión del Primer método: La imagen de 128x140 se transforma a una de 3600x2400 y luego a una de 150x66.

CONJUNTO	PERDIDA	PRECISIÓN (%)
Normal	13,58	59,35
Neutral duplicado	14,10	64,30
Sin Neutral	13,73	63,79

Cuadro 4.5: Valores de pérdida y precisión del Segundo método: La imagen de 128x140 se redimensiona a una de 150x66 directamente.

CONJUNTO	PERDIDA	PRECISIÓN (%)
Normal	12,73	58,17
Neutral duplicado	13,24	63,01
Sin Neutral	12,83	62,49

Cuadro 4.6: Valores de pérdida y precisión del Tercer método: La imagen de 128x140 se redimensiona a una de 90x98 respetando el aspect ratio.

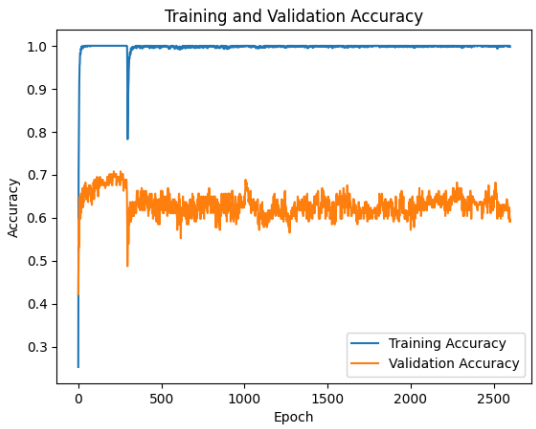

Figura 4.3: Evolución de la precisión (Accuracy) de los conjuntos train y validation con los hiperparámetros de Slimi (2020).

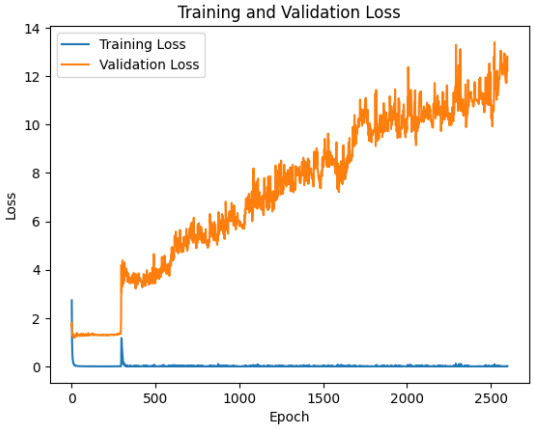

Figura 4.4: Evolución de la pérdida (Loss) de los conjuntos train y validation con los hiperparámetros de Slimi (2020).

4.2. Matriz de confusion

Una vez habiendo descrito como varía la precisión del modelo al cambiar los hiperparámetros y datos de entrada, se procede a discutir la matriz de confusión resultante de la Shallow ANN. Como se puede observar en el Cuadro 4.7 y Figura 4.5, la red neuronal optimizada con imágenes 90x98 es incapaz de clasificar las 8 emociones del conjunto de datos con la misma precisión y Recall en cada una. Al haber decidido duplicar el conjunto "neutral" en el modelo, este ha aprendido a clasificarlos relativamente bien, obteniendo 75 y 94 % en Precisión y Recall respectivamente. Es decir, el 94 % de los datos "neutrales" son clasificados como tal. Además, cuando el modelo predice que un ejemplo es 'neutral", acierta en el 75 % de los casos. Estos altos resultados en la primera emoción son probablemente causados por haber duplicado dicho subconjunto durante el entrenamiento.

Por otro lado, las emociones que peor clasifica el modelo son la "tristeza" y el "miedo", obteniendo 46 % de precisión en la primera y 50 % de Recall en la segunda. Como se muestra en la matriz de confusión de la Figura 4.5, el 11,43 y 10,71 % de los datos "tristeza" son clasificados como "neutral" y "calma" respectivamente. Es decir, el modelo tiene dificultad para diferenciar el miedo de esas dos clases. En el caso del "miedo", la confusión surge con la emoción "Disgusto", siendo esta vez el 16,88 % de los datos los que se clasifican incorrectamente.

