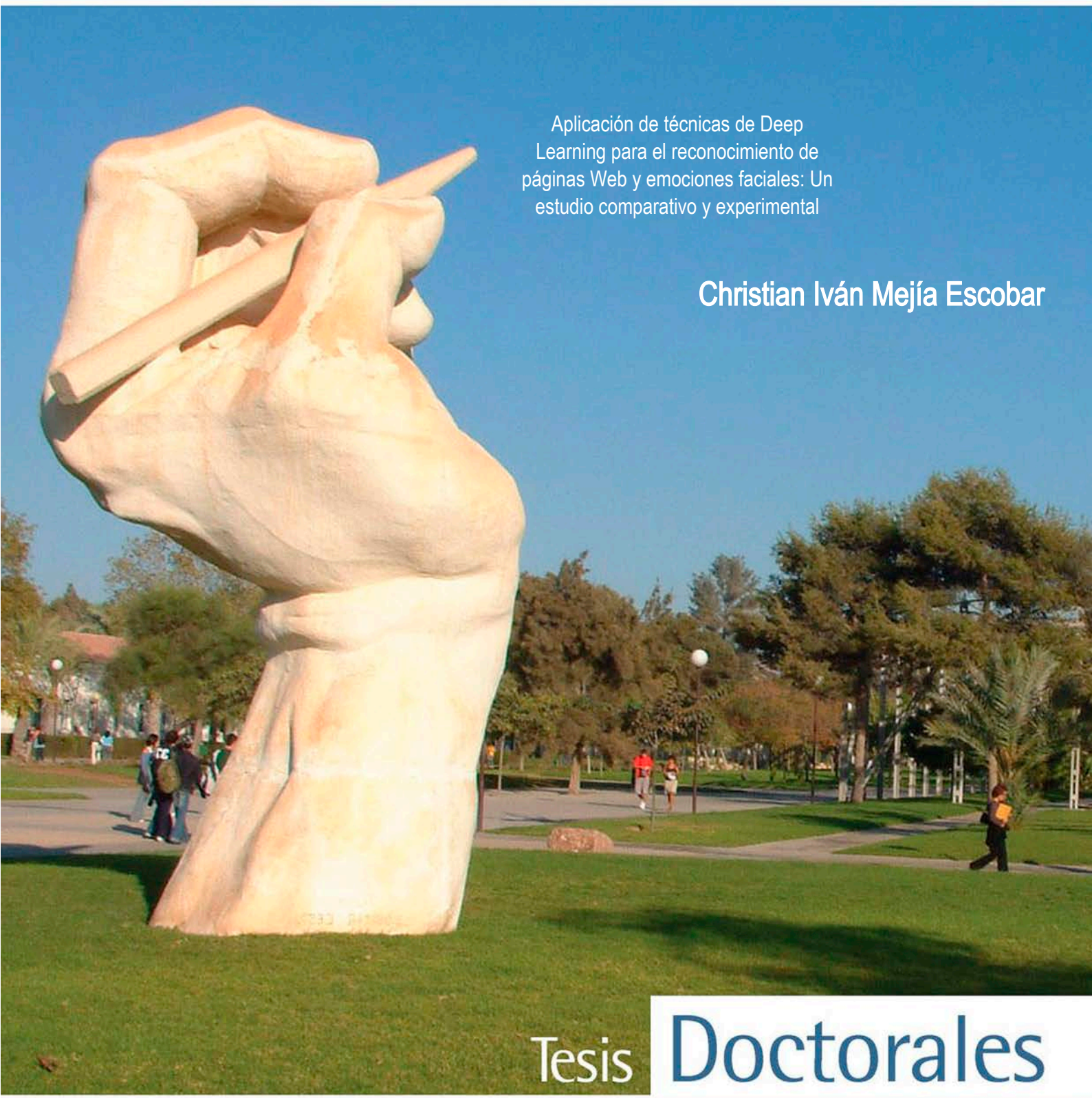
Universitat d'Alacant
Universidad de Alicante



Aplicación de técnicas de Deep Learning para el reconocimiento de páginas Web y emociones faciales: Un estudio comparativo y experimental

Christian Iván Mejía Escobar

Tesis Doctorales

UNIVERSIDAD de ALICANTE

Unitat de Digitalització UA
Unidad de Digitalización UA

Universitat d'Alacant
Universidad de Alicante

Aplicación de técnicas de Deep Learning para el reconocimiento de páginas Web y emociones faciales: Un estudio comparativo y experimental

Christian Mejía Escobar

TESIS PRESENTADA PARA ASPIRAR AL GRADO DE DOCTOR POR LA UNIVERSIDAD DE ALICANTE

PROGRAMA DE DOCTORADO EN INFORMÁTICA

Dirigida por: Dr. Miguel Ángel Cazorla Quevedo
Dra. Ester Martínez-Martín

Junio, 2023

Dedicatoria

A quienes me han acompañado en esta travesía, han sido incondicionales, y me han brindado todo su ánimo y apoyo, en especial:
Mi esposa María Dolores, por hacer de mí el ser más afortunado del planeta
Mis pequeñas hijas Paula y Triana, por ser el motor de mi existencia y a quienes espero nunca defraudar
Mi madre Mariana, por su infinito amor y sacrificio
Mi padre Luis, por su gran ejemplo académico y profesional
Mi hermana Mónica, Esteban, Mathías, Julián y Samanta, mi hermosa y amada familia.

Declaración

Por la presente declaro que, salvo cuando se hace referencia específica al trabajo de otros, el contenido de esta disertación es original y no ha sido presentado en su totalidad o en parte para la obtención de otro título o titulación en esta u otra universidad. Esta disertación es mi propio trabajo y no contiene nada que sea el resultado de un trabajo realizado en colaboración con otros, salvo lo especificado en el texto y en los Agradecimientos.

Christian Mejía Escobar
Junio, 2023

Agradecimientos

Un agradecimiento eterno al Dr. Miguel Cazorla, Catedrático de la Universidad de Alicante del más prestigioso nivel, su labor de excelencia ha sido una inspiración para aventurarme en el mundo de la inteligencia artificial y su tutoría ha sido un gran privilegio para mí.

A la Dra. Ester Martínez Martín por sus valiosas directrices y recomendaciones, que han sido fundamentales para el desarrollo de este trabajo de investigación.

Al Dr. José García Rodríguez, Coordinador del Programa de Doctorado en Informática de la Universidad de Alicante.

Agradezco a la Universidad Central del Ecuador por la beca concedida para mis estudios de doctorado.

El desarrollo y difusión científica de este trabajo de investigación ha sido financiado por el proyecto MEEBAI del programa Prometeo-CIPROM/2021/017 de la Conselleria de Innovación, Universidades, Ciencia y Sociedad Digital de la Generalitat Valenciana (España) y la Universidad Central del Ecuador a través de la certificación presupuestaria No. 34 del 25 de marzo de 2022 con el código DOCT-DI-2020-37.

Resumen

El progreso de la Inteligencia Artificial (IA) ha sido notable en los últimos años. Los impresionantes avances en imitar las capacidades humanas por parte de las máquinas se deben especialmente al campo del Deep Learning (DL). Este paradigma evita el complejo diseño manual de características. En su lugar, los datos pasan directamente a un algoritmo, que aprende a extraer y representar características jerárquicamente en múltiples capas a medida que aprende a resolver una tarea. Esto ha demostrado ser ideal para problemas relacionados con el mundo visual.

Una solución de DL comprende datos y un modelo. La mayor parte de la investigación actual se centra en los modelos, en busca de mejores algoritmos. Sin embargo, aunque se prueben diferentes arquitecturas y configuraciones, difícilmente mejorará el rendimiento si los datos no son de buena calidad. Son escasos los estudios que se centran en mejorar los datos, pese a que constituyen el principal recurso para el aprendizaje automático. La recolección y el etiquetado de extensos datasets de imágenes consumen mucho tiempo, esfuerzo e introducen errores.

La mala clasificación, la presencia de imágenes irrelevantes, el desequilibrio de las clases y la falta de representatividad del mundo real son problemas ampliamente conocidos que afectan el rendimiento de los modelos en escenarios prácticos. Nuestra propuesta enfrenta estos problemas a través de un enfoque data-centric.

A través de la ingeniería del dataset original utilizando técnicas de DL, lo hacemos más adecuado para entrenar un modelo con mejor rendimiento y generalización en escenarios reales. Para demostrar esta hipótesis, consideramos dos casos prácticos que se han convertido en temas de creciente interés para la investigación.

Por una parte, Internet es la plataforma mundial de comunicación y la Web es la principal fuente de información para las actividades humanas. Las páginas Web crecen a cada segundo y son cada vez más sofisticadas. Para organizar este complejo y vasto contenido, la clasificación es la técnica básica. El aspecto visual de una página Web puede ser una alternativa al análisis textual del código para distinguir entre categorías.

Abordamos el reconocimiento y la clasificación de páginas Web creando un dataset de capturas de pantalla apropiado desde cero.

Por otro lado, aunque los avances de la IA son significativos en el aspecto cognitivo, la parte emocional de las personas es un desafío. La expresión facial es la mejor evidencia para manifestar y transmitir nuestras emociones. Aunque algunos datasets de imágenes faciales existen para entrenar modelos de DL, no ha sido posible alcanzar el alto rendimiento en entornos controlados utilizando datasets in-the-lab. Abordamos el reconocimiento y la clasificación de emociones humanas mediante la combinación de varios datasets in-the wild de imágenes faciales.

Estas dos problemáticas plantean situaciones distintas y requieren de imágenes con contenido muy diferente, por lo que hemos diseñado un método de refinamiento del dataset según el caso de estudio.

En el primer caso, implementamos un modelo de DL para clasificar páginas Web en determinadas categorías utilizando únicamente capturas de pantalla, donde los resultados demostraron un problema multiclase muy difícil. Tratamos el mismo problema con la estrategia One vs. Rest y mejoramos el dataset mediante reclasificación, detección de imágenes irrelevantes, equilibrio y representatividad, además de utilizar técnicas de regularización y un nuevo mecanismo de predicción con los clasificadores binarios. Estos clasificadores operando por separado mejoran el rendimiento, en promedio incrementan un 26.29% la precisión de validación y disminuyen un 42.30% el sobreajuste, mostrando importantes mejoras respecto al clasificador múltiple que opera con todas las categorías juntas. Utilizando el nuevo modelo, hemos desarrollado un sistema en línea para clasificar páginas Web que puede ayudar a diseñadores, propietarios de sitios Web, Webmasters y usuarios en general.

En el segundo caso, la estrategia consiste en refinar progresivamente el dataset de imágenes faciales mediante varios entrenamientos sucesivos de un modelo de red convolucional. En cada entrenamiento, se utilizan las imágenes faciales correspondientes a las predicciones correctas del entrenamiento anterior, lo que permite al modelo captar más características distintivas de cada clase de emoción. Tras el último entrenamiento, el modelo realiza una reclasificación automática de todo el dataset. Este proceso también nos permite detectar las imágenes irrelevantes, pero nuestro propósito es mejorar el dataset sin modificar, borrar o aumentar las imágenes, a diferencia de otros trabajos similares. Los resultados experimentales en tres datasets representativos demostraron la eficacia del método propuesto, mejorando la precisión de validación en un 20.45%, 14.47% y 39.66%, para FER2013, NHFI y AffectNet, respectivamente. Las tasas de reconocimiento en las versiones reclasificadas de estos datasets son del 86.71%, el 70.44% y el 89.17%, que

alcanzan el estado del arte. Combinamos estas versiones mejor clasificadas para aumentar el número de imágenes y enriquecer la diversidad de personas, gestos y atributos de resolución, color, fondo, iluminación y formato de imagen. El dataset resultante se utiliza para entrenar un modelo más general. Frente a la necesidad de métricas más realistas de la generalización de los modelos, creamos un dataset evaluador combinado, equilibrado, imparcial y bien etiquetado. Para tal fin, organizamos este dataset en categorías de género, edad y etnia. Utilizando un predictor de estas características representativas de la población, podemos seleccionar el mismo número de imágenes y mediante el exitoso modelo Stable Diffusion es posible generar las imágenes faciales necesarias para equilibrar las categorías creadas a partir de las mencionadas características.

Los experimentos single-dataset y cross-dataset indican que el modelo entrenado en el dataset combinado mejora la generalización de los modelos entrenados individualmente en FER2013, NHFI y AffectNet en un 13.93%, 24.17% y 7.45%, respectivamente. Desarrollamos un sistema en línea de reconocimiento de emociones que aprovecha el modelo más genérico obtenido del dataset combinado. Por último, la buena calidad de las imágenes faciales sintéticas y la reducción de tiempo conseguida con el método generativo nos motivan para crear el primer y mayor dataset artificial de emociones categóricas. Este producto de libre acceso puede complementar los datasets reales, que son difíciles de recopilar, etiquetar, equilibrar, controlar las características y proteger la identidad de las personas.

Abstract

The progress of Artificial Intelligence (AI) has been remarkable in recent years. The impressive advances in mimicking human capabilities by machines are especially due to the field of Deep Learning (DL). This paradigm avoids complex manual feature design. Instead, data is passed directly to an algorithm, which learns to extract and represent features hierarchically in multiple layers as it learns to solve a task. This has proven to be ideal for problems related to the visual world.

A DL solution comprises of data and a model. Most of the current research focuses on models, in search of better algorithms. However, even if different architectures and configurations are tested, performance will hardly improve if the data is not of good quality. Few studies focus on improving data, even though it is the main resource for machine learning. Collecting and labeling large image datasets is time-consuming, labor-intensive and error-prone.

Misclassification, presence of irrelevant images, class imbalance and lack of real-world representativeness are widely known problems that affect the performance of models in practical scenarios. Our proposal addresses these problems through a data-centric approach.

By engineering the original dataset using DL techniques, we make it more suitable for training a model with better performance and generalisation in real-world scenarios. To demonstrate this hypothesis, we consider two case studies that have become topics of growing research interest.

On the one hand, the Internet is the world's communication platform and the Web is the main source of information for human activities. Web pages are growing by the second and are becoming more and more sophisticated. To organise this complex and vast content, classification is the basic technique. The visual appearance of a Web page can be an alternative to textual analysis of the code to distinguish between categories. We approach the recognition and classification of Web pages by creating an appropriate screenshot dataset from scratch.

On the other hand, although AI advances are significant on the cognitive side, the emotional side of people is a challenge. Facial expression is the best evidence to manifest and transmit our emotions. Although some facial image datasets exist to train DL models, it has not been possible to achieve high performance in controlled environments using in-the-lab datasets. We address human emotion recognition and classification by combining several in-the-wild datasets of facial images.

These problems pose different situations and require images with very different content, so we have designed a method of dataset refinement depending on the case of study.

In the first case, we implemented a DL model to classify Web pages into certain categories using only screenshots, where the results showed a very difficult multiclass problem. We addressed the same problem with the One vs. Rest strategy and improved the dataset by reclassification, irrelevant image detection, balancing and representativeness, as well as using regularisation techniques and a new prediction mechanism with binary classifiers. These classifiers operating separately improve performance, on average increasing validation accuracy by 26.29% and decreasing overfitting by 42.30%, showing significant improvements over the multiple classifier operating with all categories together. Using the new model, we have developed an online system for classifying Web pages that can help designers, Web site owners, Webmasters and general users.

In the second case, the strategy consists of progressively refining the facial image dataset through several successive trainings of a convolutional network model. In each training, the facial images corresponding to the correct predictions of the previous training are used, allowing the model to capture more distinctive features of each class of emotion. After the last training, the model performs an automatic reclassification of the entire dataset. This process also allows us to detect irrelevant images, but our aim is to improve the dataset without modifying, deleting or augmenting the images, unlike other similar work. Experimental results on three representative datasets demonstrated the effectiveness of the proposed method, improving the validation accuracy by 20.45%, 14.47% and 39.66%, for FER2013, NHFI and AffectNet, respectively. The recognition rates on the reclassified versions of these datasets are 86.71%, 70.44% and 89.17%, which reach the state of the art. We combined these better classified versions to increase the number of images and enrich the diversity of people, gestures and attributes of resolution, color, background, lighting and image format. The resulting dataset is used to train a more general model. Faced with the need for more realistic metrics of model generalisation, we created a combined, balanced, unbiased and well-labelled evaluator dataset. To this end, we organised this dataset into categories of gender, age and ethnicity. Using a predictor of these representative features of the population, we can select the

same number of images and by means of the successful Stable Diffusion model, it is possible to generate the facial images needed to balance the categories created from the aforementioned features.

Single-dataset and cross-dataset experiments indicate that the model trained on the combined dataset improves the generalisation of the individually trained models on FER2013, NHFI and AffectNet by 13.93%, 24.17% and 7.45%, respectively. We developed an online emotion recognition system that leverages the more generic model obtained from the combined dataset. Finally, the good quality of the synthetic facial images and the time reduction achieved with the generative method motivate us to create the first and largest artificial dataset of categorical emotions. This open-access product can complement real datasets, which are difficult to collect, label, balance, control features and protect people's identity.

Contenido

Lista de figuras

Lista de tablas

Nomenclatura

Acrónimos

AU Action Units

AUC Area Under the Curve

AVFER Aggregation for ViT on Facial Emotion Recognition

BOTW Best of the Web

CIRCL Computer Incident Response Center Luxembourg

HTTPS Hypertext Transfer Protocol Secure

CK+ Extended Cohn-Kanade Dataset

CLIP Contrastive Language Image Pretraining

CNN Convolutional Neural Network

CSS Cascading Style Sheets

CSV Comma Separated Values

CV Computer Vision

DL Deep Learning

ER Emotion Recognition

FACS Facial Action Coding System

FER Facial Expression Recognition

FIT Facial Image Threshing

GAN Generative Adversarial Network

HMI Human–Machine Interaction

HTML HyperText Markup Language

IA Inteligencia Artificial

iSPL Intelligent Signal Processing Lab

JPG Joint Photographic Experts Group

MAE Mean Absolute Error

ML Machine Learning

Multi-PIE Multi Pose, Illumination, Expressions

NHFI Natural Human Face Images

NLP Natural Language Processing

OvR One vs. Rest

PCA Principal Component Analysis

PNG Portable Network Graphics

ReLU Rectified Linear Unit

RGB Red, Green and Blue

ROC Receiver Operating Characteristic

SSL Secure Sockets Layer

SVM Support Vector Machine

URL Uniform Resource Locator

VAE Variational Autoencoder

VGG Visual Geometry Group

ViT Vision Transformer

Capítulo 1

Introducción

Este primer capítulo expone una visión general de nuestra investigación. Está compuesto de la Sección 1.1 que incluye los conceptos fundamentales del tema de interés, la necesidad que justifica esta tesis y la motivación detrás del trabajo desarrollado. La Sección 1.2 identifica el problema que se desea afrontar a través de dos casos de estudio concretos que se describen en la Sección 1.3. La Sección 1.4 explica brevemente el método de investigación y la estrategia metodológica diseñada que será detallada más adelante. La Sección 1.5 enuncia los objetivos que se pretende alcanzar. Finalmente, la Sección 1.6 proporciona la organización del contenido del presente manuscrito.

1.1 Antecedentes

Actualmente es la etapa de mayor éxito de la *Inteligencia Artificial* (IA), la cual ha sido impulsada por la tecnología computacional cada vez más potente, la creciente disponibilidad de datos, el auge de la investigación y el desarrollo de algoritmos "inteligentes", así como el trabajo colaborativo e interdisciplinar para abordar problemas en diferentes campos [82]. Las capacidades de las máquinas para realizar tareas similares a las humanas aumentan y mejoran continuamente. Esta revolución se experimenta hace aproximadamente una década y se debe principalmente al rápido desarrollo del *Aprendizaje Profundo* (Deep Learning, DL). A diferencia del *Aprendizaje Automático* (Machine Learning, ML) tradicional, en el que definir las características relevantes del problema y la representación adecuada de los datos para el funcionamiento de un algoritmo depende de un experto humano, en DL el cambio de paradigma es radical, ya que evita estas complejas tareas manuales.

La extracción y representación automática de características realizada por un algoritmo de DL, al mismo tiempo que aprende a resolver una tarea, ha tenido especial éxito en el campo de la visión por computadora. Las computadoras pueden "captar" el mundo como lo ven nuestros ojos mediante modelos que aprenden patrones y características en datos visuales. Una de las aplicaciones principales es reconocer objetos en imágenes y clasificarlos dentro de diferentes categorías. En muchas ocasiones, esta tarea se realiza de forma más ágil y precisa, ya que la identificación automática de cualquier característica observable que permita identificar un objeto o distinguirlo de otro podría pasar desapercibida al ojo humano. Numerosos sistemas de visión artificial basados en DL se han desarrollado para la robótica social, la conducción autónoma, la asistencia médica, el reconocimiento facial, o la seguridad pública y privada, entre otros usos.

Una solución de DL se compone de datos y un modelo, ambos elementos son fundamentales para lograr buenos resultados. A pesar de que los datos son el recurso principal para el proceso de aprendizaje, la investigación casi exclusivamente sigue un enfoque centrado en los modelos (*model-centric*, centrados en el modelo). Su propósito es encontrar nuevos algoritmos de aprendizaje que, al ser implementados con un lenguaje de programación, permitan obtener modelos de alto rendimiento con un dataset fijo [117]. En este contexto, aunque las redes *transformer* han causado una revolución en el *Aprendizaje Automático*, especialmente en el *Procesamiento del Lenguaje Natural* (Natural Language Processing, NLP), aún no han conseguido dominar del todo la *Visión por Computadora* (CV) y comparten protagonismo en diversas tareas desafiantes dentro de este campo. Cuando se trata de clasificación de imágenes, las herramientas de vanguardia siguen siendo las Redes Neuronales Convolucionales (Convolutional Neural Networks, CNNs). Están especializadas para realizar análisis automáticos de imágenes, aprender características específicas de cada clase, asociar imágenes y clases, y predecir la categoría a la que pertenecen nuevas imágenes [114]. Varias arquitecturas de CNN han sido propuestas, tanto personalizadas (creadas desde cero) como preentrenadas mediante *Aprendizaje por transferencia* (Transfer Learning) y *Ajuste fino* (Fine-Tuning). Todas estas técnicas prueban distintos hiperparámetros de configuración e incorporan mecanismos de regularización como el aumento de datos, dropout y la normalización por lotes [53]. En la práctica, el enfoque model-centric es demandante en términos de experimentación y tiempo de cómputo. Sin embargo, no ha alcanzado el objetivo de un rendimiento ideal.