	Precision	**Recall**	**F1-Score**	**Support**
Neutral	0.75	0.94	0.84	230
Calma	0.76	0.73	0.75	290
Felicidad	0.73	0.56	0.64	190
Tristeza	0.46	0.52	0.49	140
Enojo	0.74	0.67	0.70	220
Miedo	0.63	0.50	0.56	160
Disgusto	0.58	0.58	0.58	180
Sorpresa	0.59	0.71	0.65	140
Accuracy			0.67	1550
Macro Avg	0.66	0.65	0.65	1550
Weighted Avg	0.68	0.67	0.67	1550

Cuadro 4.7: Resultados de la clasificación de la Shallow ANN con Imágenes 90x98.

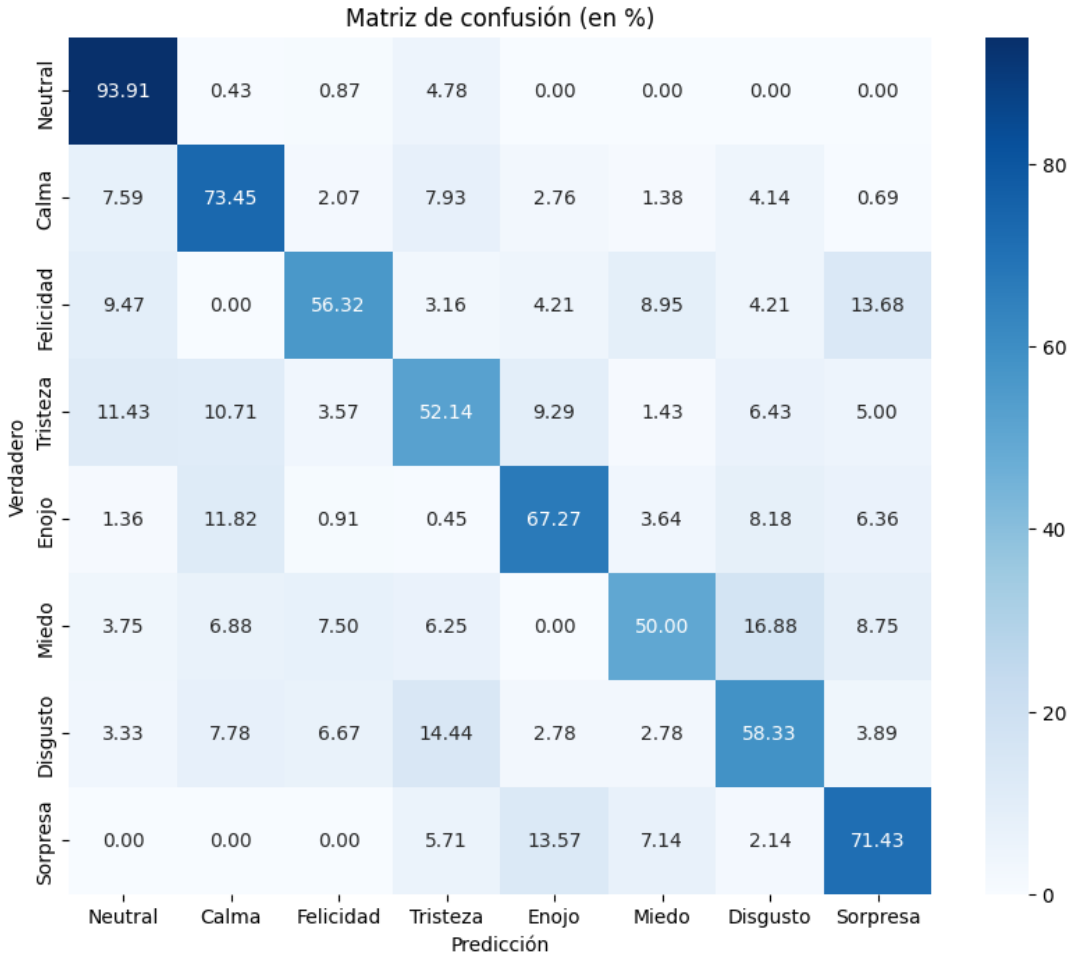

Figura 4.5: Matriz de confusión de la Shallow ANN con Imágenes 90x98.

Como se explica en la sección anterior, emplear Gradient Boosting para filtrar los 4000 pixels más importantes de las imágenes 90x98 mejora la precisión de la Shallow ANN de un 68,81 a un 71,09%. Esta mejoría no tiene solo un carácter global, sino que también se puede medir de manera local en el Cuadro 4.6 y Figura 4.8. Como se puede observar en el Cuadro 4.6, la precisión de todas las emociones mejora entre un 2 y un 6% respecto al Cuadro 4.7. La única excepción es el "Miedo", que pasa de valer 63 a 59.

	Precision	Recall	F1-Score	Support
Neutral	0.80	0.96	0.87	230
Calma	0.80	0.72	0.76	290
Felicidad	0.75	0.59	0.66	190
Tristeza	0.53	0.66	0.59	140
Enojo	0.80	0.72	0.76	220
Miedo	0.59	0.52	0.55	160
Disgusto	0.64	0.61	0.62	180
Sorpresa	0.61	0.76	0.67	140
Accuracy			0.71	1550
Macro Avg	0.69	0.69	0.69	1550
Weighted Avg	0.71	0.71	0.70	1550

Cuadro 4.8: Resultados de la clasificación de la Shallow ANN con los 4000 pixels más importantes filtrados con Gradient Boosting.

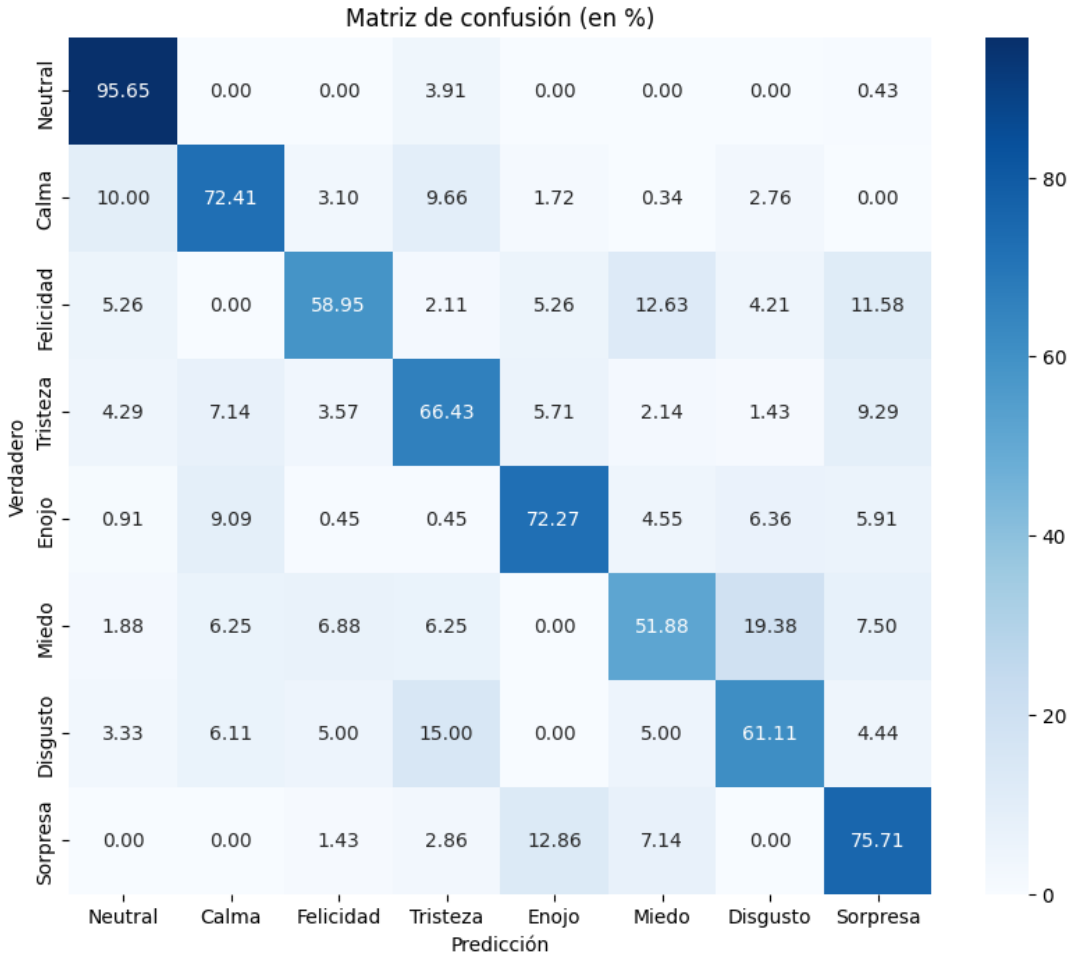

Figura 4.6: Matriz de confusión de la Shallow ANN con los 4000 pixels más importantes filtrados con Gradient Boosting.

Capítulo 5

Conclusiones

5.1. Redes neuronales de una capa oculta

En este trabajo se ha desarrollado y entrenado una red neuronal de una sola capa oculta capaz de reconocer emociones a partir de archivos de voz del conjunto de datos RAVDESS. Como se puede observar en el Cuadro 4.2, una red de 1500 neuronas y 200 epochs que utiliza los 4000 pixels más importantes de un espectograma de dimensiones 90x98 es capaz de clasificar los ejemplos con un **71,09 %** de precisión. Este resultado es un **2,28 %** mejor que los que se pueden obtener con una red que no emplea un método de selección de características según su importancia. Es decir, se ha demostrado que simplificando los datos de entrada se mejora la capacidad de clasificación del modelo.

Sin embargo, es relevante remarcar la transcendencia de elegir un buen modelo para ordenar los pixels en base a su importancia. Por ejemplo, cuando se utilizan los pixels más importantes de acuerdo a Random Forest, los resultados son peores que los que se obtienen con **Gradient Boosting**. Esto es especialmente notorio cuando se usan 1300 pixels en vez de 4000, dado que Random Forest no solo es incapaz de superar el resultado óptimo de una red sin