A lo largo de este trabajo de tesis, hemos podido experimentar con muchas arquitecturas y configuraciones de modelos de CNN, incluso con redes de tipo transformer, pero la mejora del rendimiento no ha sido significativa. La hipótesis establece que si la calidad

de los datos es deficiente, el aumento de la precisión del reconocimiento es muy difícil. La investigación centrada en los datos (*data-centric*) está mucho menos explorada. Está guiada por el principio de que los datos son el recurso más importante y consiste en fijar el modelo mientras se busca un conjunto de datos de alta calidad [83]. En nuestro caso, los datos corresponden a imágenes estáticas o también extraídas de vídeos, cuya colección recibe el nombre de *dataset*. Muy pocos estudios se han propuesto mejorar los datasets de imágenes, a pesar de que los mismos creadores admiten los problemas en la calidad de los datos [73]. Esto se debe probablemente a que el manejo y tratamiento de un gran número de imágenes es un trabajo arduo y complicado en términos de tiempo y esfuerzo. Consecuentemente, la falta de resultados notables del enfoque model-centric, los escasos trabajos data-centric y la premisa de que los datos serían más importantes que el modelo, nos motivan a proponer un novedoso método data-centric. Utilizando técnicas de DL, efectuamos la ingeniería del dataset original para hacerlo más fiable, lo que mejora el rendimiento del reconocimiento y la clasificación de imágenes en aplicaciones del mundo real.

1.2 Problemática

La primera preocupación cuando se trabaja con técnicas de DL son los datos. Idealmente, un dataset de entrenamiento debe satisfacer los criterios de cantidad y calidad. Debe estar compuesto por un número suficiente de muestras y proporcionar una representación uniforme de las características, sin sesgos, sin errores, y disponer de etiquetas precisas en el caso supervisado [32]. Dichas propiedades permitirán que el proceso de aprendizaje alcance una precisión conveniente y posteriormente realice predicciones adecuadas.

En la realidad, estas condiciones no siempre se cumplen y, en algunos casos, se vuelven altamente complejas. Cuando se trata de datasets de imágenes relacionados con problemas del mundo visual, podemos encontrarnos con dos situaciones: 1) la inexistencia de un dataset apropiado para el tema de interés, por lo que debemos crearlo desde cero; y 2) la disponibilidad de uno o más datasets, pero que no satisfacen los mencionados criterios de cantidad y calidad.

Debido a que la recolección manual de extensos datasets de imágenes es difícil y consume tiempo, la recolección automática se ha convertido en la alternativa preferida. Sin embargo, este proceso involucra problemas que se reflejan en los datasets y afectan el rendimiento de los modelos de DL, que se entrenan y evalúan en estos datasets.

La Figura 1.1 representa la difícil problemática que abordamos mediante un árbol jerárquico [38]. La complejidad del problema principal puede descomponerse en subprob-

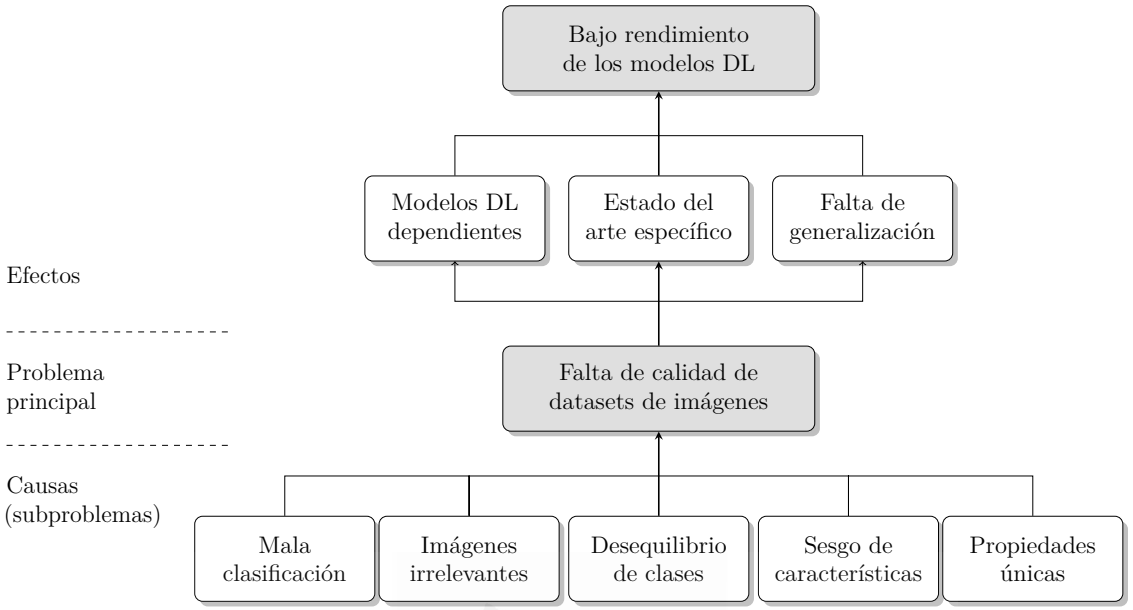


Figura 1.1: Representación gráfica del problema.

lemas más pequeños, identificados como causas, para los que podemos diseñar estrategias separadas, pero que pueden funcionar en conjunto para producir efectos positivos que contribuyan a la consecución del objetivo general.

Hemos mencionado que la práctica común es recolectar imágenes desde Internet de manera automática. El uso de scripts de programación o plugins de navegadores sin condiciones de control, verificación de errores ni estrategias de selección de muestras, así como la multitud de colaboradores para la tarea de etiquetado (*crowdsourcing*) son los factores principales que originan inconvenientes ampliamente citados en la literatura como:

- La mala clasificación, debida especialmente a la subjetividad del etiquetador humano y la existencia de clases similares o difíciles de diferenciar.
- La presencia de imágenes irrelevantes, cuyo contenido no es útil ni está relacionado con la tarea de reconocimiento concreta.
- Un desequilibrio en el número de imágenes de las distintas clases, unas mayoritarias y otras minoritarias, incluso con enormes diferencias.
- La falta de uniformidad o sesgo que puede conducir a una representatividad excesiva o insuficiente de ciertas características relevantes.

- Cada dataset tiene propiedades únicas como el número de imágenes, el color, la resolución y el formato de archivo, que pueden ser muy diferentes de otros datasets del mismo ámbito.

Un modelo de DL entrenado en un dataset que adolece de estas limitaciones aprende de ellas y se vuelve extremadamente dependiente. Esta dependencia termina sesgando el estado del arte en la tarea específica de reconocimiento y clasificación, pues el modelo con mayor precisión en un dataset, generalmente obtiene un rendimiento pobre cuando se prueba con datos nuevos y es peor en escenarios del mundo real donde las imágenes son captadas en entornos no controlados. Esto afecta la capacidad de generalizar y reconocer objetos raros o ausentes en el dataset de entrenamiento. Por lo tanto, la falta de calidad de los datasets de imágenes es el problema principal del bajo rendimiento de los modelos de DL. Si no se reducen los inconvenientes inherentes al dataset, es muy difícil mejorar el rendimiento en la tarea de reconocimiento y clasificación de imágenes en aplicaciones del mundo real.

1.3 Casos de estudio

A pesar de la revolución que ha supuesto el DL en problemas de visión artificial, igualando o incluso superando el rendimiento humano en ciertos casos, aún existen retos importantes. En este trabajo, consideramos dos aplicaciones que hasta el momento no alcanzan el rendimiento deseado en escenarios prácticos, motivo por el cual han despertado el interés científico en los últimos años, creando campos de investigación en sus respectivos ámbitos. En ambos casos, el uso de imágenes es el recurso esencial para aprovechar técnicas de DL que permitan tratar el problema de categorización de forma más rápida y confiable. Primeramente, abordamos la mejora del reconocimiento y la categorización de páginas Web mediante la creación de un dataset de capturas de pantalla. En segundo lugar, aprovechamos los datasets existentes para mejorar el reconocimiento y la categorización de emociones humanas.

1.3.1 Categorización de páginas Web

Es difícil imaginar un mundo sin Internet, específicamente sin la Web, que se ha convertido en la mayor plataforma de comunicación a nivel global y la principal fuente de información para casi todas las actividades humanas. Su contenido se estructura en páginas Web que crecen exponencialmente en número y son cada vez más sofisticadas. Esto supone un serio reto para los motores de búsqueda como Google y Bing, rastreadores, sistemas de

recomendación y directorios Web. Para organizar este complejo y vasto contenido, la técnica básica es la clasificación. Tradicionalmente, suele basarse en el análisis textual del código HTML. Sin embargo, el aspecto visual de una página Web desempeña un papel importante y puede ser una nueva alternativa para distinguir entre categorías. En lugar de un análisis del código y contenido cada vez más complejos, usando la imagen de la página Web tal como aparece ante el usuario la evaluación será independiente del lenguaje y la tecnología de implementación. La categorización automática es el método viable para hacer frente al problema de escalabilidad y depuración de la Web, pues hay muchos sitios inactivos o de baja calidad. Incluso, este método basado en la imagen podría combinarse con el convencional basado en contenido para obtener mejores resultados.

1.3.2 Reconocimiento de emociones humanas

Aunque la IA ha logrado avances impresionantes imitando las capacidades cognitivas y físicas del ser humano, el aspecto emocional es aún distintivo de las personas. Automatizar la percepción de las emociones humanas puede beneficiar a muchas áreas. Sin embargo, esta tarea es difícil para las máquinas, incluso para los humanos, ya que su naturaleza subjetiva hace que las personas interpreten emociones diferentes a partir de la misma información. La expresión del rostro es el indicador que mejor define el estado emocional de la persona, por lo que analizar una imagen facial es el enfoque más extendido para tratar este problema. Sin embargo, las complejas características de cada tipo de emoción y las sutiles distinciones entre clases similares de emociones, hace que sea difícil para los modelos de DL generalizar a partir de estas imágenes y lograr un alto rendimiento en aplicaciones prácticas.

1.3.3 Criterios de selección

La selección de ambos casos de estudio, tanto la categorización de páginas Web como el reconocimiento de emociones, se debe a los siguientes criterios:

- Cada caso plantea una situación diferente con respecto a la existencia de un dataset de imágenes adecuado. La disponibilidad de una gran cantidad de datos es el requisito actual de la investigación y el desarrollo en áreas relacionadas con la Web. En respuesta a esta necesidad, proponemos un nuevo dataset que incluye una representación visual, así como atributos cualitativos y cuantitativos para una mejor caracterización de las páginas Web. Estos elementos también nos permiten extraer información útil acerca del diseño Web a partir del análisis estadístico presentado

en el Apéndice A. En contraste, tenemos una amplia gama de datasets de emociones disponibles, pero no han sido efectivos trabajando de manera individual, así que hemos combinado algunos de los más reconocidos. Analizar ambas situaciones nos permite diseñar una propuesta más general.

- El contenido de la imagen es un aspecto importante en el proceso de mejora de los datasets. En relación a las emociones, las imágenes contienen rostros humanos tomadas de fotografías o vídeos, mientras que las imágenes de páginas Web difieren de fotografías que incluyen personas, animales, plantas u objetos, donde sus características son mejor definidas, sino que son capturas de pantalla de sitios Web, cuyos elementos son más difíciles de detectar pues se mezclan contenidos de tipo textual, gráfico y multimedia.

- La gran relevancia que tienen para sus respectivas áreas. Por una parte, la categorización de páginas Web basada en capturas de pantalla es una potencial alternativa para optimizar la organización y depuración del enorme y complejo contenido Web. Por otro lado, el reconocimiento de emociones humanas basada en imágenes faciales tiene implicación directa en ámbitos como la robótica, medicina, seguridad, marketing, educación, empleo, entre otros. Un detalle más amplio de estas aplicaciones se encuentra en el Apéndice B.

A pesar de ser dos temáticas muy diferentes, cuando son tratadas con técnicas de DL, el problema tiene una raíz común: los datasets de entrenamiento adolecen de desventajas, lo que hace difícil a los modelos aprender de manera efectiva a partir de estos datos.

1.4 Propuesta

La estrategia para mejorar la calidad del dataset de entrenamiento de un modelo de DL para el reconocimiento y la clasificación de imágenes sigue los pasos ilustrados en la Figura 1.2. Básicamente, hemos diseñado mecanismos de ingeniería de datos apoyados en diversas técnicas de DL para abordar los subproblemas identificados en el apartado 1.2.

Una vez que definimos la problemática que será tratada con DL, comenzamos con la búsqueda de los datasets in-the-wild existentes en el dominio del problema. Además de los datasets que se citan en los trabajos similares, algunas de las fuentes más conocidas para la obtención de datasets de imágenes son el motor de búsqueda de *Google Dataset*

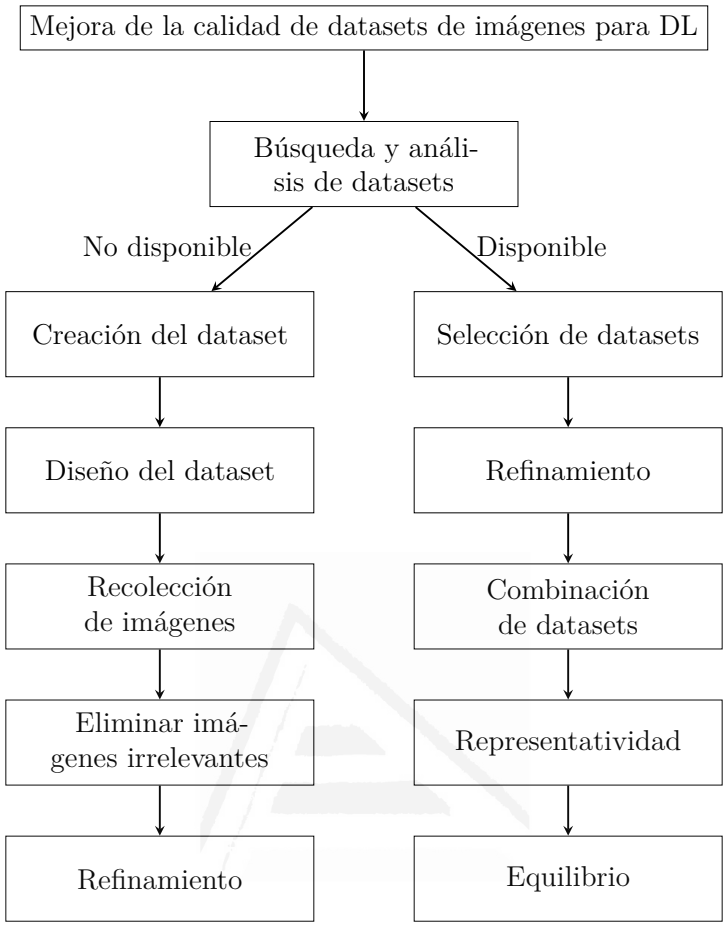


Figura 1.2: Flujograma para mejorar la calidad de los datasets de imágenes.

Search[1], los sitios Web de *Open Science Foundation*[2] (OSF) y *Papers with Code*[3], el repositorio de imágenes de *ImageNet*[4], y plataformas como *Kaggle* y *GitHub*.

Como resultado de esta búsqueda, podríamos encontrarnos con que no hay ningún dataset disponible o, si lo hay, no se adapta totalmente a las necesidades del problema. A continuación, abordamos cada una de estas posibilidades.

1.4.1 Dataset no disponible

Acerca de páginas Web, los datasets existentes son escasos y están limitados a capturas de pantalla que muchas veces aparecen recortadas, lo que elimina información visual útil para la categorización. Además, no se conoce el origen ni la denominación de la

[1] https://datasetsearch.research.google.com/
[2] https://osf.io/
[3] https://paperswithcode.com/datasets
[4] https://www.image-net.org/

página Web. En raras ocasiones se incluyen las URLs respectivas. Por tales motivos, no es posible aprovechar estos datasets, así que nuestro propósito es crear uno adecuado desde cero.

Primero, es necesario diseñar la estructura del dataset y definir sus componentes. No sólo incorporamos la representación visual de la página Web a través de una captura de pantalla completa, sino que decidimos combinarla con parámetros cualitativos y cuantitativos extraídos del código fuente HTML subyacente, de forma que una página Web esté mejor caracterizada. Este dataset mixto que integra datos de tipo visual, textual y numérico es una contribución a la investigación sobre problemas relacionados con la Web, la cual demanda datasets más sofisticados y de mayor tamaño.

Dado que la recolección manual de datos es una tarea compleja y requiere demasiado tiempo, hemos automatizado la mayor parte del proceso escribiendo varios programas en Python y R. Primero, recolectamos automáticamente las URLs desde múltiples fuentes a nivel mundial a través de las técnicas de búsqueda (Google) y navegación (directorio Web). Descargamos y guardamos el código HTML, y extraemos los atributos cuantitativos y cualitativos mediante *scraping*, junto con el aspecto visual de las páginas Web mediante *webshooting*.

Aunque hemos generado un nuevo y extenso dataset para la Web, la descarga automática de imágenes trae consigo imágenes irrelevantes que no son útiles. Es aquí donde empieza la mejora del dataset utilizando técnicas de DL. Procedemos a depurar el dataset por medio de un clasificador de páginas Web válidas y de error. Estas últimas son eliminadas del dataset inicial, disminuyendo el tamaño del mismo, pero aumentando su confiabilidad.

Otro problema frecuente es la clasificación errónea. Las imágenes dentro de una misma clase (intraclase) tienen una variabilidad importante entre sí y una similitud entre las imágenes de distintas clases (interclase). Las páginas Web son cada vez más polifacéticas, ya que muchas son parte de sitios, portales, repositorios y plataformas que engloban contenidos diversos, por lo que puede haber confusión entre categorías, por ejemplo, páginas Web científicas que incluyen hallazgos en forma de un sitio de noticias o incluso publicidad. En particular, los resultados del buscador de Google están más orientados al contenido que a la apariencia, por lo que varias capturas de pantalla de páginas Web no tienen la apariencia esperada de la categoría. Para reducir este problema, diseñamos un proceso de refinamiento basado en los criterios de representatividad y equilibrio. Vamos a conformar un dataset más pequeño, pero equilibrado y de menor variabilidad intraclase y similitud interclase. Dado que las capturas de pantalla procedentes del directorio Web son más fiables debido a la estricta revisión a la que están sometidas, de este conjunto

seleccionamos manualmente el mismo número de imágenes, las cuales deben ser las más representativas de cada categoría. Este dataset sirve para entrenar un modelo de DL que captura las características más distintivas de las diferentes categorías y que permite clasificar el resto de imágenes. Así, obtenemos una versión mejor clasificada del dataset original.

1.4.2 Dataset disponible

La búsqueda de datasets de imágenes para el problema de interés puede dar como resultado una o más opciones disponibles. Existen varios datasets faciales de emociones. Sin embargo, es muy difícil obtener la misma precisión de entornos controlados (in-the-lab) en entornos no controlados (in-the-wild). Actualmente, no hay un dataset que represente la compleja heterogeneidad de las expresiones emocionales de toda la población de forma completa y equilibrada. Este sería el recurso ideal para entrenar un modelo de reconocimiento en aplicaciones reales, pero crear un dataset con estos requisitos es un problema de enormes proporciones. Por tal razón, proponemos la combinación de varios datasets conocidos del ámbito de emociones.

Empezamos seleccionando los datasets más convenientes para la combinación con base en la disponibilidad y distintas propiedades de tamaño, resolución, color y número de categorías para ampliar la diversidad del dataset resultante. En esta investigación, consideramos los datasets FER2013 [33], NHFI [105] y AffectNet [81]. Todos ellos presentan desequilibrio, imágenes irrelevantes y mal clasificadas. Estos problemas son tratados de manera individual.

La estrategia consiste en un refinamiento progresivo del dataset ejecutando varios entrenamientos del mismo modelo de red convolucional. Después de cada entrenamiento, se realiza la predicción de todas las imágenes faciales y sólo se seleccionan las correctas para formar el dataset del siguiente entrenamiento. Este proceso se repite hasta que hay pocas predicciones incorrectas, normalmente de un solo dígito. Como resultado, el último modelo entrenado alcanza una precisión muy alta, por lo que se encarga de reetiquetar todas las imágenes del dataset original. Por tanto, se genera una nueva distribución del dataset sin alterar su tamaño ni modificar las imágenes. En el último paso, se entrena el mismo modelo CNN en la versión reclasificada del dataset, y la precisión resulta mayor en comparación con el dataset original. Los experimentos muestran un aumento del 20.45%, 14.47% y 39.66% para FER2013, NHFI y AffectNet, respectivamente. En estos datasets también se alcanzó un rendimiento del estado del arte.

Cabe señalar que este proceso de refinamiento iterativo también nos permite identificar las imágenes irrelevantes. Sin embargo, nuestro trabajo se ha impuesto no eliminar

imágenes para evitar modificar el tamaño del dataset original. Utilizamos las versiones reclasificadas de estos datasets para combinarlas de forma que actúen como uno solo. El objetivo es introducir más variabilidad de rostros y gestos faciales, así como también mezclar distintas propiedades de la imagen como la dimensión, la calidad, el color, el fondo y la iluminación. Cabe señalar que juntar estas características plantea dificultades, sobre todo a la hora de entrenar y evaluar los modelos. El resultado es un único dataset más extenso y diverso, el mayor de emociones categóricas que conocemos de carácter mixto. Un modelo de DL entrenado en este nuevo dataset puede convertirse en el primero y más genérico hasta el momento.

No sólo nos preocupamos por disponer de un gran dataset más diverso para entrenar modelos de reconocimiento más robustos, sino que también debemos probar lo bueno y genérico que es un modelo entrenado en este y contrastarlo con otros datasets. Para lograrlo, es necesario un dataset evaluador que sea representativo capaz de proporcionar una métrica de rendimiento más objetiva y genérica para las aplicaciones prácticas. Tal dataset no existe en la actualidad, por lo que nos proponemos crearlo siguiendo una estructura equitativa definiendo categorías de las variables más relevantes de la población y que se pueden extraer de la imagen facial como el género, la edad y la etnia de las personas retratadas [92]. Para cada categoría, elegimos el mismo número de imágenes utilizando un modelo predictor de estas características para asignar las imágenes a las carpetas correspondientes. El tamaño de este dataset evaluador no es considerable, sin embargo, es mucho más importante la representatividad que la cantidad. Dado que faltarán imágenes para ciertas categorías que resultan raras, las carpetas incompletas se equilibran con imágenes sintéticas generadas por *Stable Diffusion*, un modelo de difusión generativa de IA de reciente éxito [36][95]. La alta calidad de las imágenes sintéticas avala que este enfoque es una alternativa conveniente y eficaz a las técnicas tradicionales como el aumento de datos y GAN [5][52]. Esto nos motiva a crear el primer y mayor dataset totalmente artificial en el campo de emociones categóricas.

En esta sección, presentamos la propuesta para mejorar la calidad de los datasets de imágenes considerando su creación desde cero y la combinación de algunos conocidos. En cada caso, diseñamos y aplicamos varias técnicas basadas en DL para reducir los inconvenientes de los datasets originales y hacerlos más apropiados para la tarea de interés. Para validar nuestra propuesta, emplearemos estos datasets mejorados para entrenar y evaluar modelos basados en redes convolucionales, tanto personalizados como preentrenados. Esta serie de experimentos nos permite comprobar un mejor rendimiento y generalización en dos problemas de gran complejidad: la categorización de páginas Web y el reconocimiento de emociones humanas.

1.5 Objetivos

1.5.1 General

Proponer un novedoso método data-centric basado en técnicas de Deep Learning para mejorar el reconocimiento y la clasificación de imágenes en aplicaciones del mundo real.

1.5.2 Específicos

1. Diseñar estrategias basadas en técnicas de DL para mejorar los datasets de imágenes con el fin de aumentar el rendimiento y la generalización de los modelos de reconocimiento y clasificación de imágenes.

2. Implementar modelos de DL, basados en redes neuronales convolucionales, que alcancen un mejor rendimiento y generalización en la categorización de páginas Web y reconocimiento de emociones humanas.

3. Emplear los datasets y los modelos de deep learning mejorados en el desarrollo de sistemas de categorización de páginas Web y reconocimiento de emociones humanas.

1.6 Estructura de la tesis

El contenido del documento se organiza de la siguiente manera: en el Capítulo 1, hemos presentado el contexto general de nuestro trabajo de investigación. Identificamos la problemática de la falta de calidad de los datos y su importancia en el DL. Resaltamos la necesidad de métodos que busquen mejorar dicha calidad pues influye altamente en el buen rendimiento de los modelos. Describimos la metodología y los procesos diseñados para reducir los problemas más conocidos de los datasets de imágenes en la categorización de páginas Web y el reconocimiento de emociones. Por último, planteamos lo que de manera general y específica pretendemos conseguir.

El Capítulo 2 revisa el estado actual del tema que se está abordando. Analizamos los esfuerzos de investigación y trabajos relacionados con el enfoque data-centric para mejorar la calidad de los datasets de imágenes, especialmente en el ámbito de emociones, la combinación de datasets y la generación de imágenes faciales de manera artificial.

El Capítulo 3 está dedicado al difícil problema de la categorización de páginas Web basada en capturas de pantalla. Se estudian los actuales datasets de imágenes en este dominio y se justifica la necesidad de crear uno mejor. Se detalla la metodología de trabajo, así como el proceso de refinamiento del nuevo dataset mediante técnicas de DL.

Se expone la categorización Web temática multiclase, un problema que demuestra alta complejidad y que es tratado mediante un enfoque de clasificación binaria para aumentar el nivel de precisión y reducir el sobreajuste.

El Capítulo 4 expone el tratamiento del problema de reconocimiento de emociones humanas basada en imágenes faciales. Presenta un análisis de los datasets de emociones considerados en esta investigación. Describe en detalle la metodología de trabajo, los métodos y materiales utilizados. Explica la parte experimental, muestra y discute los resultados obtenidos.

El Capítulo 5 se refiere a la implementación de dos aplicaciones del mundo real, en las cuales ponemos en práctica los datasets y modelos obtenidos en los capítulos anteriores.

Finalmente, en el Capítulo 6 se mencionan las conclusiones generales de nuestro trabajo, las contribuciones concretas de esta tesis, los repositorios de los recursos utilizados y productos generados, las publicaciones científicas realizadas, y las posibles líneas de trabajo futuro.

Capítulo 2

Estado del arte

En este capítulo, exponemos todo lo relacionado con la investigación data-centric en el ámbito de ML y los datasets de imágenes. La Sección 2.1 se refiere a los conceptos fundamentales y la necesidad de realizar aportes bajo este enfoque ante la supremacía de la investigación model-centric. La Sección 2.2 analiza los retos que se presentan en los datasets de imágenes y explora los métodos propuestos en los trabajos de la literatura data-centric. Al final, en la Sección 2.3 destacamos las diferencias e innovaciones con respecto a nuestro trabajo.

2.1 Enfoque data-centric de ML

En los últimos años, el cambio de paradigma de la programación tradicional hacia el aprendizaje automático ha sido la clave para que la IA empiece a salir de la ciencia ficción y pase a formar parte de nuestro día a día. La rápida evolución en cuanto a la disponibilidad de los datos, el desarrollo de algoritmos que aprenden de estos datos y el creciente poder computacional, han sido los factores de éxito para los impresionantes logros del ML.

Existe una amplia gama de aplicaciones basadas en ML en cualquier ámbito de la vida, por ejemplo, los asistentes virtuales, traductores de idiomas, autocompletado y respuestas automáticas de mensajería, sistemas de recomendación y sugerencias personalizadas de noticias, vídeos, productos y servicios, reconocimiento de voz y rostros, aplicaciones de Facebook, Instagram y Twitter, entretenimiento, juegos, etc. Sin embargo, esta masiva puesta en práctica también ha desvelado ciertas aplicaciones que funcionan bien en un entorno controlado de laboratorio, pero suelen fallar en escenarios reales.

Para abordar esta problemática, la tendencia que ha seguido la investigación es de tipo model-centric, que busca la mejora de la precisión fijando los datos y mejorando iterativamente el algoritmo de entrenamiento o la arquitectura del modelo [43][67]. Sin embargo, no cabe esperar avances significativos cuando los datos utilizados no son fiables.

Los modelos de DL se caracterizan por una fuerte dependencia, es decir, son útiles cuando los datos son muy parecidos a los de entrenamiento, pero pueden fallar en entornos distintos. En particular, en el mundo real se enfrentan a imperfecciones de los datos. Los modelos aprenden estas limitaciones, por lo que trabajar sólo bajo el enfoque model-centric puede ser el camino menos conveniente.

En contraste, la mejora de la calidad del conjunto de datos a menudo proporciona una mayor precisión que pasar tiempo experimentando con arquitecturas más profundas, ajuste de hiperparámetros y técnicas de regularización de los modelos [40]. A pesar de esta realidad conocida, los esfuerzos de investigación data-centric son escasos. Una razón es que disponer de un conjunto de datos de alta calidad a gran escala tiene un elevado coste y consume demasiado tiempo [66]. Este problema se incrementa en el ámbito del DL debido a la gran dependencia de enormes volúmenes de datos de entrenamiento, especialmente imágenes y anotaciones, lo que añade otra dificultad que es el considerable espacio de almacenamiento. Por lo tanto, las contribuciones en esta dirección son fundamentales para lograr que el buen rendimiento en laboratorio se traslade al mundo real. Precisamente, nuestra intención es enfrentar los retos asociados a los datos de imágenes en escenarios reales.

2.2 Retos del enfoque data-centric

Se han identificado problemas genéricos comunes a todos los datasets in-the-wild creados para entrenar modelos de DL, tales como la escasez de datos, datos incompletos, la falta de representatividad, sesgo de características, el desequilibrio de clases y las etiquetas erróneas. Seguidamente, profundizamos en los métodos propuestos por la investigación data-centric para enfrentar estas problemáticas y los contrastamos con las estrategias diseñadas en nuestro trabajo de tesis.

2.2.1 Aumento de datos

Hay aplicaciones en las que los datasets de imágenes que constituyen la materia prima se han quedado pequeños, por ejemplo, el diagnóstico médico de enfermedades raras o el control de calidad en la industria [98]. Los trabajos realizados para atenuar la necesidad

de grandes cantidades de datos se enmarcan principalmente dentro de dos líneas de investigación: el aumento de datos y la generación artificial o sintética de datos . Una tercera alternativa que consideramos útil en esta tesis es la combinación de los datasets pertenecientes al dominio del problema. Seguidamente, analizamos en detalle cada una de ellas.

El clásico *data augmentation* es un método eficaz para incrementar los datos de entrenamiento y mejorar el rendimiento de la generalización. Para datos de imágenes, consiste en aplicar transformaciones geométricas como rotación, giros horizontales y verticales, traslación, reflexión, escalado y recorte aleatorio [42]. Aunque estos efectos se pueden conseguir de forma personalizada utilizando librerías gráficas, se prefieren herramientas automáticas como las proporcionadas por TensorFlow y Keras. Aquí no hay necesidad de entrenar una red neuronal profunda en un dataset de imágenes. Sin embargo, las imágenes generadas terminan siendo sólo copias ligeramente modificadas de las imágenes originales.

En años recientes, se han propuesto técnicas más sofisticadas de aumento de datos que producen una variedad de imágenes a partir de las originales. Por ejemplo, la oclusión aleatoria eliminando regiones al azar de la imagen con *CutOut* [18], la sustitución de una parte de la imagen por otra diferente mediante *CutMix* [113], *MixUp* [115] produce una mezcla de dos imágenes en forma de collage artístico, y *AugMix* [45] que realiza una composición de varias operaciones de aumento para producir una nueva imagen sin alejarse demasiado de la original. Finalmente, *GridMix* [4] es un método basado en cuadrículas, que permite una mezcla discontinua de dos imágenes. Las imágenes son divididas en $n \times n$ celdas y, en cada celda de la cuadrícula, se selecciona aleatoriamente una porción de cada imagen para llenar esa celda.

Estos métodos son especialmente eficaces para conjuntos de imágenes con una categorización bien diferenciada o donde las clases pertenecen a temáticas muy diferentes; sin embargo, en el caso de las capturas de pantalla de las páginas Web o las imágenes faciales, se podrían eliminar o mezclar características distorsionando el contenido y la semántica. Por ende, nos enfocamos en un método más conveniente para el tipo de imágenes de nuestro estudio.

2.2.2 Generación de imágenes artificiales

El estado del arte en la generación de imágenes artificiales es avanzado y está en constante desarrollo. Existen varios métodos que permiten generar imágenes faciales con un alto grado de realismo y control sobre las características de las imágenes generadas. Pueden dividirse en dos grandes grupos:

- Los métodos geométricos basados en imágenes.
- Los que utilizan técnicas de síntesis de imágenes.

2.2.2.1 Métodos geométricos

Además del aumento de datos convencional, también existen técnicas de modelado 2D y 3D que requieren imágenes de referencia e información geométrica asociada. Estas técnicas permiten un control preciso de la posición, la expresión y la iluminación de las imágenes generadas.

Jin et al. [49] utilizan FACSGen, un programa facial 3D basado en Facial Action Coding System (FACS), para generar un dataset sintético de expresiones faciales que consta de 1000 sujetos y cada sujeto tiene 7 expresiones. Este dataset se utiliza como datos no etiquetados y se combina con los datos etiquetados como entrada para una red neuronal profunda. El entrenamiento se realiza bajo un mecanismo de aprendizaje por asociación, que recompensa o penaliza una asociación en función de la similitud entre la característica etiquetada y la no etiquetada. De este modo, es posible mejorar la capacidad de discriminación de la red profunda en el mismo dataset. Los resultados experimentales con los datasets RaFD y Oulu-CASIA muestran que el rendimiento del reconocimiento de expresiones faciales mejora con el aprendizaje por asociación.

Vonikakis et al. [107] abordan la escasez de datasets dimensionales sobre emociones. Utilizan la técnica del *morphing* entre imágenes faciales de expresiones categóricas para generar imágenes sintéticas que puedan asignarse al espacio polar AV (valencia e intensidad), con control total de la distribución y etiquetadas dimensionalmente de forma automática y coherente. Se presenta el dataset MorphSet con 167 sujetos y aproximadamente 342 expresiones por sujeto, lo que da un total de 57K+ imágenes. Se entrena un modelo ResNet-18 con MorphSet, AffectNet y Aff-Wild para predecir la valencia y la expresión, y se evalúa con el conjunto de validación de AffectNet y un 20% de imágenes no vistas de Aff-Wild y MorphSet. Las métricas de rendimiento favorecen a Morphset, lo que sugiere que es adecuado para el aprendizaje supervisado de emociones de tipo continuo.

Kollias et al. [62] utilizan una imagen facial con una expresión neutra y sintetizan una nueva imagen facial de la misma persona, pero con una expresión categórica o dimensional diferente. El proceso realiza la detección facial y la localización de puntos de referencia en la imagen de entrada, ajusta un modelo 3D deformable en la imagen resultante, deforma el rostro reconstruido, añade la expresión deseada y mezcla el nuevo rostro con la imagen original. Las imágenes faciales artificiales se utilizan para aumentar los datos y entrenar

redes neuronales profundas en varios datasets dimensionales o categóricos, verificando la mejora del rendimiento en el reconocimiento de emociones.

2.2.2.2 Síntesis de imágenes

Es otro método muy eficaz para generar artificialmente imágenes que contengan una expresión facial fotorrealista deseada. Una de las técnicas más utilizadas en la síntesis de imágenes es el uso de GAN (Generative Adversarial Networks). Se trata de una arquitectura que combina una red generativa y una red discriminativa y que puede ser entrenada sobre datasets reales para generar imágenes que formen datasets sintéticos [63]. Dado que la GAN original no podía generar imágenes faciales con una expresión facial específica referida a una persona concreta, se han propuesto algunos métodos condicionados a categorías de emoción.

Se han creado datasets faciales sintéticos a partir de redes GAN entrenadas en datasets reales, por ejemplo, Colbois et al. [16] utilizan StyleGAN2 como generador de un dataset sintético. Mediante la edición semántica del espacio latente, se generan imágenes faciales con variaciones controladas de expresión, pose e iluminación. El dataset sintético es Syn-Multi-PIE e imita a Multi-PIE[1] (Multi Pose, Illumination, Expressions). Los resultados experimentales indican que el dataset sintético satisface los requisitos de privacidad y precisión. Se confirma que las identidades generadas son novedosas, es decir, diferentes de las del dataset de entrenamiento, y la evaluación sugiere que se puede sustituir el dataset real por el sintético y seguir obteniendo conclusiones similares sobre el rendimiento de diversos sistemas de reconocimiento facial.

Boutros et al. [10] crearon un dataset de rostros sintéticos respetuoso con la privacidad llamado SFace. Utiliza un modelo generativo StyleGAN2-ADA que se entrena en el dataset CASIA-WebFace, el cual se utiliza para tareas de reconocimiento facial. El entrenamiento se realiza con un enfoque condicional, en el que las etiquetas de clase son las etiquetas de identidad. La evaluación sugiere que SFace se puede utilizar para entrenar modelos de reconocimiento facial, logrando un alto rendimiento de verificación, y se demuestra que asociar una identidad del dataset auténtico a otra con la misma etiqueta de clase del dataset sintético es difícilmente posible.

Bozorgtabar et al. [11] proponen un método para sintetizar imágenes faciales fotorrealistas condicionadas por la expresión facial. Un codificador-decodificador utiliza la representación latente compartida entre dominios de imagen y un mapa de calor de puntos de referencia faciales como representación de la expresión facial. Aumentar el dataset Oulu-CASIA VIS con imágenes de expresión sintéticas y entrenar un clasificador

[1] https://www.cs.cmu.edu/afs/cs/project/PIE/MultiPie/Multi-Pie/Home.html

de expresiones que alcance una precisión media superior a la del estado del arte. A partir de evaluaciones cualitativas, cuantitativas, y a nivel de usuario, demuestran que las imágenes sintéticas son de mayor calidad en comparación con IcGAN y CycleGAN, que pueden utilizarse para aumentar los datos y que el aprendizaje a partir de imágenes faciales sintéticas puede mejorar el reconocimiento de expresiones faciales (FER).

La síntesis de imágenes mediante GANs es la solución más notable para generar imágenes artificiales y sus diversas variantes han logrado resultados del estado del arte [11][49]. Los estudios citados avalan la utilidad de las imágenes sintéticas como recurso complementario para mejorar el rendimiento de los modelos de visión computacional. De ahí que los datasets faciales artificiales surjan como una alternativa muy prometedora a los datasets de imágenes faciales reales [84], cuya recopilación, preparación y etiquetado demanda mucho tiempo y trabajo, y son propensas a errores. La distribución de clases es desequilibrada, los atributos no son controlables y la información facial requiere el consentimiento de la persona debido a problemas de privacidad y uso indebido [10][29].

En lugar de recopilar, se trata de generar las imágenes por computadora reduciendo tiempo y esfuerzo. Es una técnica escalable según las necesidades, el etiquetado es automático y más fiable, es un proceso controlable bajo los parámetros especificados por el usuario, minimiza los problemas de privacidad al no tratarse de personas reales en situaciones reales y permite equilibrar casos raros. Esto último es muy útil para el caso de etnias, edades o género menos frecuentes. En resumen, los datasets faciales reales tienden a ser más ruidosos y menos controlados que los artificiales, lo que resulta potencialmente beneficioso para mejorar el rendimiento en el reconocimiento de emociones.

Sin embargo, sigue habiendo retos como generar imágenes que sean verdaderamente realistas y representen la complejidad y diversidad de las expresiones faciales en el mundo real. Por lo tanto, las imágenes sintéticas no pueden sustituir totalmente a las imágenes reales, ya que éstas no representan el mundo real. Como conclusión, aún es necesario entrenar y evaluar los modelos con imágenes reales para garantizar su capacidad de generalización a escenarios prácticos [49].

Hemos aprovechado esta cualidad de complementariedad de las imágenes sintéticas para equilibrar nuestro dataset evaluador, que es un conjunto más pequeño y no requiere una cantidad significativa de imágenes artificiales. En cambio, la necesidad de un gran dataset de entrenamiento es tratada mediante la combinación de los datasets existentes, en lugar del intenso trabajo que implica la recopilación y el etiquetado desde cero de decenas de miles de imágenes.

2.2.3 Combinación de datasets de FER

El enfoque tradicional de FER es con un único dataset (*single-dataset*), que consiste en entrenar y evaluar en el mismo dataset, reservando una parte para los datos de entrenamiento y otra para los de evaluación. Esto limita la capacidad de generalización de dicho modelo y su rendimiento en entornos reales. Nos interesa el trabajo *cross-dataset*, que trata de ampliar la variedad combinando datasets, ya sea para el entrenamiento, la evaluación, o ambos. Pocos trabajos utilizan e investigan este enfoque en profundidad.

Ramis et al. [92] proponen un protocolo para unir múltiples datasets utilizando 4 conocidos (BU-4DFE, CK+, JAFFE, WSEFEP), 2 nuevos (FEGA y FE-Test) y descartan FER+ y AffectNet debido a los problemas de los datasets in-the-wild. La precisión alcanza el 70% cuando el modelo basado en una CNN se prueba con FE-test, no visto durante el entrenamiento. Se concluye que la combinación mejora los resultados respecto a un único dataset de entrenamiento, ya que cada dataset añade información importante en el entrenamiento al haber sido recogidos en diferentes condiciones y amplía la diversidad de género, edad y etnia con diferentes fondos e iluminación.

Meng et al. [79] utilizan los datasets ExpW y Sfew in-the-field, FERPlus y Raf-db in-the-wild, y CK+ y Jaffe in-the-lab. Realizan varios experimentos single- y cross-dataset (diferentes datasets de origen y destino). Para conseguir un modelo más general, combinan los datasets de forma sistemática y sucesiva variando de 2 a 6. Estos datasets combinados sirven como entrenamiento (fuente múltiple) y sólo uno para la prueba (objetivo único). Aunque los resultados muestran que el entrenamiento con datasets combinados puede mejorar la generalización de un modelo, se concluye que no necesariamente la mayor cantidad de datos puede mejorar la precisión, sino que la necesidad de combinarlos se debe a que la capacidad de generalización de los datasets pequeños debe ser baja.

Chaudhari et al. [14] combinan FER2013, CK+ y AffectNet en uno denominado AVFER (Aggregation for ViT on Facial Emotion Recognition), que se divide en entrenamiento, validación, y prueba para evaluar Vision Transformer (ViT) y contrastarlo con ResNet-18. La precisión alcanza el 50.05% y el 53.10% para ResNet-18 y ViT, respectivamente. Se trata de un resultado aceptable teniendo en cuenta que es uno de los primeros enfoques que utilizan Transformer para el reconocimiento de emociones faciales.