filtrado de características, es decir, 68,81 %, sino que empeora la precisión para cada tupla de valores {neuronas, epochs} dada. Es decir, en este caso, es preferible usar la imagen 90x98 completa a filtrar 1300 pixels de acuerdo al criterio de Random Forest. Por este motivo, aunque se destaca el potencial de utilizar modelos como Gradient Boosting para reducir las características y así mejorar tanto los resultados como el consumo de recursos, se sugiere ser cuidadoso al seleccionar el modelo de filtrado y el número de características utilizadas.

También es importante recalcar que hay muchos más artículos en el estado del arte donde se presentan clasificadores superando el 71,09 % de aciertos con RAVDESS. Sin embargo, varios de ellos no emplean el conjunto de datos entero, sino que eliminan los ejemplos de algunas emociones concretas. Al reducir el número de clases en el espacio muestral, se reduce la complejidad, aumentando así las posibilidades de que el modelo aprenda a distinguir patrones y los diferentes subconjuntos. No obstante, siguen habiendo muchas propuestas que obtienen resultados inferiores independientemente de si utilizan las 8 emociones en su totalidad o no. Por ejemplo; Iqbal *et al.*[33], Zamil *et al.*[34], Jannat *et al.*[29] y Zeng *et al.*[38] que utilizaron Gradient Boosting, LMT, Inception-v3 y G-ResNet respectivamente.

En resumen, las redes neuronales de una sola capa presentan resultados aceptables en comparación a trabajos anteriores resumidos en los Cuadros 2.3 y 2.4. Mediante una buena selección de características según su importancia y una optimización de hiperperámetros, se puede alcanzar un **71,08 %** de precisión utilizando las 8 emociones del conjunto RAVDESS. Aunque estos resultados determinan que las NN de estas características son útiles y preferibles a algunos algoritmos convencionales de ML, se ratifica no son tan buenos como algunas propuestas de DL donde se ha llegado a alcanzar el 89,58 %.[36]

5.2. Propuestas para vías futuras de investigación

Habiendo explicado los resultados experimentales y como estos se comparan al estado del arte resumido en los Cuadros 2.3 y 2.4, se deben discutir las líneas futuras de investigación. Se propone investigar en mayor profundidad como varía la precisión de la red neuronal al utilizar distintos modelos para ordenar los pixels según su importancia al igual que reducir el número de pixels filtrados. En este trabajo se han realizado experimentos únicamente con Gradient Boosting y Random Forest al igual que con 4000 y 1300 pixels de una imagen 90x98. Esto supone un total de 4 casos diferentes que se han testeado únicamente con el fin de determinar si la precisión variaba significativamente. En el futuro, se debería intentar determinar que modelo y número de pixels es óptimo, al igual que si vale la pena consumir tiempo y recursos filtrando los datos en relación a la mejoría que proporciona en la precisión.

Por otro lado, aunque este trabajo no ha obtenido resultados superiores a los del estado del arte, se ha demostrado que la capacidad de abstracción de un modelo sobre el conjunto de datos RAVDESS puede mejorar más de un 2 % mediante un método de selección de características según su importancia. Por tanto, se puede investigar si estos métodos de filtrado de pixels también mejoran los resultados de otros clasificadores, en específico aquellos que han obtenido una mejor precisión.

Bibliografía

[1] N.L.W. Keijsers, Neural Networks, Editor(s): Katie Kompoliti, Leo Verhagen Metman, Encyclopedia of Movement Disorders, Academic Press, 2010, Pages 257-259, ISBN 9780123741059, https://doi.org/10.1016/B978-0-12-374105-9.00493-7.

[2] Larry Hardesty, MIT News Office, Explained: Neural networks, Ballyhooed artificial-intelligence technique known as "deep learning" revives 70-year-old idea. 2017

[3] Poggio, T., Mhaskar, H., Rosasco, L. et al. Why and when can deep-but not shallow-networks avoid the curse of dimensionality: A review. Int. J. Autom. Comput. 14, 503–519 (2017). https://doi.org/10.1007/s11633-017-1054-2

[4] Mr. N. Ratna Kanth and Dr. S. Saraswathi: A Survey on Speech Emotion Recognition. Advances in Computer Science and Information Technology (ACSIT) Print ISSN: 2393-9907; Online ISSN: 2393-9915; Volume 1, Number 3; November, 2014 pp. 135-139.

[5] P. Hajek, A. Barushka and M. Munk, Neural networks with emotion associations, topic modeling, and supervised term weighting for sentiment analysis, Int. J. Neural Syst. 31(10) (2021) 2150013.

[6] E. Delfino, A. Pastore, E. Zucchini, M. F. P. Cruz, T. Ius, M. Vomero, A. D'Ausilio, A. Casile, M. Skrap, T. Stieglitz and L. Fadiga, Prediction of speech onset by micro-electrocorticography of the human brain, Int. J. Neural Syst. 31(7) (2021) 2150025.

[7] Yafeng Niu, Dongsheng Zou, Yadong Niu, Zhongshi He, Hua Tan: A breakthrough in Speech emotion recognition using Deep Retinal Convolution Neural Networks.