Ghosh et al. [30] utilizan FER2013 y CK+ (omitiendo las categorías de neutralidad y desprecio) para desarrollar un modelo de aprendizaje federado compartido. Se consideran dos dispositivos cliente y un servidor central, todos ellos con la red MobileNet. Durante el proceso, las imágenes permanecen fuera del servidor, lo que garantiza la privacidad frente a posibles actividades maliciosas. Un cliente entrena su modelo con FER2013 y el otro con CK+, y los parámetros (pesos) se transfieren al servidor, donde se calculan los

promedios y se asignan como pesos al modelo global. La evaluación single- y cross-dataset muestra mejores indicadores de precisión (0.7657) y recall (0.7450) de MobileNet-federado, en comparación con los modelos entrenados únicamente con FER2013 o CK+, lo que lo convierte en una forma más robusta y segura de entrenar con imágenes faciales.

Abou et al. [2] crean el dataset 3RL que contiene 24000 imágenes faciales etiquetadas según 5 tipos de emoción, a partir de la combinación de FER2013, CK+ y un dataset generado por estudiantes[2] de ML. Los experimentos single- y cross-dataset aplicando SVM y CNN a los datasets individuales y combinados, indican una mejor generalización utilizando el método de DL.

Kim et al. [58] combinan FER2013, CK+ e iSPL[3] (intelligent Signal Processing Lab), normalizando automáticamente las imágenes faciales mediante la máquina FIT, una red neuronal en cascada multitarea y un programa de redimensionamiento. Los resultados indican que, tras fusionar los datasets, el rendimiento del modelo basado en Xception aumenta la precisión de validación en un 15.33%.

La combinación de datasets y la generación de imágenes sintéticas son las estrategias que nos han servido para abordar los problemas de escasa cantidad de datos, datos incompletos, datos que no están presentes y desequilibrio. Añadimos una organización del dataset de imágenes faciales basada en el género, la edad y la etnia, que son características relevantes de la población para reducir el sesgo de características y la falta de representatividad.

2.2.4 Mejora de imágenes, depuración y reetiquetado

Nuestra búsqueda bibliográfica sobre la mejora del reconocimiento de emociones basada en imágenes faciales también incluye técnicas de preprocesamiento de imágenes, eliminación de ruido, eliminación de imágenes con errores y reetiquetado. Por ejemplo, Liu et al. [69] analizaron el reconocimiento de expresiones teniendo en cuenta la importancia del preprocesamiento de datos mediante la mejora del contraste de la imagen. Se obtienen características faciales más discriminativas utilizando un método híbrido para la extracción y una red de clasificación que combina VGG-16 y ResNet. En los experimentos en tres datasets de referencia: CK+, FER2013 y AR, se obtuvieron los mejores índices de reconocimiento: 98.6%, 94.5% y 97.2%, respectivamente.

Kim et al. [59] diseñaron un sistema de preprocesamiento de imágenes y vídeos denominado máquina FIT (Facial Image Threshing) capaz de eliminar imágenes faciales irrelevantes, recortar, redimensionar, y reorganizar la clasificación de imágenes faciales

[2] https://github.com/muxspace/facial_expressions
[3] https://ispl.korea.ac.kr

antes de entrenar el algoritmo Xception, mejorando la precisión de validación en un 16.95% con FER2013.

Mazen et al. [73] aplicaron las siguientes operaciones al dataset: (1) eliminar las imágenes no faciales, las imágenes de texto y las imágenes de perfil, (2) reetiquetar las imágenes erróneamente anotadas utilizando una CNN, y (3) aumentar los datos para superar el desequilibrio de clases, generando nuevas imágenes faciales para las clases minoritarias con una CycleGAN. Como resultado, la precisión media de la evaluación aumentó del 64% para FER2013 original al 91.76% para la versión equilibrada modificada.

Los trabajos citados abordan el preprocesamiento del dataset antes del entrenamiento de un modelo, sin embargo, las operaciones aplicadas cambian el número total de imágenes, ya sea eliminándolas o aumentándolas. Además, las imágenes se modifican recortándolas, redimensionándolas o retocando el contraste. Nuestro objetivo es preservar las imágenes y el tamaño del dataset, por lo que nos enfocamos en la clasificación errónea, uno de los problemas más influyentes en el menor rendimiento de los modelos de FER in-the-wild.

Kim y Wallraven [57] presentaron un estudio sobre la calidad del etiquetado en AffectNet. Debido al gran tamaño del dataset, se seleccionó un subconjunto con 800 imágenes difíciles de reconocer de las diferentes expresiones categóricas para ser reetiquetadas por 13 anotadores humanos. Tras la reanotación colectiva, el 83.25% del número total de votos no coincidía con las etiquetas originales del dataset. Además, las predicciones de varias ResNets entrenadas en el AffectNet original se comparan con las etiquetas asignadas por los anotadores humanos, encontrando que no hay una buena coincidencia para la expresión categórica. Esta prueba piloto sugiere la baja calidad de etiquetado del dataset original para estas imágenes faciales difíciles, lo que influye en el bajo rendimiento de un modelo de DL. Se menciona que se están realizando trabajos de reanotación más extensos para comprobar un rendimiento más preciso, sin embargo, la anotación manual exige un gran esfuerzo y tiempo.

Nuestro trabajo no requiere ningún tipo de preparación o modificación de las imágenes, y evita disminuir o aumentar su número. Pretendemos reclasificar automáticamente las imágenes para reducir la variabilidad intraclase y el solapamiento interclase del dataset original. Consecuentemente, mejorar el rendimiento en el reconocimiento de emociones.

2.3 Conclusión

Un número limitado de artículos en relación a métodos data-centric se han realizado para mitigar los problemas de los datasets de imágenes. Generalmente se enfocan en un problema específico, a diferencia de nuestra investigación que aborda la falta de calidad

de los datos de manera integral. Planteamos diversas estrategias que utilizan DL para tratar la carencia de imágenes, el mal etiquetado, imágenes con contenido irrelevante, el desequilibrio, el sesgo y la falta de imparcialidad.

Los trabajos citados coinciden en la conveniencia de combinar datasets para entrenar y mejorar la capacidad de generalización de un modelo de FER. Nuestro trabajo propone combinar datasets in-the-wild, a diferencia de estos trabajos que utilizan preferentemente datasets in-the-lab y, ocasionalmente, FER2013 y AffectNet. Aunque el entrenamiento de los modelos se realiza sobre los datasets combinados, no se aprecia una evaluación sobre múltiples datasets.

La clasificación incorrecta y la presencia de imágenes irrelevantes son inconvenientes que enfrentamos con un refinamiento iterativo del dataset utilizando entrenamientos sucesivos de una CNN de clasificación con múltiples categorías y una clasificación binaria, creando una categoría de error frente a las demás.

La investigación relacionada con la expresión facial ha generado una amplia variedad de aplicaciones en la vida real. Nuestro trabajo proporciona un gran dataset combinado a partir de los datasets in-the-wild reclasificados y mejorados, que resulta útil para aplicaciones de FER más cercanas a un contexto práctico. Adicionalmente, proporcionamos un dataset mixto, equilibrado e insesgado como métrica de generalización para la evaluación de modelos entrenados con diferentes datasets. Como alternativa a la generación de imágenes mediante las técnicas convencionales de aumento de datos y GAN, nuestra estrategia se basa en un modelo de difusión de vanguardia para la generación de imágenes sintéticas de alta calidad y realismo.

Capítulo 3

Categorización Web

Este capítulo trata el problema de la categorización de páginas Web a partir de capturas de pantalla. El contenido está basado en nuestros artículos sobre la creación de un dataset visual, cualitativo y cuantitativo de páginas Web [78] y categorización Web utilizando DL [76]. La Sección 3.1 destaca la importancia de la Web y los conceptos fundamentales, identifica la necesidad y beneficios de la categorización Web, describe la mejora del dataset de entrenamiento mediante técnicas de DL, y los resultados esperados. La Sección 3.2 revisa los datasets existentes de páginas Web, el uso dado a estos datasets, así como sus ventajas y desventajas. La Sección 3.3 explica la creación de un novedoso dataset de páginas Web que combina elementos visuales, textuales y numéricos. La Sección 3.4 detalla la depuración de este dataset mediante un detector de páginas Web erróneas, mientras que la Sección 3.5 trata la categorización Web en múltiples categorías temáticas. Ambas aplicaciones sólo con capturas de pantalla y modelos de red convolucional. En la Sección 3.6, se aplica un enfoque de clasificación binaria que mejora el rendimiento de la categorización Web multiclase. Finalmente, la Sección 3.7 se dedica a las conclusiones.

3.1 Introducción

El mundo moderno depende de Internet. Muchas actividades humanas como el comercio, la educación, el entretenimiento y la interacción social tienen aplicaciones digitales soportadas por esta tecnología. Internet y la Web son conceptos estrechamente relacionados. El primer término se refiere a la gran red de redes (la infraestructura), mientras que el segundo se refiere al contenido, que consiste en *sitios Web*, que son una colección

de *páginas Web* interconectadas sobre un tema específico [9]. Desde su invención en la década de 1990, la Web ha revolucionado el acceso a grandes cantidades de datos e información. Factores como la facilidad de uso, la interfaz fácil de usar, la popularidad y el aumento de la conectividad han hecho de la Web una herramienta cotidiana para personas y organizaciones de todos los ámbitos [44].

La Web es una plataforma de comunicación global en constante evolución. El volumen de información es enorme y crece rápidamente, haciéndose más compleja y abarcando todos los temas. La gestión eficaz de tal cantidad y variedad de información es una tarea cada vez más difícil para las técnicas tradicionales. Por ejemplo, organizar los contenidos válidos y filtrar los no válidos (depurar Internet) son retos a los que se enfrenta la Web actual [118]. La inmensa presencia en Internet de páginas Web de error (en construcción, mantenimiento, oferta de dominio, cuenta suspendida, página no encontrada, incompatibilidad del navegador, virus, suplantación de identidad o fallo del servicio), que siguen siendo indexadas y devueltas por los motores de búsqueda, afecta a los webmasters y a los usuarios en general.

Una vez filtradas las páginas Web no válidas, hay que organizar el contenido válido. La clasificación es la técnica básica para gestionar este abundante contenido y es esencial para el trabajo de los motores de búsqueda, directorios Web y los sistemas de recuperación de información [64]. Para hacer frente a este problema, los trabajos de investigación tratan de mejorar los mecanismos de clasificación. Hacerlo manualmente es poco práctico, por lo que la clasificación automática es el método recomendado. Esto suele hacerse analizando el contenido textual de la página Web y el código HTML subyacente. Sin embargo, las páginas Web modernas, que incluyen objetos multimedia, streaming de vídeo e imágenes que se comparten, hacen que la extracción de información sea cada vez más complicada [106]. Esto ha motivado el interés científico y nuevas áreas de investigación y desarrollo.

Los problemas de clasificación son una especialidad de la IA. Aquí presentamos el desarrollo de aplicaciones basadas en DL que se ocupan de la depuración y organización de páginas Web. Implementamos el reconocimiento automático de páginas Web erróneas como un problema de clasificación binaria y la categorización de páginas Web válidas dentro de varios temas como un problema multiclase. Para disminuir la alta dificultad del problema multiclase, se reformula como varias clasificaciones binarias, y se aplican técnicas de regularización y un mecanismo de predicción que considera todos los clasificadores binarios. Ambas aplicaciones utilizan exclusivamente capturas de pantalla como entrada para entrenar modelos basados en CNNs.

3.2 Datasets de páginas Web

Nuestra revisión de la literatura sobre los datasets de páginas Web existentes se resume en la Tabla 3.1. Para facilitar el análisis y la comparación, los hemos dividido en dos grupos según el tamaño del dataset utilizado: pequeños (menos de 1000 instancias) y grandes.

Tabla 3.1: Resumen de los datasets de páginas Web y sus características principales.

Autor, año	Tamaño	Categorías	Tipo de datos	Propósito
De Boer *et al.*, 2011	Pequeño: 60 screenshots	Noticias, hoteles, conferencias y celebridades	Imágenes	Estético y temático clasificación con ML
Reinecke *et al.*, 2014	Pequeño: 430 screenshots	Genérico	Imágenes	Estético clasificación
Khani *et al.*, 2016	Pequeño: 430 screenshots	Buena o mala	Imágenes	Estético clasificación con ML
López *et al.*, 2017	Pequeño: 280 páginas Web	Comida, animales, moda, naturaleza, hogar y vehículos	URL e imágenes extraídas desde HTML	Temático clasificación con ML
López *et al.*, 2019	Pequeño: 365 páginas Web	Comida, vehículos, animales, moda, diseño interior y paisajismo	URL e imágenes extraídas desde HTML	Temático clasificación con ML
CIRCL, 2019	Pequeño: 460 screenshots	Phishing verificado o potencial	Imágenes	Análisis de eventos de seguridad
ImageNet, 2009	Extenso: 1840 screenshots	Genérico	Imágenes	Recurso para investigación en imágenes
Nordhoff *et al.*, 2018	Extenso: 80901 screenshots	Genérico	URL, métricas e imágenes	Estético y diseño Web
CIRCL, 2019	Extenso: 37500 screenshots	Onion Website (Web oculta, no indexada)	Imágenes	Análisis de eventos de seguridad
Universidad de Alicante, 2019	Extenso: 8950 screenshots etiquetados	Buen y mal diseño	Imágenes etiquetadas	Estético Categorización

El interés científico en la categorización Web está aumentando. DeBoer *et al.* [9] utilizan un pequeño dataset exclusivamente de capturas de pantalla (*screenshots*) de páginas Web y algoritmos de ML distintos de las redes neuronales. Aunque realizan experimentos de clasificación binaria de la estética y la antigüedad de las páginas, nos enfocamos en la categorización multiclase dentro de cuatro temas (noticias, hoteles, conferencias y celebridades). Estas categorías son muy diferentes entre sí, por lo que

el problema de categorización es menos complicado. En cuanto a la disponibilidad del dataset, no hay ningún enlace para descargar las capturas de pantalla.

López *et al.* [70] y [71] tienen datasets con más páginas Web que incluyen sus respectivas imágenes y enlaces URL. Sin embargo, estas imágenes no son capturas de pantalla, sino elementos de la página Web. La URL se utiliza para descargar las imágenes a partir del código HTML y analizarlas para su categorización. Los autores realizan una clasificación en categorías temáticas, con aprendizaje de transferencia, ajuste fino y el algoritmo de k vecinos más cercanos, utilizando como entrada la URL para descargar las imágenes que forman parte del sitio Web. Aunque hay más categorías que en el trabajo anterior, siguen cubriendo temas muy diferentes. En ninguno de los dos trabajos se ofrece un enlace de descarga.

Entre los datasets más pequeños, Reinecke *et al.* [93] proponen el más significativo, que podría ser un recurso útil para la investigación y desarrollo a pequeña escala. Abarca varios países del mundo y diversos temas, además se puede descargar. Sin embargo, es estrictamente visual e insuficiente para las necesidades actuales. Su finalidad está más orientada al análisis estético y la clasificación.

Khani *et al.* [55] utilizan el dataset anterior como entrada para una CNN, se reduce la dimensionalidad de características mediante PCA y se clasifican las páginas Web con el algoritmo SVM como buena o mala estética, pero no tratan la categorización multiclase.

El Centro de Respuesta a Incidentes Informáticos de Luxemburgo (CIRCL) es una iniciativa gubernamental creada para responder a amenazas e incidentes de seguridad informática. CIRCL [24] ofrece un dataset con más de 400 capturas de pantalla de sitios Web de phishing verificados o potenciales. Además, está disponible un extenso dataset con más de 37000 imágenes [23], correspondientes a capturas de pantalla de sitios Web pertenecientes a la *Dark-Web*, la faceta problemática de la Web asociada a la ciberdelincuencia, el odio y el extremismo [28]. Ambos datasets pueden descargarse fácilmente, pero las imágenes representan falsificaciones y páginas Web ocultas, por lo que estos datasets tendrían una aplicación limitada.

ImageNet [17] es la más popular de las bases de datos de imágenes. Incluye millones de imágenes organizadas según la jerarquía *WordNet*[1]. La sección destinada a sitios Web contiene 1840 capturas de pantalla de diferentes países e idiomas sin categorizar. Algunas capturas de pantalla aparecen recortadas y su descarga requiere registro y autorización.

El dataset creado por la Universidad de Alicante [110] recopila 8950 capturas de pantalla de páginas Web para analizar y evaluar la calidad del diseño Web. La mitad

[1]Una gran base de datos léxica del inglés. Los sustantivos, verbos, adjetivos y adverbios se agrupan en conjuntos de sinónimos cognitivos (synsets). https://wordnet.princeton.edu

de las imágenes proceden del sitio *Awwwards*[2] y están etiquetadas con "buen diseño", mientras que la otra mitad se extraen de páginas amarillas, etiquetadas con "mal diseño". Este dataset sirve para el trabajo académico de la institución.

Por su parte, Nordhoff *et al.* [85] destaca porque abarca un mayor número de páginas Web. Sin embargo, estas sólo proceden de 44 países, los parámetros son puramente estéticos y la descarga de imágenes no es directa.

Los datasets actuales sobre categorización de páginas Web son generalmente pequeños, se limitan a capturas de pantalla y, en raras ocasiones, proporcionan las URLs respectivas. Las categorías se restringen a la estética y unos pocos temas. La investigación sobre problemas relacionados con la Web necesita crear datasets más sofisticados y de mayor tamaño. Nuestro trabajo pretende contribuir a estos aspectos con un dataset que considere todos los países del mundo, incluya atributos relacionados con la estructura de la página Web, sea de uso más general y esté disponible para descarga.

3.3 Creación del dataset de páginas Web

En esta sección, presentamos una metodología para crear desde cero un dataset amplio y disponible, que incorpora la representación visual, complementada con atributos cualitativos y cuantitativos, de modo que una página Web sea mejor caracterizada. El flujo de trabajo se representa en la Figura 3.1.

3.3.1 Diseño del dataset

El primer paso consiste en definir los elementos que componen el dataset según el objetivo propuesto. Nuestro interés se centra en la estructura y el contenido de una página Web, por lo que seleccionamos una serie de atributos cualitativos y cuantitativos para la estructura, y un webshot para el aspecto visual. De este modo, creamos un único dataset mixto en tipo de datos, diseñado para combinar datos visuales, textuales y numéricos, más extenso y descriptivo que los actuales datasets de páginas Web. La Tabla 3.2 muestra la lista de los elementos de nuestro dataset, y que se detallan a continuación.

Un *webshot* es una imagen digital de la página Web completa, a diferencia de una captura de pantalla, que puede aparecer recortada porque sus dimensiones superan las del dispositivo de visualización, obligando al usuario a desplazar. El elemento *name* que se da al webshot es clave, sigue una convención para identificar la fuente, la categoría y el país de la página Web. También es el vínculo entre el webshot y los atributos

[2] https://www.awwwards.com

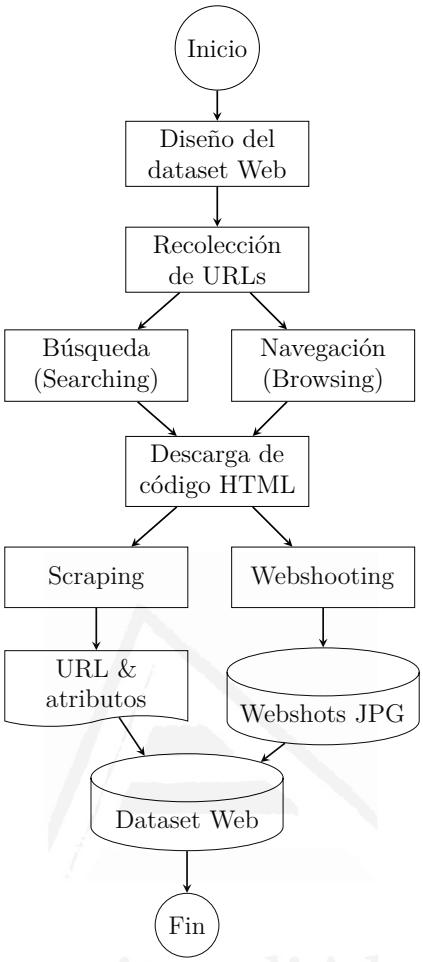


Figura 3.1: Metodología para producir el dataset de páginas Web.

cualitativos y cuantitativos. La *URL* (Localizador Uniforme de Recursos) es la dirección de Internet de la página Web junto con el mecanismo de recuperación (*http/https*). Se coloca en la barra de direcciones de los navegadores (browsers), que son los programas que muestran el contenido al usuario. Recopilamos URLs de todo el mundo para cubrir diferentes características y preferencias culturales, de modo que el dataset incluye atributos relacionados con las ubicaciones geográficas, como el país y el continente. Las páginas Web obtenidas pertenecen a las siguientes categorías: arte y entretenimiento, negocios y economía, educación, gobierno, noticias y medios de comunicación, y ciencia y ambiente. Consideramos estas categorías porque forman parte del directorio Web utilizado en este trabajo y que se explica en la Sección 3.3.2.2. Adicionalmente, incluimos los siguientes parámetros cuantitativos, que proporcionan una visión general de la estructura y la calidad de una página Web:

Tabla 3.2: Estructura del dataset de páginas Web.