[8] Javier De Lope, Manuel Graña. (2022) *A Hybrid Time-Distributed Deep Neural Architecture for Speech Emotion Recognitiong*. International Journal of Neural Systems, Vol. 32, No. 6 (2022) 2250024, doi: 10.1142/S0129065722500241

[9] Anwer Slimi, Mohamed Hamroun, Mounir Zrigui, and Henri Nicolas. 2020. *Emotion recognition from speech using spectrograms and shallow neural networks*. The 18th International Conference on Advances in Mobile Computing & Multimedia (MoMM2020), November 30- December 2, 2020, Chiang Mai, Thailand. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3428690.3429153

[10] Livingstone SR, Russo FA (2018) The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLoS ONE 13(5): e0196391.

[11] S. R. Livingstone, K. Peck, and F. A. Russo, “Ravdess: The ryersonaudio-visual database of emotional speech and song,” in Annual Meeting of the Canadian Society for Brain, Behaviour and Cognitive Science, 2012.

[12] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, B. Weiss: Database of German Emotional Speech. INTERSPEECH 2005 - Eurospeech, 9th European Conference on Speech Communication and Technology, Lisbon, Portugal, September 4-8, 2005

[13] Wang, Kunxia & An, Ning & Li, Bing Nan & Zhang, Yanyong & Li, Lian. (2015). Speech Emotion Recognition Using Fourier Parameters. Affective Computing, IEEE Transactions on. 6. 69-75. 10.1109/TAFFC.2015.2392101.

[14] Gao, Yuanbo & Li, Baobin & Wang, Ning & Zhu, Tingshao. (2017). Speech Emotion Recognition Using Local and Global Features.

[15] Jianfeng Zhao, Xia Mao, Lijiang Chen, Speech emotion recognition using deep 1D & 2D CNN LSTM networks, Biomedical Signal Processing and Control, Volume 47, 2019, Pages 312-323, ISSN 1746-8094, https://doi.org/10.1016/j.bspc.2018.08.035.

[16] Javier de Lope, Manuel Graña, An ongoing review of speech emotion recognition, Neurocomputing, Volume 528, 2023, Pages 1-11, ISSN 0925-2312, https://doi.org/10.1016/j.neucom.2023.01.002.

[17] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, Rada Mihalcea. (2018) MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. arXiv:1810.02508

[18] Sheng-Yeh Chen, Chao-Chun Hsu, Chuan-Chun Kuo, Ting-Hao (Kenneth)Huang, Lun-Wei Ku. (2018) EmotionLines: An Emotion Corpus of Multi-Party Conversations. https://doi.org/10.48550/arXiv.1802.08379

[19] Busso, C., Bulut, M., Lee, C., Kazemzadeh, A., Mower Provost, E., Kim, S., Chang, J., Lee, S., Narayanan, Shrikanth. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. Language Resources and Evaluation. 42. 335-359. 10.1007/s10579-008-9076-6.

[20] Yenigalla, P., Kumar, A., Tripathi, S., Singh, C., Kar, S., Vepa, J. (2018) *Speech Emotion Recognition Using Spectrogram & Phoneme Embedding*. Proc. Interspeech 2018, 3688-3692, doi: 10.21437/Interspeech.2018-1811

[21] Sarma, M., Ghahremani, P., Povey, D., Goel, N.K., Sarma, K.K., Dehak, N. (2018) Emotion Identification from Raw Speech Signals Using DNNs. Proc. Interspeech 2018, 3097-3101, doi: 10.21437/Interspeech.2018-1353

[22] Z. Lu, L. Cao, Y. Zhang, C. Chiu, and J. Fan. ICASSP, page 7149-7153. IEEE, (2020). Speech Sentiment Analysis via Pre-Trained Features from End-to-End ASR Models.