Elemento	**Tipo**	**Descripción**
Webshot	Visual	Screenshot de una página Web entera en formato JPG
Name	Texto	Identificador único dado al webshot
URL	Texto	Enlace para localizar y desplegar una página Web
Country	Cualitativa	País de origen de la página Web
Continent		Región que agrupa países
Category		Tópico principal de la página Web
Time	Cuantitativa	Tiempo de descarga del código de la página Web
Bytes		Tamaño en bytes del código de la página Web
Images		Número de imágenes de la página Web
Script_files		Número de archivos ejecutables de la página Web
CSS_files		Número de archivos de estilo de la página Web
Tables		Número de etiquetas *table* en el código fuente
iframes		Número de etiquetas *iframe* en el código fuente
Style_tags		Número de etiquetas *style* en el código fuente
Img_bytes		Peso del webshot en bytes
Img_width		Ancho del webshot en píxeles
Img_height		Alto del webshot en píxeles

- El tiempo de descarga del código fuente, ya que los usuarios quieren esperar lo menos posible para ver una página Web [60].
- El tamaño en bytes, pues cuanto mayor sea, más lentas serán la descarga y la visualización de la página Web.
- La cantidad de imágenes ya que aumentarán el tiempo de descarga. Una página Web no es necesariamente más atractiva por tener más imágenes, es recomendable un equilibrio entre todos los tipos de información [80].
- La cantidad de *scripts*, los cuales son archivos externos para dotar a la página Web de una funcionalidad más compleja. Es aconsejable reducir su cantidad porque aumentan el tráfico de red y el tiempo de descarga.
- El número de archivos *CSS*, mismos que maquetan el estilo. Se vuelven una carga extra y retrasan la visualización de la página Web, lo ideal es que haya uno.
- La cantidad de tablas, que a menudo se utilizan para estructurar el contenido de una página Web, pero esto se desaconseja debido a elementos apropiados como las etiquetas *div*.

- Las etiquetas *iFrames* insertan una página Web dentro de otra, lo que actualmente no es una buena práctica. Las etiquetas *style* no son recomendables ya que hay archivos CSS.
- El peso en bytes y las dimensiones (ancho y alto) en píxeles de cada webshot.

3.3.2 Recolección de URLs

Una vez estructurado el dataset, el siguiente paso es recopilar URLs de todo el mundo relacionadas con las categorías dadas. De cada URL, se descargó el código HTML, se extrajeron los atributos cuantitativos y cualitativos mediante *scraping* y se hizo un webshot de toda la página Web. Para obtener un mayor número de URLs, utilizamos las dos formas de buscar información en la Web: la *Búsqueda* y la *Navegación*. La búsqueda requiere que el usuario traduzca una necesidad de información en consultas, mientras que la navegación es una actividad humana básica y natural, que se produce en un entorno de información donde los objetos de información son visibles y están organizados. A continuación, describimos cómo ambas técnicas nos permitieron recopilar URLs y, a partir de ellas, extraer atributos y capturar webshots, todo ello mediante scripts en Python y R.

3.3.2.1 Búsqueda

El motor de búsqueda de Google pregunta por palabras o frases relacionadas con el tema de interés. Para evitar la tarea repetitiva de escribir la consulta en la página de búsqueda de Google y recuperar manualmente las URLs de respuesta, automatizamos el proceso mediante un script de Python, donde:

1. El nombre del país y su código de Internet se extraen iterativamente de un archivo de texto[3].
2. La consulta de búsqueda tiene la siguiente estructura: 'site:' + *código de país* + ' business OR economy OR marketing OR computers OR internet OR construction OR financial OR industry OR shopping OR restaurant' + ' ext:html'

   OR, *site* y *ext* son operadores o palabras reservadas que pueden utilizarse en frases de consulta dentro del motor de búsqueda de Google. El operador "OR" concatena varias palabras de búsqueda relacionadas con la categoría. El operador "site" especifica el dominio geográfico de primer nivel de Internet asignado para cada

[3] https://osf.io/yrmx8

país, por ejemplo, ".es" para España. El operador "ext:html" produce resultados exclusivamente con esta extensión de archivo.

3. La petición devuelve la página de resultados de Google con los 100 primeros enlaces, que se utiliza para conseguir una distribución aproximadamente uniforme de las páginas Web según el país y la categoría.

4. Los enlaces de las páginas Web se extraen escaneando automáticamente el código fuente de la página de resultados (scraping), generando un archivo de texto que contiene las URLs y sus atributos de país, continente y categoría.

5. Para el resto de categorías, las consultas son:

   'site:' + *código de país* + ' arts OR entertainment OR dance OR museums OR theatre OR literature OR artists OR galleries' + ' ext:html'

   'site:' + *código de país* + ' education OR academy OR university OR college OR school' + ' ext:html'

   'site:' + *código de país* + ' government OR military OR presidency ' + ' ext:html'

   'site:' + *código de país* + ' news OR media OR magazine OR radio OR television OR newspaper' + ' ext:html'

   'site:' + *código de país* + ' science OR environment OR archaeology' + ' ext:html'

3.3.2.2 Navegación

Esta técnica utiliza un *Directorio Web*, el cual es un sitio Web especializado que consiste en un catálogo de enlaces a otros sitios Web. La construcción, el mantenimiento y la organización por categorías y subcategorías corren a cargo de expertos humanos, a diferencia de los motores de búsqueda, que lo hacen automáticamente. Para incluir una URL, los especialistas realizan un proceso de revisión, análisis y evaluación para verificar los requisitos determinados por el directorio Web. Algunos directorios Web han sobrevivido a la popularidad de buscadores como Google. Podemos destacar *Best of the Web (BOTW)* (Figura 3.2), uno de los más reconocidos por su calidad, alcance global, amplio abanico de categorías y subcategorías, nivel de tráfico (visitas al mes), fiabilidad, número de enlaces y exigentes requisitos.

En lugar de una consulta o frase de búsqueda, es necesario conocer la estructura jerárquica de los directorios y subdirectorios hasta la URL de interés. Aprovechamos la organización por países y categorías definida por BOTW, por ejemplo, para Grecia:

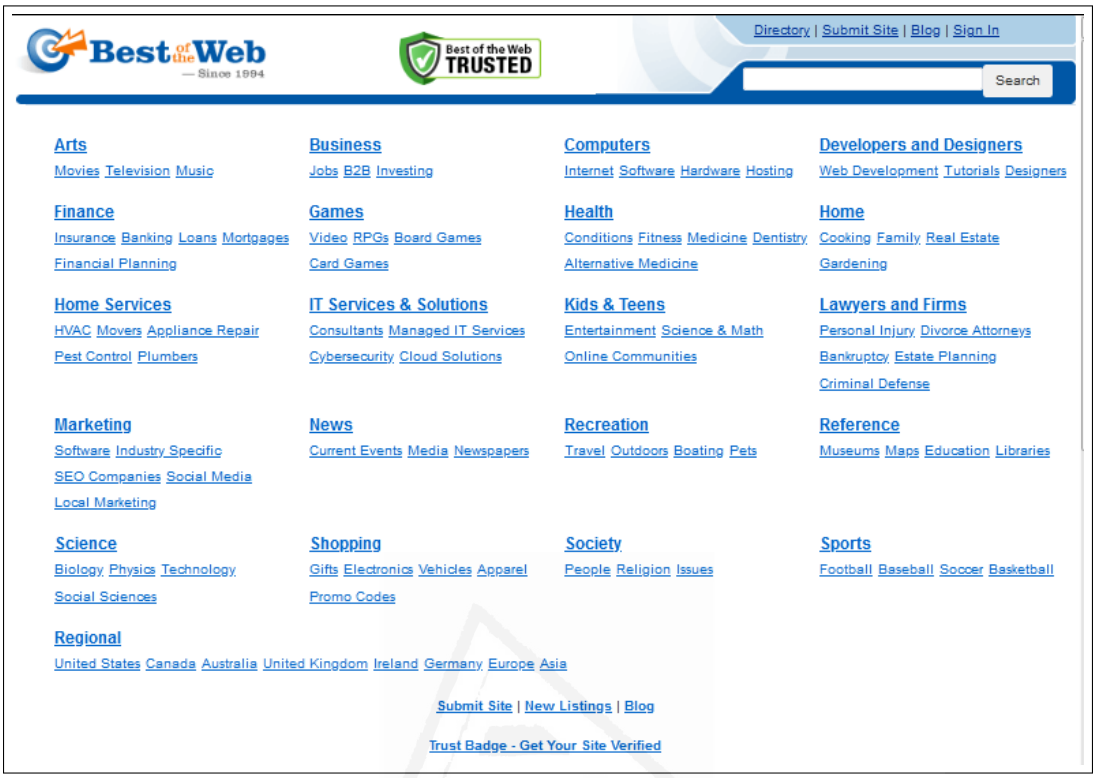


Figura 3.2: Directorio Web BOTW (https://botw.org/).

```
https://botw.org/top/Regional/Europe/Greece/Arts-and-Entertainment/
https://botw.org/top/Regional/Europe/Greece/Business-and-Economy/
https://botw.org/top/Regional/Europe/Greece/Education/
https://botw.org/top/Regional/Europe/Greece/Government/
https://botw.org/top/Regional/Europe/Greece/News-and-Media/
https://botw.org/top/Regional/Europe/Greece/Science-and-Environment/
```

Recopilamos las URLs publicadas dentro de cada categoría mediante un script de Python[4] que implementa las siguientes acciones:

1. Iterativamente lee el nombre de cada país dentro de un archivo de texto.
2. Establece la ruta correspondiente a la categoría, que siempre tendrá la misma estructura, en la que sólo cambia el nombre del país, por ejemplo: 'https://botw.org/top/Regional/' + *nombre del país* + '/Science-and-Environment/'
3. Se realiza una conexión a la dirección Web formada para obtener el código fuente de la página de resultados y extraer las URLs de cada categoría.

[4] https://osf.io/73sc2

4. Los enlaces se almacenan en un archivo de texto[5], que se importa a una hoja de cálculo donde se aplica un filtro para seleccionar sólo aquellos enlaces que pertenecen a un país en concreto.

3.3.3 Recolección de atributos

Tras recopilar y almacenar las URLs de las dos fuentes descritas, implementamos un script para:

1. Leer secuencialmente los enlaces URL almacenados en el archivo de texto[6].
2. Realizar una conexión a través del navegador a cada uno de estos enlaces.
3. Descargar y analizar el código fuente de la página Web, obteniendo los atributos especificados en el diseño del dataset, es decir, tiempo de descarga en segundos, tamaño total en bytes, número de imágenes, archivos de script, archivos CSS, tablas, etiquetas iFrames y etiquetas de estilo.

Este script incluye la librería de scraping Web denominada *Beautiful Soup*, que nos permite definir y extraer los atributos a partir del código HTML de cada página Web.

3.3.4 Recolección de webshots

Utilizamos los paquetes *Webshot* y *PhantomJS* para crear un script en R que lee cada URL dentro de los archivos de texto generados en las secciones de búsqueda y navegación, toma una instantánea de toda la página Web y la guarda como imagen JPG. El nombre asignado a la imagen es un identificador único para conocer la fuente, la categoría y el país al que pertenece la página Web, por ejemplo, *B2PaísesBajos_791.jpg* indica un webshot obtenido mediante navegación, perteneciente a la categoría número dos (negocios y economía) y que procede de los Países Bajos. El número tras el guión bajo sólo establece un orden secuencial. Nótese que este identificador es la clave para asociar el webshot a sus respectivos atributos cualitativos y cuantitativos.

De este modo, ha sido posible vincular dos soportes de almacenamiento diferentes, es decir, una hoja de datos y una carpeta de imágenes. Además, este script incluye funciones para obtener tanto el peso del webshot en bytes como sus dimensiones (anchura y altura) en píxeles, con el fin de caracterizar mejor una página Web. Todas las imágenes y la hoja de datos están disponibles para su visualización y descarga. La Figura 3.3

[5] https://osf.io/hjwgm
[6] https://osf.io/gk3p2

presenta una pequeña muestra del dataset obtenido, con un ejemplo de cada categoría que está compuesto por el webshot de la página Web y sus respectivos atributos de tipo cualitativo y cuantitativo.

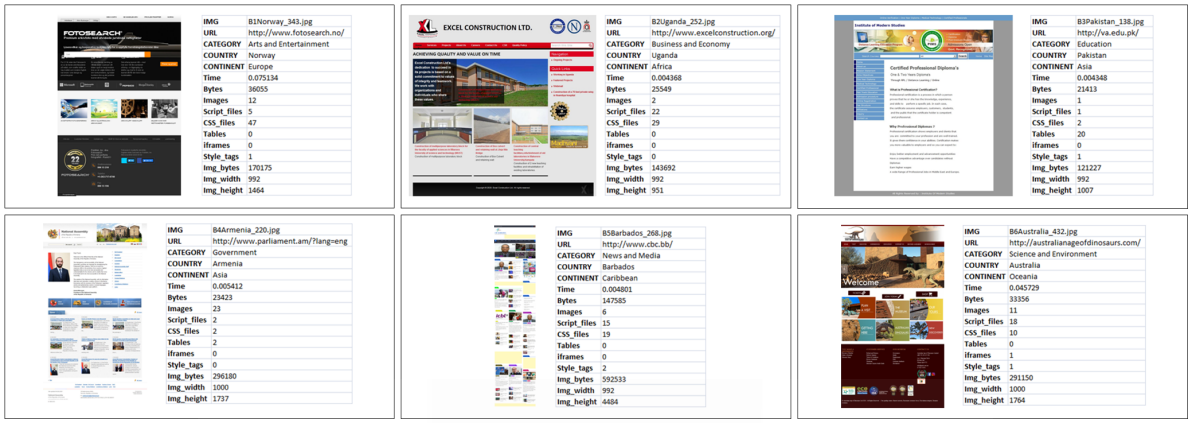


Figura 3.3: Muestra del dataset de páginas Web con un ejemplo de cada categoría.

3.4 Reconocimiento de páginas Web de error

Durante nuestra recopilación automática de datos, algunos eventos bloquearon la descarga del código HTML o la obtención del webshot. Las causas incluyen la solicitud de aceptación manual de *cookies* y certificados *SSL*, mensajes de error como *HTTP 403 Prohibido*, *HTTP 404 No encontrado*, *HTTP 406 No aceptable*, *HTTP 909 Permiso denegado*, y la superación del tiempo de espera. Utilizamos el manejo de excepciones dentro de los scripts para evitar interrupciones en la ejecución de los programas. Cuando se producía un error, se asigna a los campos asociados a los atributos o al webshot el valor '−1'. De este modo, los programas pueden continuar su ejecución y solventar la inexistencia de webshots o atributos.

Para el dataset final, consideramos sólo las URLs que tenían el webshot respectivo, ya que éste es el elemento más importante de nuestro trabajo. Sin embargo, tras una breve revisión visual, se detectaron varias páginas Web de error como sitios en construcción, mantenimiento, oferta de dominio, cuenta suspendida, página no encontrada, incompatibilidad del navegador, riesgos de virus o phishing. Algunas de ellas se muestran en la Figura 3.4.

Los webshots mostrando mensajes de error no son útiles para el dataset, por lo que son eliminados. Aunque el tamaño del dataset se vuelve menor, obtenemos un dataset más limpio. Dado que las conexiones URL correspondientes a estos webshots no devolvían

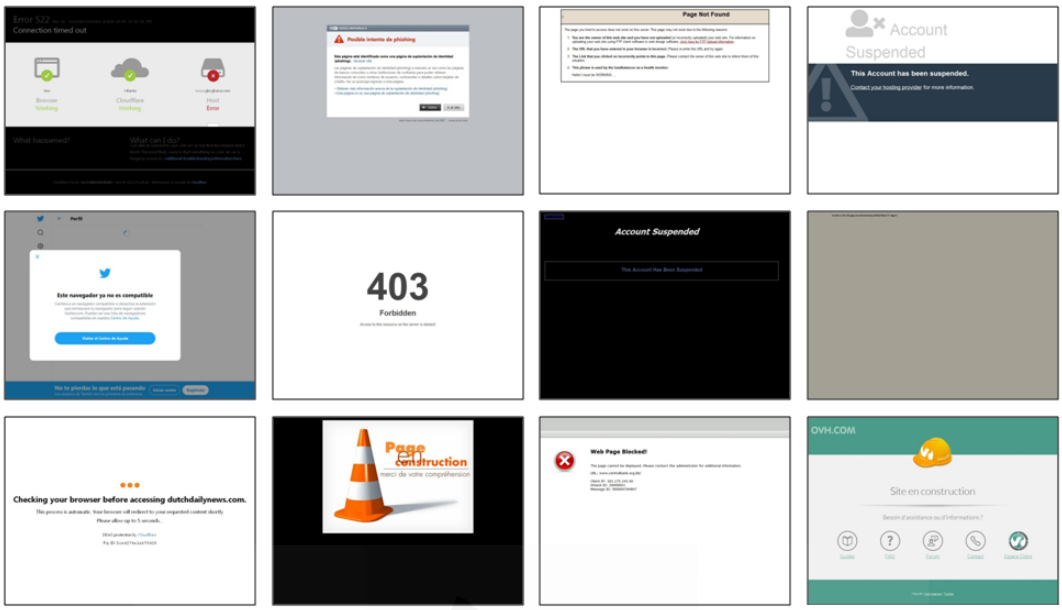


Figura 3.4: Una muestra de las páginas Web de error.

mensajes de error HTTP 403 o HTTP 404, ni el código HTML contenía frases como "cuenta suspendida" o "página en construcción", no fue posible realizar un análisis de tipo textual. Por tal razón, implementamos un analizador de imágenes para evitar la verificación manual y visual de miles de webshots, que requiere un tiempo y un esfuerzo excesivos.

Utilizamos una CNN para detectar las páginas Web de error y separarlas en una carpeta específica, todo ello de forma automática. En este sentido, las páginas Web que no contienen información útil se destinan a la carpeta “ERROR”, mientras que las páginas Web que contienen información valiosa son ubicadas en la carpeta “VALID”. Aquí presentamos una detección automática de páginas Web de error basada exclusivamente en sus webshots. Consiste en determinar si una página Web pertenece a la categoría “VALID” o a una categoría “ERROR”, es decir, es un problema de clasificación binaria. Para ello, seguimos la metodología que se muestra en la Figura 3.5, cada una de las fases se explica posteriormente en detalle.

3.4.1 Selección de datos

El principal recurso para el proceso de aprendizaje son los datos. En nuestro caso, los datos son las imágenes que servirán de entrada para un entrenamiento que pretende, de forma iterativa, obtener un resultado conocido de validez o error y, si se alcanza una precisión aceptable, realizar predicciones. El entrenamiento requiere imágenes asociadas

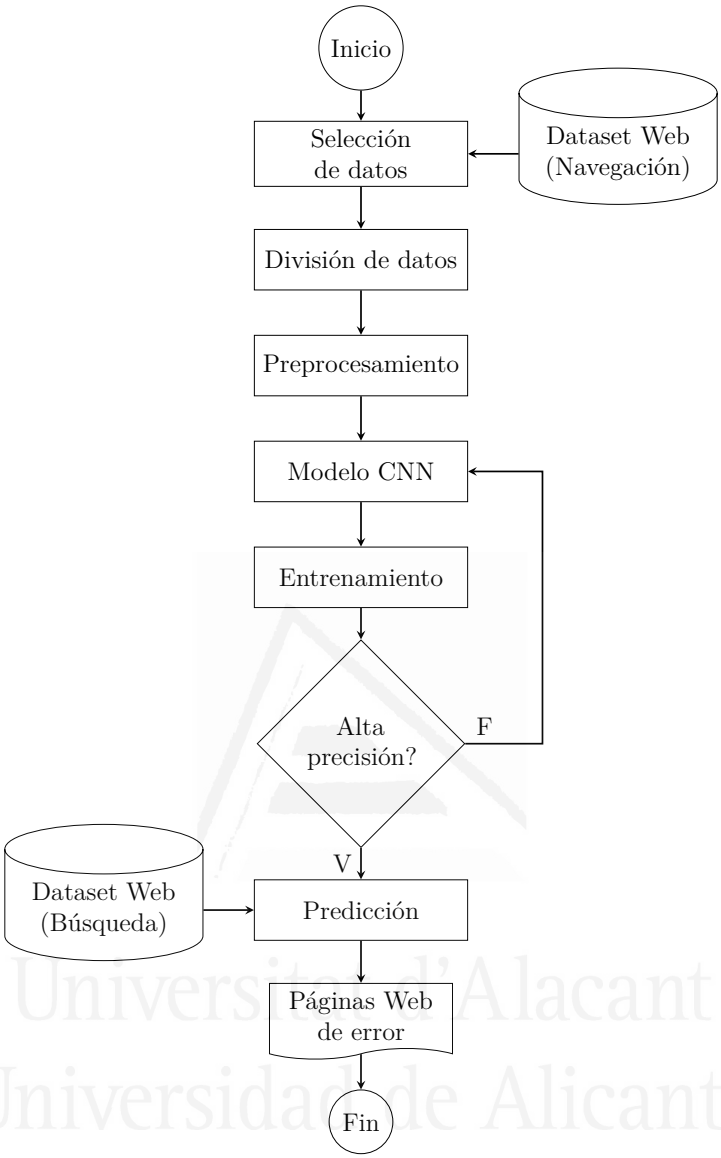


Figura 3.5: Metodología para detectar páginas Web de error.

a su categoría respectiva: válida o error. Dado que nuestro dataset consta de dos grupos de imágenes (navegación y búsqueda), seleccionamos los webshots del subconjunto más pequeño (navegación) para realizar una inspección visual exhaustiva y clasificar las imágenes manualmente, obteniendo los resultados que se muestran en la Tabla 3.3. Una vez ajustado el modelo de red neuronal, este clasifica cada webshot del subconjunto mayor (búsqueda) como página Web válida o de error.

El dataset para entrenar el modelo de detección de páginas Web de error tiene 3609 imágenes, 427 webshots de error y 3182 webshots válidas, que se almacenaron en *Google Drive* en carpetas separadas: "VALID" y "ERROR" (Figura 3.6).

Tabla 3.3: Dataset para detección de páginas Web de error (webshots obtenidos por navegación).

	Navegación		
Categoría	**Webshots**	**Válidas**	**Error**
Arte y entretenimiento	447	397	50
Negocios y economía	1058	892	166
Educación	419	368	51
Gobierno	730	669	61
Noticias y medios	458	394	64
Ciencia y ambiente	497	462	35
Total	**3609**	**3182**	**427**

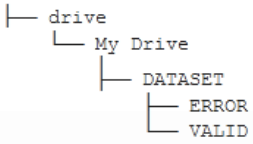


Figura 3.6: Organización del dataset para detección de páginas Web de error.

Utilizamos *Google Colaboratory* por ser una plataforma gratuita que ofrece un potente hardware y no requiere instalación ni configuración, soporta Python a través de un cuaderno de programación online, e incluye los paquetes y las librerías para facilitar el trabajo con DL como *TensorFlow*, *Keras*, y *Sklearn*, entre otros. El paso inicial es la conexión a la fuente de datos de *Google Drive*, en la que se han almacenado las imágenes de nuestro dataset dentro de las carpetas y subcarpetas de la Figura 3.6. Esta estructura facilita el etiquetado de las imágenes con su categoría correspondiente.

3.4.2 División de datos

Una de las tareas que caracteriza el proceso de aprendizaje automático es la división de los datos. Dado que sólo disponemos de 3609 imágenes, consideramos dos subconjuntos: entrenamiento y validación (Tabla 3.4). El subconjunto de entrenamiento contiene el mayor número de imágenes (80%) y se utiliza para aprender y ajustar los parámetros del modelo, mientras que el subconjunto de validación (20%) se utiliza para evaluar la capacidad del modelo.

Tabla 3.4: Partición del dataset para la detección de páginas Web de error.

Conjunto	**Webshots**	**Válidas**	**Error**	**Porcentaje**
Entrenamiento	2886	2545	341	80%
Validación	723	637	86	20%
Total	3609	3182	427	100%

Aunque lo más adecuado es un dataset equilibrado, es decir, con el mismo número de casos de error y de casos válidos, utilizamos todas las imágenes para tratar de ampliar la generalización. Para dividir automáticamente en carpetas de entrenamiento y validación, es útil instalar e importar el paquete *split-folders*[7]. Es necesario especificar el directorio de imágenes, el directorio de salida y la proporción a dividir (80% y 20%, respectivamente). El resultado es una nueva estructura de directorios. Dentro de la carpeta "SPLIT", se crean las carpetas "train" y "val", y dentro de cada una de éstas, las carpetas "ERROR" y "VALID".

3.4.3 Preprocesamiento de datos

Las imágenes deben ser preparadas convenientemente antes del proceso de entrenamiento. Primero, normalizamos los píxeles pasando de valores enteros entre 0 y 255 a una escala entre 0 y 1. La clase *ImageDataGenerator* de Keras divide todos los valores de los píxeles por el valor máximo (255).

Luego, las imágenes tienen dimensiones diferentes de ancho y alto, por lo que se redimensionan todas a 256x256 píxeles mediante el parámetro *target_size* del método *flow_from_directory*. Esta operación se realizó en grupos de 32 imágenes (*batch_size*) que se etiquetan para la clasificación binaria (*class_mode*) según la carpeta donde se almacenan (válida o error) dentro del directorio de entrenamiento.

De este modo, los valores pequeños tanto de los píxeles como de las dimensiones ayudan a acelerar el proceso de entrenamiento. El procedimiento anterior también se aplica a los datos de validación, cambiando únicamente el directorio.

3.4.4 Modelo CNN

La arquitectura del modelo se basa en la CNN propuesta por Liu *et al.* para detectar sitios Web maliciosos [68]. Al tratarse de un problema similar, sólo hemos aplicado pequeñas adaptaciones. Su estructura (Figura 3.7) se compone de dos partes:

- Base convolucional, para la extracción automática de características. La imagen de 256x256 píxeles se separa en 3 canales de color RGB. Esta entrada es procesada por 3 capas convolucionales con su respectiva función de activación ReLU (Rectified Linear Unit) y capas de maxpooling. Las dos primeras convoluciones utilizan 32 filtros (kernels), mientras que la tercera tiene 64, con un tamaño de 3x3, en contraste con el tamaño del pool de 2x2.

[7] https://pypi.org/project/split-folders/

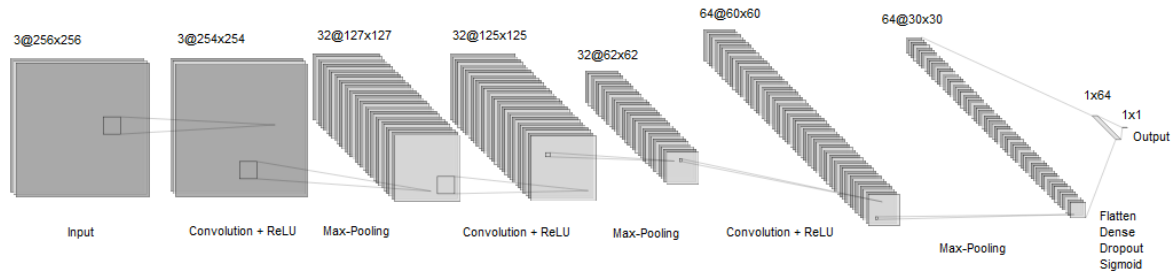


Figura 3.7: Arquitectura CNN para la detección de páginas Web de error.

- Clasificador binario, que es una capa totalmente conectada para recibir las características de forma aplanada y aplicar dropout para reducir el sobreajuste. La función sigmoide genera la predicción como un valor de probabilidad entre 0 y 1. Si el valor es superior a 0.5, la página Web es válida y, en caso contrario, se trata de una página Web de error.

El framework de Keras proporciona funciones para implementar esta arquitectura desde cero de forma sencilla, simplemente añadiendo en secuencia las capas convolucionales, de activación, pooling, dropout, flatten y dense, y especificando sus respectivos parámetros.

3.4.5 Entrenamiento

Antes de comenzar el entrenamiento, debemos definir explícitamente los *hiperparámetros* requeridos por la red neuronal de clasificación binaria. Podemos establecer la función de pérdida que será minimizada por el algoritmo de optimización y la precisión de la clasificación como la métrica que será recogida y reportada por el modelo.

Algunas horas de cálculo se consumen en la ejecución de 20 iteraciones (épocas) del dataset de entrenamiento de 2886 imágenes. Cada iteración consta de 90 grupos de 32 imágenes. La precisión alcanzada es del 96.6% (iteración 20), mientras que en la etapa de validación la precisión es del 97.16% (iteración 16). La evolución del proceso se resume en los gráficos de las curvas de aprendizaje mostradas en la Figura 3.8.

Las fases de entrenamiento y validación alcanzaron un alto nivel de precisión, progresando ambas al mismo nivel, lo cual es deseable. El modelo se ajusta muy bien a las imágenes proporcionadas, pero no se sabe cómo se comporta con imágenes nuevas (generalización). Esta preocupación se aborda analizando la diferencia entre las pérdidas de entrenamiento y las de validación. Esta última, a pesar de oscilar, no varía mucho de la otra hasta la iteración 17, después de la cual, empiezan a separarse, con la posibilidad de sobreajuste. Por tanto, el modelo es almacenado con la precisión y los pesos de la

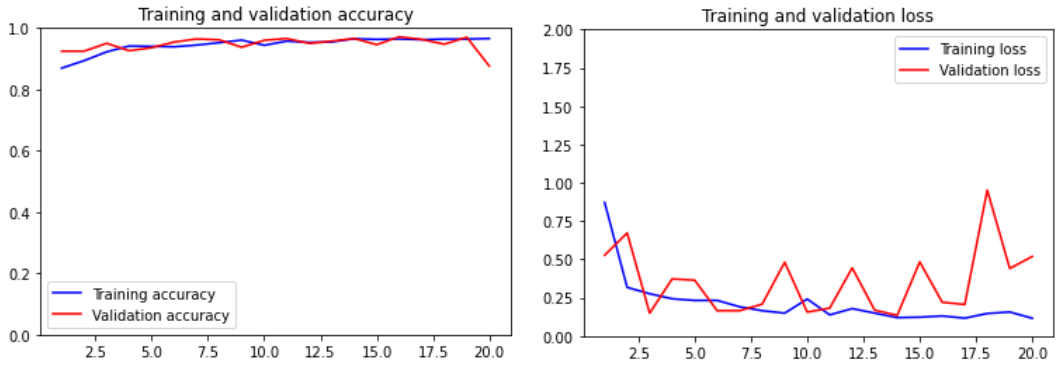


Figura 3.8: Precisión y pérdida en las fases de entrenamiento y validación.

iteración número 16. Podemos decir que el modelo es capaz de distinguir aceptablemente las páginas Web de error y las válidas, por lo que pasamos a la fase de predicción.

3.4.6 Predicción

Las imágenes del conjunto más grande (búsqueda) de nuestro dataset se convierten en la entrada del modelo ya entrenado y validado. Utilizamos la librería *google.colab* para cargar el archivo (webshot) desde la unidad local con un botón de selección.

Una vez cargado el archivo de imagen, se preprocesa utilizando las librerías *NumPy* y *preprocessing* para transformar la imagen en un array con una forma adecuada y unos valores de píxel normalizados para el modelo, el cual realiza la predicción.

El resultado para las imágenes seleccionadas de una en una se muestra en la Figura 3.9. Aparece la imagen Web redimensionada y la predicción como valor de probabilidad entre 0 y 1. En el lado izquierdo, el valor es inferior a 0.5, por lo que la categoría asignada es de error; mientras que, en el lado derecho, el valor es muy próximo a 1, por lo que es un caso de página Web válida.

Además de hacer predicciones una a una, nuestro propósito es generar predicciones para grupos de imágenes. Aprovechamos que nuestro dataset está organizado por categorías temáticas, por lo que basta con escoger una lista de archivos mediante el botón de selección. Por ejemplo, podemos seleccionar todas las imágenes de la categoría "Arte y entretenimiento". Para un grupo de archivos, los resultados se presentan de manera textual. Un extracto se muestra en la Figura 3.10.

Esta lista de predicciones se pasa a una hoja de cálculo y, mediante un filtro, se seleccionan las páginas Web de la categoría de error y se guardan en un archivo con formato de texto (*list.txt*). Este archivo es el argumento para ejecutar el siguiente comando:

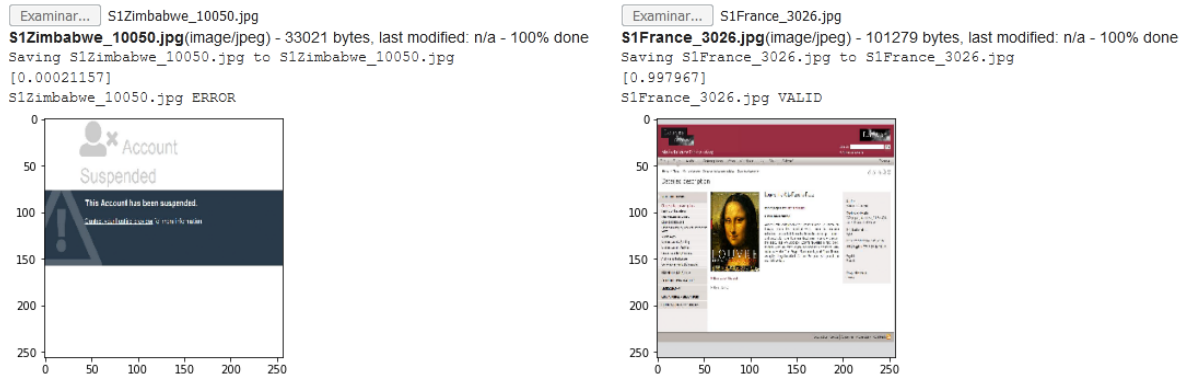


Figura 3.9: Predicción de error (izquierda) y página Web válida (derecha).

```
S1Belize_1298.jpg VALID
S1Denmark_2410.jpg VALID
S1Spain_2654.jpg VALID
S1Moldova_5622.jpg ERROR
S1Guadeloupe(France)_3334.jpg VALID
S1Nicaragua_6280.jpg VALID
S1Denmark_2409.jpg VALID
S1Jamaica_4543.jpg VALID
S1Thailand_8478.jpg VALID
S1Taiwan_9033.jpg VALID
S1UnitedArabEmirates_103.jpg VALID
S1NewZealand_6630.jpg VALID
S1Uruguay_9424.jpg VALID
S1Lithuania_5276.jpg VALID
S1Mongolia_5891.jpg VALID
S1Bahamas_1173.jpg VALID
S1Reunion(France)_7335.jpg VALID
S1Zimbabwe_10064.jpg ERROR
S1SaintVincentandtheGrenadines_9571.jpg VALID
S1Iran_4324.jpg VALID
S1Andorra_75.jpg VALID
S1Liechtenstein_5133.jpg VALID
S1Bangladesh_702.jpg ERROR
S1Iran_4334.jpg VALID
```

Figura 3.10: Predicción para un grupo de páginas Web.


```
cat list.txt | ForEach {mv $_ ERROR}}
```

Esta línea de comandos mueve todas las imágenes de su carpeta original a la carpeta "ERROR". Se ejecuta en *Windows* (interfaz PowerShell), aunque es fácilmente adaptable a diferentes sistemas operativos como *Linux*.

Como resultado de la predicción, se encontraron 822 páginas Web erróneas y 7747 válidas. Una vez clasificadas y separadas las imágenes, la verificación visual es mucho más rápida, y es posible determinar manualmente los aciertos y fallos del clasificador. Identificamos 1214 páginas Web de error reales y 7355 páginas Web válidas reales. Se realiza el mismo procedimiento con el resto de imágenes de las demás categorías. Los resultados se resumen en la Tabla 3.5.

Tabla 3.5: Resultados de la clasificación de páginas Web de error.

	Búsqueda					
Categoría	Webshots	Válida (Predicción)	Error (Predicción)	Válida (Real)	Error (Real)	Precisión
Arte y entretenimiento	8569	7747	822	7355	1214	94.68%
Negocios y economía	8699	8004	695	7546	1153	93.79%
Educación	8742	8083	659	7524	1218	92.80%
Gobierno	8088	7363	725	6685	1403	90.85%
Noticias y medios	11574	10597	977	9650	1924	90.80%
Ciencia y ambiente	8893	8137	756	7496	1397	92.34%
Total	54565	49931	4634	46256	8309	92.47%

Utilizamos la *matriz de confusión* para evaluar y determinar la precisión de nuestro modelo. Esta matriz compara la realidad y la predicción y, basándose en los aciertos y los fallos, calcula un valor de precisión. Por ejemplo, para la categoría "Arte y entretenimiento", el clasificador predijo 822 páginas Web erróneas, pero falló en 32 casos. Hubo 7747 predicciones de páginas Web válidas, pero 424 fueron incorrectas. Estos valores se colocan en la Tabla 3.6 y se sustituyen en la fórmula de precisión.

Tabla 3.6: Matriz de confusión para la categoría de arte y entretenimiento.

		Predicción	
		Error	Válida
Real	Error	790	424
	Válida	32	7323

$$\text{Precisión} = \frac{790+7323}{790+32+424+7323} = 94.68\%$$

El clasificador alcanzó una precisión del 94.68% para esta categoría, lo que es bueno teniendo en cuenta el pequeño número de imágenes implicadas en el proceso de entrenamiento. La Tabla 3.7 muestra la matriz de confusión para cada categoría y una matriz total que indica una precisión del 92.47% para todo el dataset de la técnica de búsqueda.

Después de ejecutar la detección automática de errores de páginas Web, la composición y el tamaño final de nuestro dataset se muestran en la Tabla 3.8. Las imágenes se obtuvieron automáticamente mediante dos técnicas: búsqueda y navegación. En la primera, la fuente es Google y el buscador pregunta por palabras clave relacionadas con la categoría de interés. En la segunda, la fuente es el directorio Web BOTW. Combinando ambas fuentes, hemos recopilado 49438 capturas de pantalla de páginas Web válidas

Tabla 3.7: Matriz de confusión para el resto de categorías y resultado global.

Negocios y economía

		Predicción	
		Error	Válida
Real	Error	654	499
	Válida	41	7505

Educación

		Predicción	
		Error	Válida
Real	Error	624	594
	Válida	35	7489

Gobierno

		Predicción	
		Error	Válida
Real	Error	694	709
	Válida	31	6654

Noticias y medios

		Predicción	
		Error	Válida
Real	Error	918	1006
	Válida	59	9591

Ciencia y ambiente

		Predicción	
		Error	Válida
Real	Error	736	661
	Válida	20	7476

Total

		Predicción	
		Error	Válida
Real	Error	4416	3893
	Válida	218	46038

cercanas a los 17 GB de espacio de almacenamiento. Está disponible públicamente[8] en *Open Science Foundation*[9].

Tabla 3.8: Composición y tamaño del dataset de páginas Web final.

Categoría	**Navegación** Webshots	**Búsqueda** Webshots	**Total**
Arte y entretenimiento	397 (147 MB)	7355 (2.58 GB)	7752
Negocios y economía	892 (300 MB)	7546 (2.48 GB)	8438
Educación	368 (126 MB)	7524 (2.64 GB)	7892
Gobierno	669 (253 MB)	6685 (2.47 GB)	7354
Noticias y medios	394 (237 MB)	9650 (3.19 GB)	10044
Ciencia y ambiente	462 (193 MB)	7496 (2.63 GB)	7958
Total	**3182** (1.22 GB)	**46256** (15.99 GB)	**49438**

3.5 Categorización Web multiclase

En esta sección, demostramos el uso del dataset presentado con un caso práctico de categorización Web en múltiples categorías temáticas. La clasificación de páginas Web, también denominada *categorización Web*, determina si una página o sitio Web pertenece a cierta categoría. Por ejemplo, juzgar si una página trata sobre “arte”, “negocios” o

[8] https://osf.io/7ghd2

[9] Una plataforma libre y abierta para apoyar, difundir y permitir la colaboración de la investigación científica

"deportes" es un caso de clasificación temática [91]. Esta tarea suele realizarse analizando tanto el contenido textual como el código HTML subyacente [50], lo que supone un reto cada vez mayor dado el diseño complejo y dinámico de las páginas Web modernas. El aspecto visual también es una parte importante de una página web, y muchos temas tienen un *look and feel* distintivo, por ejemplo, los blogs de diseño Web tienen un aspecto muy elaborado, mientras que los sitios de periódicos tendrán mucho texto e imágenes [9].

En este trabajo, realizamos una categorización automática de páginas Web según su tema y basada exclusivamente en su apariencia visual. Aprovechamos el dataset generado en este trabajo, formado por páginas Web pertenecientes a 6 categorías: arte y entretenimiento, negocios y economía, educación, gobierno, noticias y medios, y ciencia y ambiente. Por lo tanto, el problema se convierte en una categorización multiclase.

Implementamos un modelo de DL con una CNN. En esencia, se trata de un proceso de aprendizaje con las webshots recopiladas para alcanzar una precisión aceptable y realizar predicciones. Se espera captar características, difíciles de identificar manualmente, que permitan distinguir categorías, predecir a cuál de ellas pertenece una página Web, analizar la dificultad de la clasificación temática de las páginas Web y comprobar si existen patrones particulares para cada categoría.

Los resultados que se presentan a continuación se seleccionaron a partir de una serie de varios experimentos en los que se probaron diferentes modelos y arquitecturas utilizando el dataset completo y partes del mismo. El código desarrollado, así como los pesos del modelo son de acceso público[10]. Los mejores resultados se obtuvieron con la técnica de aprendizaje por transferencia y las imágenes del dataset de navegación. Esto se debe a que en un directorio Web, las páginas pasan por un riguroso proceso de registro bajo la supervisión de especialistas humanos, por lo que tienen una mejor distinción y categorización.

El dataset para la categorización multiclase se compone únicamente de las webshots del conjunto de navegación que se organizó y dividió según la Tabla 3.9. El proceso de entrenamiento dispone de una cantidad equilibrada de datos, es decir, el mismo número de imágenes para cada categoría. La categoría con menos imágenes (educación) fue la base para seleccionar aleatoriamente el mismo número de imágenes en las demás categorías. Se descartó una imagen de 3.68 MB y resolución de 992x30154 píxeles porque la librería de imágenes de Python no abre imágenes de mayor tamaño para evitar ataques maliciosos. Así, cada categoría tiene 367 imágenes, un total de 2202 imágenes, el 80% para el entrenamiento, mientras que el 20% restante se utiliza para la validación (ambos conjuntos seleccionados aleatoriamente).

[10] https://osf.io/8zfh2

Tabla 3.9: Partición del dataset para la categorización Web multiclase.

Categoría	Webshots	Dataset	Entrenamiento (80%)	Validación (20%)
Arte y entretenimiento	397	367	293	74
Negocios y economía	892	367	293	74
Educación	368	367	293	74
Gobierno	669	367	293	74
Noticias y medios	394	367	293	74
Ciencia y ambiente	462	367	293	74
TOTAL	**3182**	**2202**	**1758**	**444**

Las imágenes se almacenan dentro de la estructura de directorios que se muestra en la Figura 3.11. Dentro de la carpeta principal del dataset, se tiene la división en entrenamiento y validación, y las subcarpetas representan las categorías, que tienen los mismos nombres que los temas considerados en este trabajo.