[23] T. Anrarjon, Kwon Mustaqeem, S.: A lightweight CNN-based speech emotion recognition system using deep system using deep, Sensorts 20 (2020) 5212.

[24] Z. Xie and L. Guan, "Multimodal Information Fusion of Audiovisual Emotion Recognition Using Novel Information Theoretic Tools", International Journal of Multimedia Data Engineering and Management, vol. 4, no. 4, pp. 1-14, 2013.

[25] Kate Dupuis, M. Kathleen Pichora-Fuller: Recognition of Emotional Speech for Younger and Older Talkers: Behavioural Findings from The Toronto Emotional Speech Set, 2011

[26] Gu, Yue & Lyu, Xinyu & Sun, Weijia & Li, Weitian & Chen, Shuhong & Li, Xinyu & Marsic, Ivan. (2019). Mutual Correlation Attentive Factors in Dyadic Fusion Networks for Speech Emotion Recognition. Proceedings of the ... ACM International Conference on Multimedia, with co-located Symposium & Workshops. ACM International Conference on Multimedia. 2019. 157-166. 10.1145/3343031.3351039.

[27] N. -H. Ho, H. -J. Yang, S. -H. Kim and G. Lee. (2020) "Multimodal Approach of Speech Emotion Recognition Using Multi-Level Multi-Head Fusion Attention-Based Recurrent Neural Network,ïn IEEE Access, vol. 8, pp. 61672-61686, 2020, doi: 10.1109/ACCESS.2020.2984368.

[28] Popova, Anastasiya & Rassadin, Alexandr & Ponomarenko, Alexander. (2018). Emotion Recognition in Sound. Studies in Computational Intelligence. 736.

[29] Rahatul Jannat, Iyonna Tynes, Lott La Lime, Juan Adorno, and Shaun Canavan. 2018. Ubiquitous Emotion Recognition Using Audio and Video Data. In Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers (UbiComp '18). Association for Computing Machinery, New York, NY, USA, 956–959. https://doi.org/10.1145/3267305.3267689

[30] Matin, Rezwan & Valles, Damian. (2020). A Speech Emotion Recognition Solution-based on Support Vector Machine for Children with Autism Spectrum Disorder to Help Identify Human Emotions. 1-6. 10.1109/IETC47856.2020.9249147.

[31] B. Zhang, G. Essl and E. M. Provost, Recognizing emotion from singing and speaking using shared models, in 2015 Int. Conf. Affective Computing and Intelligent Interaction, Xi'an, China, 2015, pp. 139-145.

[32] P. Shegokar and P. Sircar, Continuous wavelet transform based speech emotion recognition, in Int. Conf. Signal Processing and Communication Systems, Gold Coast, Australia, 2016, pp. 1–8.

[33] A. Iqbal and K. Barua, A real-time emotion recognition from speech using gradient boosting, in Proc. 2019 Int. Conf. Electrical, Computer and Communication Engineering (ECCE), Baltimore, MD, USA, 2019, pp. 1–5.

[34] A.A.A.Zamil, S. Hasan, S. M.J. Baki, J. M. Adam and I. Zaman, Emotion detection from speech signals using voting mechanism on classified frames, in 2019 Int. Conf. Robotics, Electrical and Signal Processing Technique, Bangladesh, 2019, pp. 281-285.

[35] A. Huang and P. Bao, Humanvocal sentiment analysis, arXiv:1905.08632.

[36] José Antonio Nicolás Navarro (2022). Reconocimiento de Emociones a partir de la Voz utilizando Deep Learning.

[37] D. Issa, M. Faith-Demirci and A. Yazici, Speech emotion recognition with deep convolutional neural networks, Biomed. Signal Proc. Control 59 (2020) 101894.

[38] Zeng, Yuni, Mao, Hua, Peng, Dezhong and Yi, Zhang (2019) Spectrogram based multi-task audio classification. Multimedia Tools and Applications, 78 (3). pp. 3705-3722. ISSN 1380-7501