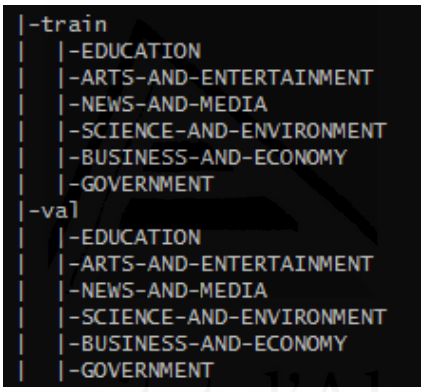


Figura 3.11: Organización del dataset por categorías para la categorización Web.

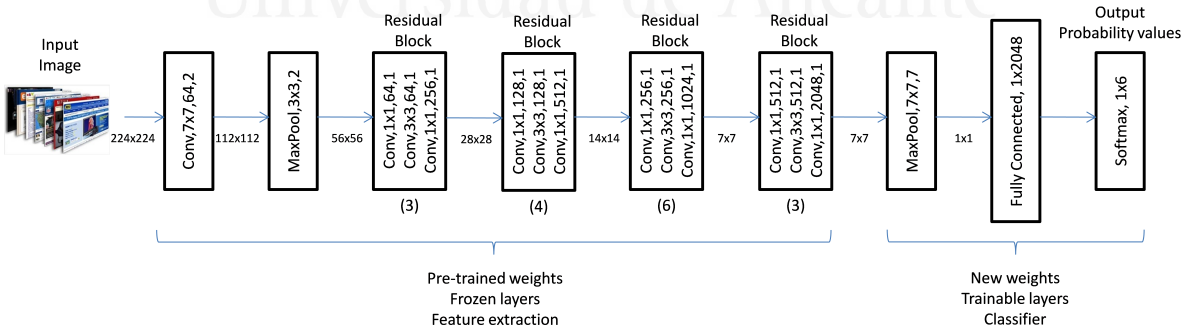


Figura 3.12: Arquitectura del modelo basado en ResNet-50 para la categorización Web.

Después de organizar las imágenes, es aconsejable un paso de preprocesamiento para normalizar los valores enteros entre 0 y 255 de píxel de la imagen a la escala de valores entre 0 y 1. También es necesario cambiar el tamaño a los 224x224 píxeles recomendados

para el modelo, porque las imágenes del dataset tienen dimensiones diferentes de ancho y alto. Ambas son prácticas habituales que ayudan a acelerar el proceso de entrenamiento.

Se probaron varios modelos con diversas opciones para lograr una mayor precisión. El modelo final utiliza *ResNet* [41], que es una CNN competitiva preentrenada en el dataset ImageNet (más de 14 millones de imágenes pertenecientes a 1000 categorías), y que fue la ganadora del reto de ImageNet en 2015. Aunque existen arquitecturas más actualizadas, ResNet sigue siendo muy popular para las implementaciones de aprendizaje por transferencia. Seleccionamos *ResNet-50*, que tiene 50 capas de profundidad y cuya base convolucional se mantiene fija para la extracción de características, mientras que la parte clasificadora se sustituye por una nueva que se encarga de predecir las probabilidades para las 6 clases existentes (Figura 3.12).

Dado que los filtros proporcionados por ImageNet son lo suficientemente genéricos para adecuarse a casi cualquier problema, sólo se entrenan las capas del clasificador en nuestro dataset sin tocar la estructura y los pesos de la parte convolucional. Tras 500 iteraciones de todo el conjunto de entrenamiento de 1758 imágenes en grupos de 32 imágenes, se obtuvo una precisión del 94.26% y del 40.38% en la fase de validación. La evolución del proceso se resume en las siguientes curvas de aprendizaje (Figura 3.13).

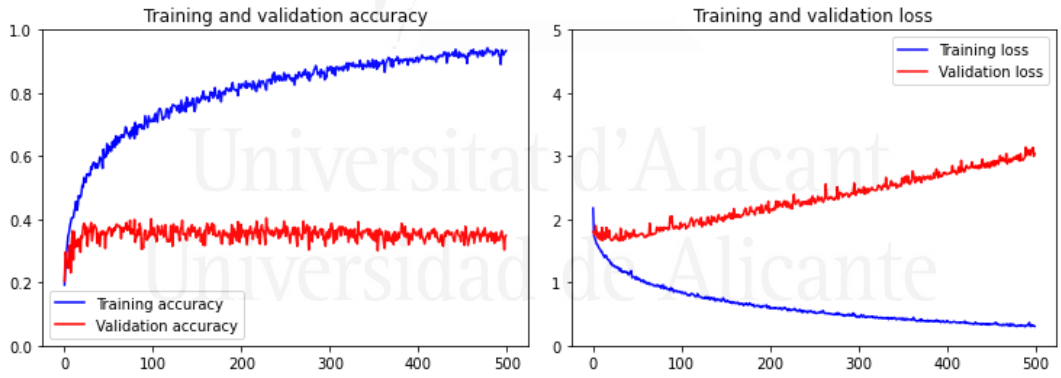


Figura 3.13: Precisión y pérdida en las fases de entrenamiento y validación para la categorización multiclase.

Una solución adecuada a este problema debe cumplir los siguientes requisitos: a) alta precisión de entrenamiento, b) curvas de validación y entrenamiento muy próximas entre sí, y c) pequeña diferencia entre el error de validación y el de entrenamiento. Los gráficos muestran que sólo se cumple el primer elemento, por lo que el modelo aprendió muy bien, pero no para la generalización, es decir, para clasificar aceptablemente nuevas imágenes. Aunque aumentamos los datos, ajustamos los hiperparámetros y aplicamos técnicas de regularización como el dropout, ni se mejora la precisión ni se reduce significativamente

el sobreajuste. Para una mejor comprensión de los resultados, la Tabla 3.10 muestra la matriz de confusión con los datos de validación.

Tabla 3.10: Matriz de confusión para los datos de validación.

		Predicción					
		Arte y entretenimiento	Negocios y economía	Educación	Gobierno	Noticias y medios	Ciencia y ambiente
Real	Arte y entretenimiento	38	6	7	17	3	3
	Negocios y economía	16	20	1	25	6	6
	Educación	11	8	16	27	5	7
	Gobierno	6	4	9	46	5	4
	Noticias y medios	26	2	3	10	31	2
	Ciencia y ambiente	17	7	4	21	6	19

Si nos centramos en las categorías de arte y entretenimiento, gobierno, y noticias y medios, el modelo acierta en la mayoría de los casos, pero el número de aciertos es bajo. Esta prueba toma los datos de validación, un total de 444 imágenes, logrando una precisión del 38.29%, según la matriz de confusión. Para el resto de categorías, el modelo se confunde significativamente. Clasificar páginas Web dentro de estas categorías es un problema de alta dificultad. La composición de las páginas Web actuales es cada vez más compleja y el contenido presenta una gran variabilidad de características visuales, incluso dentro de la misma categoría.

Nuestra recopilación automática de capturas de pantalla generó un dataset con inconvenientes para un proyecto de aprendizaje profundo, como el desequilibrio de datos, la variabilidad intraclase y la similitud interclase (solapamiento). En particular, los resultados de Google están más orientados al contenido que a la apariencia, ya que algunas imágenes recuperadas con la técnica de búsqueda no tienen la apariencia esperada de la categoría respectiva.

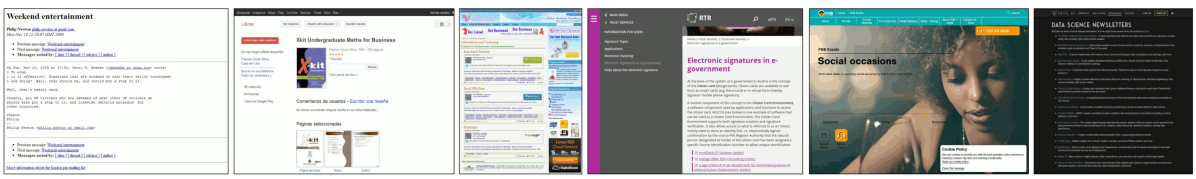


Figura 3.14: Muestra de screenshots con aspecto que no corresponde a la categoría.

Para ilustrarlo, la Figura 3.14 muestra de izquierda a derecha: una página de arte y entretenimiento sin elementos decorativos, una página de negocios y economía sin

imágenes de oficinas ni gráficos estadísticos, una página acerca de educación que no tiene ningún aspecto académico, símbolos nacionales o culturales de un país ausentes en una página gubernamental, una página de noticias sin mucho texto ni fotografías, y ninguna imagen de la naturaleza o descubrimientos en una página de ciencia y ambiente. Si las imágenes dentro de una misma clase difieren mucho y las imágenes de clases diferentes se parecen entre sí, el análisis de datos, el aprendizaje y la predicción de resultados precisos pueden complicarse [88]. A continuación, aplicamos un nuevo enfoque que pretende mejorar el rendimiento en este difícil problema de categorización.

3.6 Categorización Web One vs. Rest

Los resultados de la categorización Web multiclase arrojan una precisión del 94.26% en la fase de entrenamiento y 40.38% en la fase de validación, lo que indica una baja generalización y un sobreajuste significativo. Estos resultados demuestran un problema multiclase muy difícil. En esta sección, abordamos el mismo problema bajo un nuevo enfoque, aplicando la estrategia One vs. Rest (OvR), utilizando técnicas de regularización, redefiniendo el mecanismo de predicción y generando un dataset más fiable. Así, hemos obtenido interesantes mejoras en precisión y reducción del sobreajuste.

3.6.1 Método

Describimos en detalle la creación de un dataset más adecuado, la conversión de una categorización multiclase en múltiples problemas binarios, la regularización añadida al modelo de aprendizaje profundo y el método de predicción redefinido.

3.6.1.1 Refinar el dataset

Disponemos de un amplio dataset de capturas de pantalla de páginas Web de todos los países del mundo organizadas en 6 categorías (Tabla 3.8). Nuestro dataset original probó ser difícil de distinguir por categorías y reduce el rendimiento de un clasificador automático. Para preparar un dataset más conveniente, consideramos las propiedades de representatividad y equilibrio.

Seleccionamos las imágenes más representativas dentro de cada categoría. Este proceso es manual para el conjunto de navegación y automático para el conjunto de búsqueda. El conjunto de navegación es la base porque es más pequeño y la variabilidad dentro de cada categoría es menor, sus páginas Web pasan por una rigurosa supervisión

de especialistas humanos, tanto en contenido como aspecto visual, por lo que las páginas Web tienen una mejor categorización.

Nuestra exhaustiva inspección visual seleccionó 120 imágenes de cada categoría. Este número se debe a que la categoría con menos imágenes (educación) limitaba a las demás. Así, disponemos de un nuevo dataset equilibrado para entrenar un modelo de clasificación y determinar automáticamente las imágenes más representativas del conjunto mayor (búsqueda).

El modelo emite un valor entre 0 y 1 para cada imagen, que es interpretado como la probabilidad de que la imagen pertenezca a la categoría. Para disponer de un conjunto equilibrado, se seleccionaron automáticamente 1000 imágenes con la mayor probabilidad para cada categoría. Este número se debe a que la categoría con menor número de imágenes restringe a las demás. Aunque el tamaño del dataset final es reducido, hemos obtenido un conjunto de datos más fiable, cuya estructura se muestra en la Tabla 3.11.

Tabla 3.11: Composición del nuevo dataset equilibrado.

Categoría	**Navegación**	**Búsqueda**	**Total**
Arte y entretenimiento	120	1000	1120
Negocios y economía	120	1000	1120
Educación	120	1000	1120
Gobierno	120	1000	1120
Noticias y medios	120	1000	1120
Ciencia y ambiente	120	1000	1120
Total	**720**	**6000**	**6720**

3.6.1.2 Problema multiclase en subproblemas binarios

La clasificación de páginas Web con 6 categorías ha resultado ser una tarea muy difícil. Los resultados mostraron una baja precisión y un alto sobreajuste, por lo que intentamos reducir la complejidad descomponiendo el problema multiclase en subproblemas binarios. En el enfoque OvR, si hay K clases, se requieren K clasificadores binarios diferentes, cada uno diseñado para discriminar ejemplos de una clase dada en relación con todas las demás clases [94].

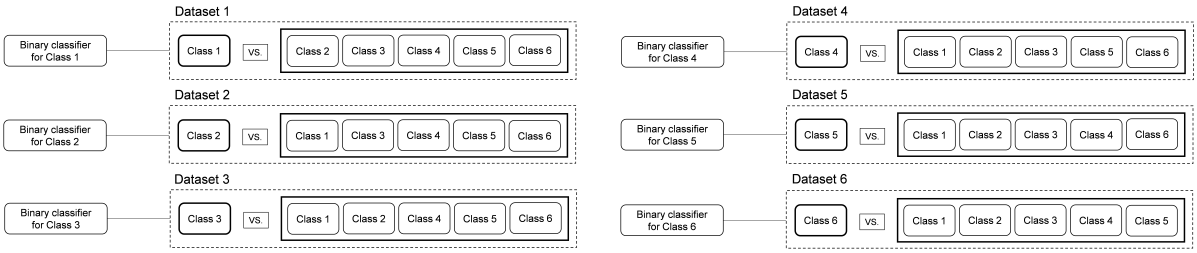


Figura 3.15: Clasificadores binarios OvR y sus respectivos datasets.

Los 6 clasificadores binarios resultantes se muestran en la Figura 3.15. Cada clasificador está diseñado para distinguir cada categoría, por lo que esperamos una mejor distinción y reconocimiento de características. El inconveniente es reorganizar el dataset para cada clasificación, por ejemplo, para el clasificador binario de la categoría "Arte y entretenimiento", las imágenes de esta categoría se almacenan en una carpeta, y el resto de las imágenes de todo el dataset en otra carpeta. Del mismo modo, debemos organizar el respectivo dataset para los demás clasificadores, generando 1120 imágenes de una categoría frente a 5600 del resto. Para resolver este desequilibrio, 1120 de las 5600 imágenes se seleccionan aleatoriamente, evitando posibles sesgos durante el entrenamiento del modelo.

El aprendizaje automático requiere dividir el dataset en dos subconjuntos independientes: entrenamiento y prueba. El subconjunto de entrenamiento se utiliza para aprender y ajustar los parámetros del modelo, mientras que el subconjunto de prueba se utiliza para evaluar la precisión. Un tercer subconjunto de validación es necesario para optimizar el modelo durante el entrenamiento. En este caso, los conjuntos de prueba y validación serán los mismos debido al limitado número de imágenes disponibles. La división del dataset para el primer clasificador se muestra en la Tabla 3.12. La misma estructura se aplica para el resto de clasificadores binarios.

Tabla 3.12: Partición del dataset para el clasificador binario de la categoría de arte y entretenimiento.

Categoría	Webshots	Entrenamiento (75%)	Validación (25%)
Arte y entretenimiento	1120	840	280
Resto de categorías	1120	840	280
TOTAL	**2240**	**1680**	**560**

Es necesario preparar las imágenes en un formato de entrada adecuado para el proceso de entrenamiento mediante la normalización de los valores de los píxeles (enteros entre 0 y 255) a la escala de valores entre 0 y 1, y redimensionando a los 224x224 píxeles

recomendados para el modelo. Esto ayuda a acelerar el proceso de entrenamiento y a evitar la saturación de la memoria de la computadora.

3.6.1.3 Regularización del modelo

Utilizamos la misma arquitectura para todos los clasificadores binarios y sólo cambia el dataset en el que se entrena cada clasificador. Aprovechamos el aprendizaje por transferencia ya que las páginas Web suelen incluir fotografías de objetos del mundo real, por lo que el modelo preentrenado de *EfficientNetB0* [103] se adapta a nuestro problema, pero considerando sólo dos categorías.

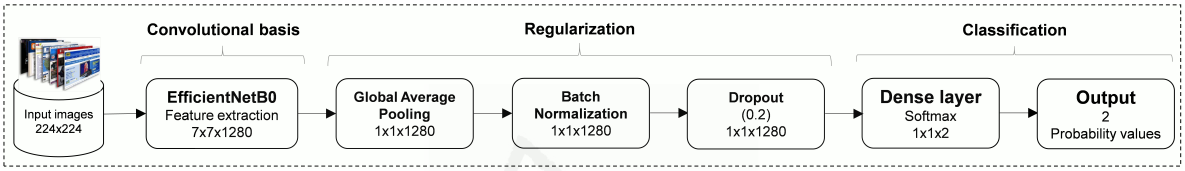


Figura 3.16: Arquitectura del modelo basada en EfficientNetB0.

La arquitectura final se representa en la Figura 3.16, y se compone de:

- Una base convolucional, original de EfficientNetB0 y útil para extraer características. Los pesos preentrenados componen los filtros para identificar características en la imagen de entrada de 224x224 píxeles. Como resultado, se generan 1280 mapas de características de 7x7.
- Capas de regularización, para minimizar el posible sobreajuste. Añadimos:
  - *Global average pooling*, para aplanar y reducir el número de parámetros. Cada mapa de características está representado por el valor medio.
  - *Batch normalization*, estabiliza el modelo haciendo que el conjunto de 1280 valores no esté disperso, con una media de 0 y una desviación estándar de 1.
  - *Dropout*, intenta evitar el sobreajuste omitiendo aleatoriamente una parte de los valores (en este caso el 20%) utilizados para la predicción final.
- El clasificador EfficientNetB0 original para 1000 clases se sustituye por uno nuevo para 2 clases. Nuestros experimentos demostraron que no es necesaria una capa totalmente conectada, sólo utilizamos una capa densa con una función de activación softmax para convertir las salidas en probabilidades de clase que suman uno.

3.6.2 Resultados experimentales

La parte experimental consiste en un total de 6 entrenamientos, uno por cada clasificador binario. En cada entrenamiento, se utiliza la misma arquitectura e hiperparámetros, y sólo cambia el dataset equilibrado de dos clases.

3.6.2.1 Entrenamiento

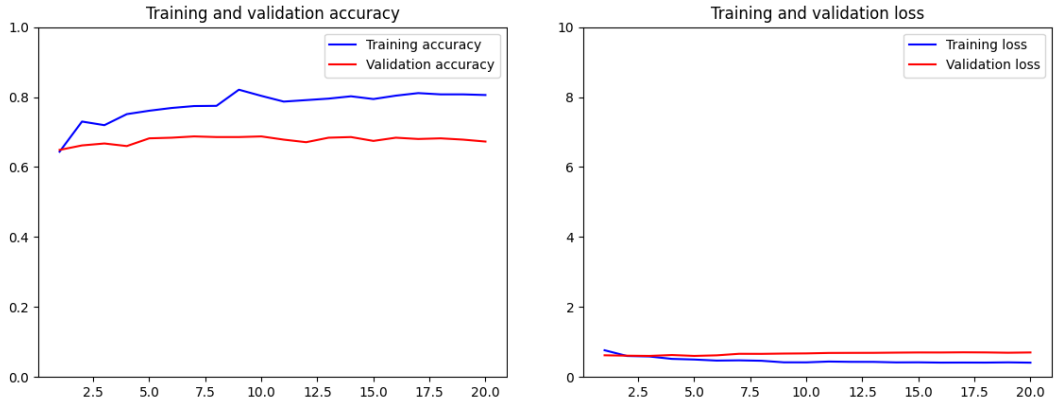
Primeramente, definimos los *hiperparámetros*, que son parámetros que una red neuronal no puede aprender: la función de pérdida (*entropía cruzada categórica*) minimizada por el método de optimización (*adam*) y las métricas recogidas (*precisión* y *pérdida*). Para ajustar el modelo, configuramos el número de imágenes a entrenar en cada lote ($batch\_size = 32$), así como la tasa de aprendizaje ($1e{-}4$) que especifica cuánto se actualizarán los pesos cada vez. En el aprendizaje por transferencia, se recomienda un valor bajo para no cambiar demasiado lo ya aprendido. También, el número de iteraciones del conjunto de imágenes de entrenamiento para actualizar los pesos ($epochs = 20$).

El proceso de entrenamiento consiste en pasar las imágenes por lotes al modelo para detectar las características, que se convierten en la entrada para que el clasificador etiquete cada imagen con una categoría. La precisión y la pérdida se calculan para cada lote, y los pesos se pueden ajustar con *backpropagation* y descenso del gradiente, excepto para la base convolucional. Una vez completados todos los lotes (1 época), se ejecuta la validación obteniendo los valores de precisión y pérdida. Este proceso se realiza de forma iterativa en función del número de épocas. En cada entrenamiento, se realizaron 20 épocas con un dataset de 1680 imágenes en lotes de 32 imágenes, y un total de 560 imágenes para la validación.

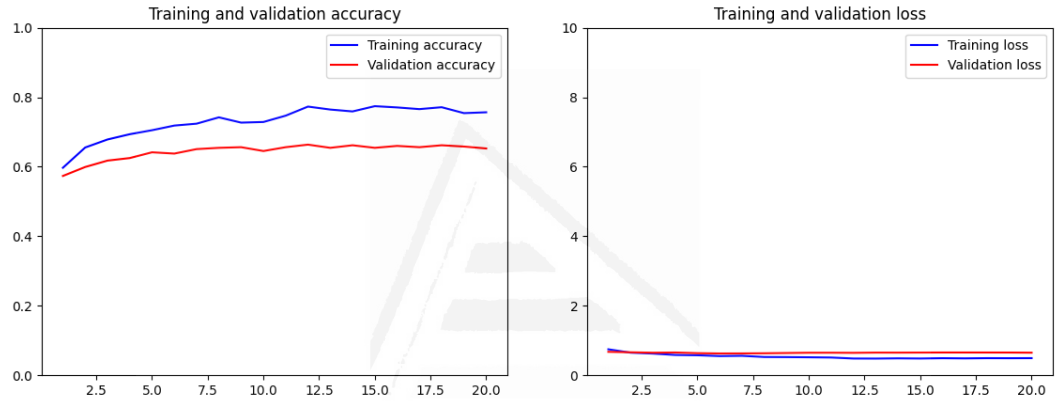
3.6.2.2 Evaluación

Una vez efectuado el entrenamiento, analizamos el rendimiento de cada clasificador binario. Tras cada época, se generaron valores de precisión (*acc*) y error (*loss*) en la fase de entrenamiento, así como valores de precisión (*val_acc*) y error (*val_loss*) en la fase de validación. Estas métricas son registradas a lo largo del proceso y se presentan en dos gráficos (entrenamiento y validación) para cada clasificador binario (Figura 3.17).

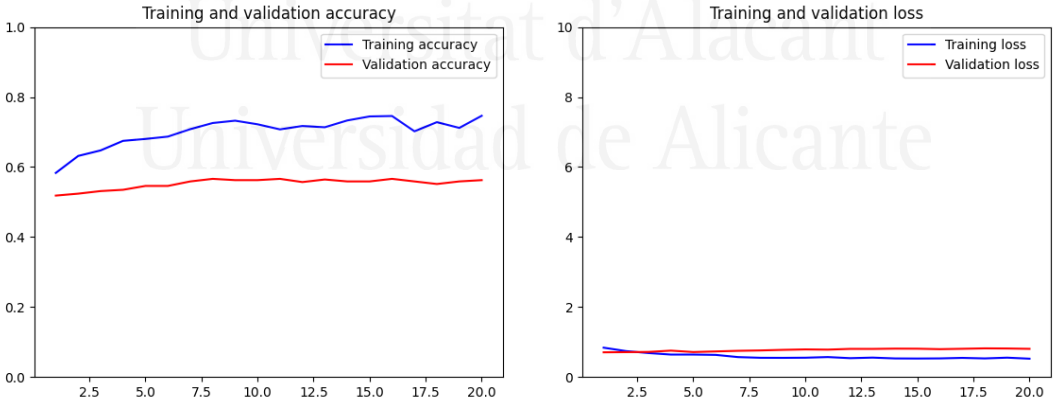
Excluyendo las categorías de educación y gobierno, las cuales aprendieron bien, pero no para clasificar aceptablemente nuevas capturas de pantalla, las demás categorías tienen un buen nivel de predicción, debido a su alta precisión de entrenamiento, curvas de validación y entrenamiento muy cercanas, tanto para la precisión como para la pérdida, y curvas de pérdida cercanas al eje horizontal, es decir, tendiendo a 0.



(a) Arte y entretenimiento vs. resto

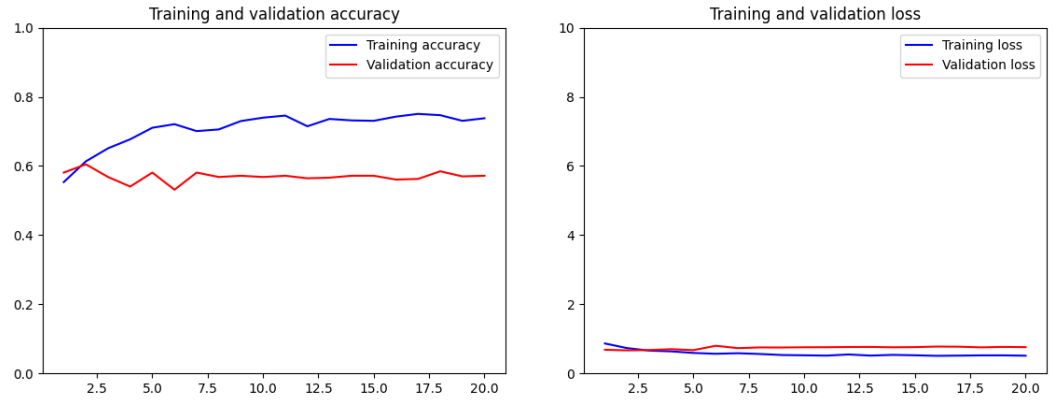


(b) Negocios y economía vs. resto

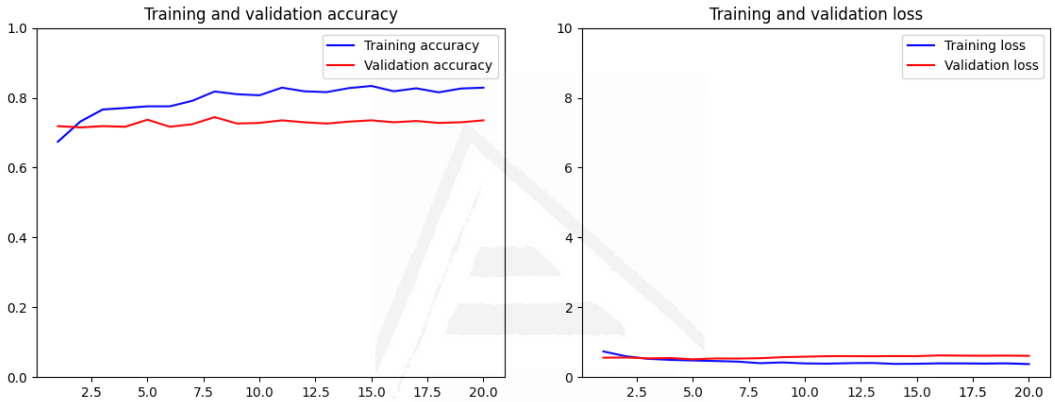


(c) Educación vs. resto

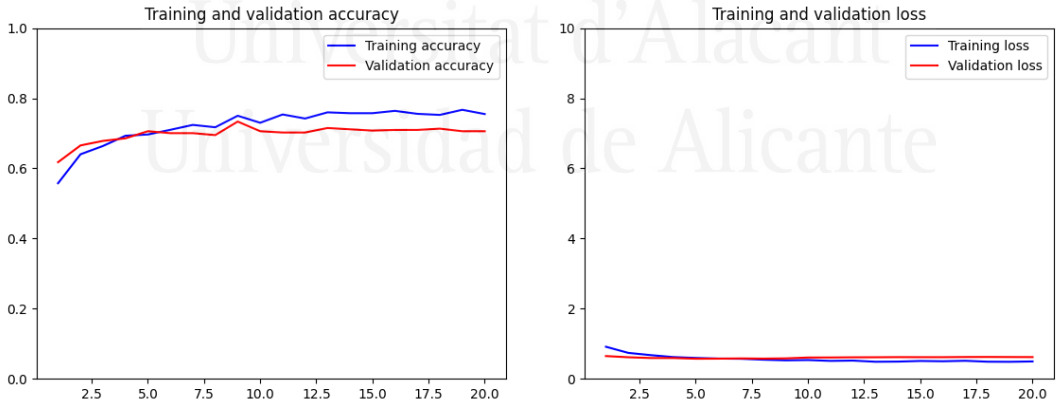
Todos los clasificadores alcanzan un valor máximo de precisión bastante bueno durante el entrenamiento (Tabla 3.13), pero la precisión de validación es más útil porque es una estimación de cómo responderá el modelo al evaluar imágenes no vistas. La diferencia entre estos indicadores es corta, excepto en las categorías de educación y gobierno. Para evaluar lo bien que nuestro clasificador identifica la categoría, utilizamos la precisión



(d) Gobierno vs. resto



(e) Noticias y medios vs. resto



(f) Ciencia y ambiente vs. resto

Figura 3.17: Precisión y pérdida en las fases de entrenamiento y validación para cada clasificador binario.

basada en verdaderos positivos (VP) y falsos positivos (FP). La categoría de noticias y medios destaca con 234 de 280 imágenes (83.57%) correctamente predichas, mientras que

Tabla 3.13: Precisión de entrenamiento y validación para cada clasificador binario.

Clasificador	Entrenamiento	Validación	VP	FP	Precisión
Arte y entretenimiento vs. resto	82.24%	68.75%	182	98	65.00%
Negocios y economía vs. resto	77.01%	66.36%	193	87	68.93%
Educación vs. resto	73.51%	56.62%	138	142	49.29%
Gobierno vs. resto	74.60%	60.48%	112	168	40.00%
Noticias y medios vs. resto	84.11%	74.45%	234	46	83.57%
Ciencia y ambiente vs. resto	78.00%	73.35%	181	99	64.64%

el peor caso es la categoría de gobierno, con 112 imágenes correctamente clasificadas, es decir, un 40%.

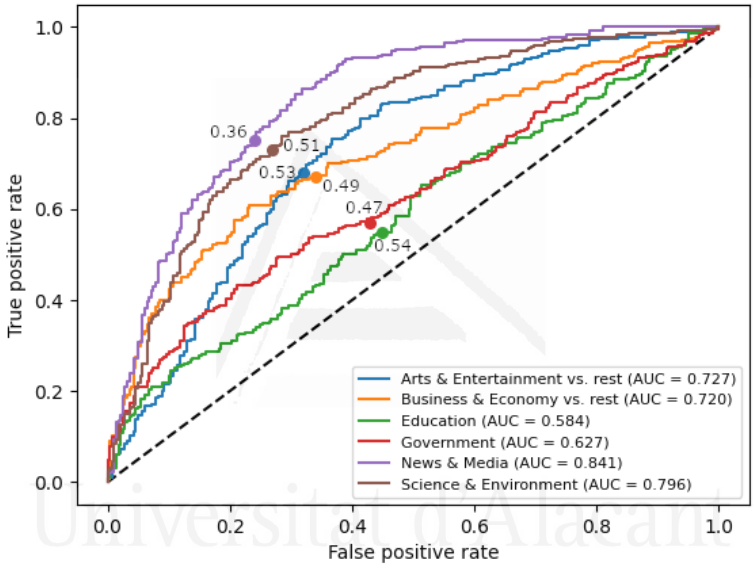


Figura 3.18: Curva ROC para todos los clasificadores binarios.

Una curva *ROC* es adecuada para visualizar el rendimiento de un clasificador y compararlo con los demás. Representa la tasa de verdaderos positivos (*sensibilidad*) y la tasa de falsos positivos (1-*especificidad*) en varios valores de umbral (Figura 3.18). La capacidad predictiva de los clasificadores es aceptable, excepto para las categorías de educación y gobierno, cuyas curvas están muy próximas a la línea diagonal. La métrica *AUC* (área bajo la curva ROC) cuantifica dicha capacidad, por lo que la categoría de noticias y medios (0.841) es el mejor clasificador. En nuestro caso, la sensibilidad es más importante que la especificidad, ya que nos centramos en detectar verdaderos positivos en la medida de lo posible. Encontramos el umbral óptimo en el punto más cercano a la esquina superior izquierda (0,1) para cada clasificador. Estos valores difieren muy poco del valor por defecto (0.5), excepto para la categoría de noticias y medios, en la que deberíamos reducirlo a 0.36.

Tabla 3.14: Comparación de los resultados de categorización multiclase y binaria.

	Categorización multiclase	Categorización por clasificadores binarios (promedio)
Precisión de entrenamiento	94.26%	78.25%
Precisión de validación	40.38%	66.67%
Precisión por categoría	38.29%	61.91%
Sobreajuste	53.88%	11.58%

La Tabla 3.14 resume y contrasta los resultados obtenidos en la categorización de páginas Web utilizando el enfoque multiclase y binario. Excepto en la fase de entrenamiento, los clasificadores binarios que operan por separado alcanzaron mejor precisión y disminuyeron el sobreajuste, mostrando importantes mejoras respecto al clasificador multiclase que opera con todas las categorías juntas.

3.7 Conclusiones

La categorización de páginas Web basada en webshots utilizando un modelo de CNN multiclase resultó un problema muy complejo. La dificultad aumenta cuando las categorías abarcan una amplia gama de temas. Nuestra investigación ha hecho un primer avance considerando un mayor número de categorías que los trabajos relacionados y utilizando capturas de pantalla completas de las páginas Web. Los resultados establecen una base que puede ser mejorada, lo que requiere mucho más trabajo, especialmente para abordar la gran variabilidad en el aspecto visual de las páginas Web dentro de cada tema. Sin embargo, contribuimos en los siguientes aspectos.

3.7.1 Dataset Web

Hemos creado un dataset mixto sobre páginas Web que combina distintos tipos de datos: texto, números e imágenes. Automatizamos el flujo de trabajo con scripts en Python y R para recopilar URLs y sus respectivas webshots, mientras que el scraping nos permitió extraer atributos de cada página Web.

Aunque pudimos recopilar automáticamente un total de 58174 webshots, el dataset final se redujo a 49438, debido a la eliminación de páginas Web de error. Implementamos una detección automática de páginas Web de error basada en un modelo CNN personalizado logrando una precisión aceptable. Esta herramienta de depuración puede contribuir para hacer frente a la importante presencia de páginas Web no válidas en Internet, que afecta a los webmasters, motores de búsqueda y usuarios en general.

Podría ser útil organizar el dataset con categorías más distintivas, aumentar el dataset con más URLs, webshots y otros atributos, así como el preprocesamiento de los webshots al mismo tamaño y resolución. Además, la exploración de otras fuentes de URLs como un motor de búsqueda distinto de Google y un directorio Web distinto de BOTW.

3.7.2 Categorización Web

Buscamos la mejora de la precisión conseguida en la categorización Web multiclase utilizando el esquema One vs. Rest, la regularización del modelo y el refinamiento del dataset. Los clasificadores binarios trabajando por separado alcanzaron mejor rendimiento con respecto al clasificador multiclase tradicional, lo que en promedio mejora la precisión y reduce el sobreajuste.

La notable precisión del modelo para ciertas categorías, nos permite inferir la existencia de patrones visuales distintivos, que pueden servir de base para futuras investigaciones. Sin embargo, proponemos la necesidad de una mejor definición de las categorías, ya que algunas de ellas siguen los mismos patrones en el diseño, por lo que existe confusión entre estas categorías. Por ejemplo, las páginas Web de ciencia muestran su contenido en forma similar a sitios de noticias. Los temas de ciencia y ambiente se incluyen a menudo en páginas relacionadas con la educación, y algunas instituciones ambientales son parte de páginas Web gubernamentales.

Se recomiendan los modelos CNN más profundos y recientes, realizar una clasificación jerárquica incluso en forma de grafo, o una combinación de la clasificación basada en el contenido, el código HTML y las capturas de pantalla puede conducir a una categorización Web más eficaz y fiable.

Capítulo 4

Reconocimiento de emociones

Este capítulo reúne el contenido de nuestros artículos sobre un novedoso método data-centric [75] y la combinación de datasets de FER [77] para abordar el problema del reconocimiento de emociones mediante DL utilizando imágenes faciales. La Sección 4.1 destaca la importancia de automatizar esta tarea enmarcada en el campo de la visión computacional, las problemáticas que caracterizan a los datasets de imágenes faciales, y la estrategia que proponemos combinando varios datasets para entrenar un modelo de mayor generalización. La Sección 4.2 describe en detalle los datasets considerados con énfasis en sus desventajas. La Sección 4.3 explica el método y los recursos utilizados para el refinamiento de los datasets y la creación de los datasets combinado, evaluador y artificial. La Sección 4.4 es para la parte experimental, que emplea estos datasets para el entrenamiento y la evaluación single- y cross-dataset para conocer la calidad y la generalización de los modelos de DL y los diferentes datasets en el reconocimiento de emociones.

4.1 Introducción

En los últimos años, la IA ha alcanzado logros que son pasos significativos hacia el objetivo de imitar la inteligencia humana, la cual integra capacidades cognitivas, fisiológicas y emocionales [27][72]. En el caso cognitivo y fisiológico, algunas de nuestras capacidades han sido igualadas e incluso superadas, por ejemplo, por robots más fuertes y rápidos, sistemas de traducción, juegos de estrategia y análisis de datos [31][56]. Sin embargo, aún existen muchos retos y limitaciones en el ámbito emocional, debido a la complejidad, naturaleza polifacética y dificultad de medir y modelizar las emociones humanas.

Las emociones son esenciales y aún distintivas de las personas, influyen en nuestro comportamiento y desempeñan un rol fundamental en la comunicación e interacción con el entorno familiar, laboral y el resto de la sociedad [47]. Además de las personas de nuestro entorno, cada vez es más común encontrarnos rodeados de máquinas tratando de imitar el comportamiento humano, por lo que es necesario interactuar. En un futuro próximo, la interacción hombre-máquina (HMI) será una práctica habitual y se pretende que esa interacción sea lo más natural posible [15][26][100].

La habilidad esencial para que la IA pueda adaptar su comportamiento y proporcionar respuestas adecuadas durante la interacción con el usuario es el reconocimiento de las emociones (Emotion Recognition, ER), que ha surgido como un tema de investigación activo. Aunque la robótica social y afectiva es la aplicación predominante en la literatura relacionada, el desarrollo de sistemas de IA capaces de reconocer emociones tiene prometedoras aplicaciones en áreas como la inteligencia emocional, medicina, sanidad, psicología, sociología, psiquiatría, seguridad pública, seguridad vial, videovigilancia, marketing y ventas, educación, artes y entretenimiento [3][30][53][69][73][108]. Un detalle más amplio de estas aplicaciones prácticas del reconocimiento de emociones se presenta en el Apéndice B.

Comprender cómo se expresan y comunican las emociones es la clave para reconocerlas. Se trata de un mecanismo complejo y multimodal en el que intervienen componentes verbales, no verbales y corporales. Un importante estudio [74] cuantificó el grado de influencia de los elementos que intervienen en la comunicación de las emociones, determinando la parte no verbal (gestos faciales y corporales) como la más influyente con un 55%, mientras que el tono de voz con un 38%, y sólo un 7% para el lenguaje verbal. En un contexto conversacional, la manifestación exclusivamente verbal de enfado o alegría debe ir acompañada de un gesto facial para transmitir credibilidad y convicción del interlocutor. Incluso el gesto bastaría para describir la emoción que estamos experimentando, ya que a menudo prestamos más atención al rostro que a las palabras. Nuestros gestos faciales hablan más que mil palabras, y la cara es la ventana del alma dice una frase muy conocida [19]. La reciente pandemia ha demostrado que cuando hay una máscara facial, la capacidad humana de inferir emociones se reduce [35]. Por lo tanto, la expresión facial es la principal forma de comunicar e identificar las emociones humanas [6][14][25][48].

Los seres humanos captan los cambios emocionales mediante la observación de las expresiones faciales. Está demostrado que los patrones de expresión facial asociados a las emociones son universales para todas las personas, en todas las culturas y contextos [21]. Es muy probable que esto se deba a una razón biológica [100][109]. Los movimientos musculares del rostro humano y la forma geométrica de sus elementos, como ojos, cejas,

nariz, boca, labios, pómulos y barbilla, se denominan *action units* (AU) y están definidas en el FACS (Facial Action Coding System) [1][22]. Algunas de las características son la dirección de la mirada, la posición de los pómulos, la apertura de la boca, la aparición de arrugas en la piel del rostro, etc. La colección de UAs que tienden a actuar juntas puede sugerir o revelar un cierto tipo de emoción [19][96][112]. En consecuencia, se puede inferir la emoción si se reconoce la expresión facial.

Del mismo modo que las personas pueden deducir el estado emocional de otras a partir de las expresiones faciales, la visión por computadora aborda el problema de automatizar el reconocimiento de expresiones faciales (Face Expression Recognition, FER) y la interpretación de emociones humanas [15]. Es una tarea crítica y difícil para las computadoras, la cual ha atraído la atención de la comunidad investigadora [51], incluso motivando competiciones como la organizada en la plataforma Kaggle [13]. FER se ha convertido en un área importante de investigación y desarrollo.

Un enfoque popular para reconocer emociones está basado en FER a partir de imágenes faciales estáticas y vídeos [3][30][37]. Para esta tarea se prefieren las técnicas de DL por su capacidad de extraer automáticamente características mediante aprendizaje supervisado, evitando el elevado coste de tiempo y esfuerzo que supone definir manualmente múltiples y complejas características de las expresiones faciales [54][79]. En particular, las CNNs han mostrado resultados prometedores en diferentes datasets de imágenes faciales. Una vez identificadas estas características, se clasifica la expresión facial y se asocia con la emoción, ya sea en términos dimensionales, por ejemplo, valencia e intensidad (lo positiva o negativa que es una emoción y la fuerza de la emoción, respectivamente), o en términos categóricos, normalmente una de las siete emociones universales básicas: feliz, sorpresa, enfado, triste, miedo, asco y neutral [20][21][62][81].

Aunque se traten diversos modelos de detección de emociones, si el dataset no es de buena calidad, no se mejorarán los resultados de precisión en aplicaciones prácticas. Nuestro trabajo se centra en los datasets, una colección de imágenes de rostros humanos captadas en cualquier entorno o también extraídas de vídeos. A nivel dimensional, hay muy pocos datasets faciales etiquetados [81], mientras que para un modelo categórico, se puede encontrar una gama más amplia de datasets disponibles.

Sea el modelo dimensional o el categórico, los datasets de imágenes faciales se recopilan principalmente de dos formas [79][109]:

***in-the-lab*:** en condiciones controladas, sin cambios y con actores que exageran los gestos de cada emoción, lo que permite un alto índice de precisión y el problema de FER se considera resuelto, pero de escasa utilidad para aplicaciones del mundo real.

***in-the-wild*:** son normalmente datasets recogidos de Internet con fines de mayor generalización, cuyas imágenes incluyen personas reales en situaciones del mundo real. Aunque existen varios datasets in-the-wild para entrenar modelos de reconocimiento de emociones, es muy difícil obtener la misma precisión de los entornos controlados en entornos no controlados.

Los datasets in-the-wild presentan problemas relacionados con la falta de calidad y generalidad. En cuanto a la calidad, la recopilación automática de imágenes desde Internet y la subjetividad del crowdsourcing son los principales factores de la presencia de imágenes irrelevantes, el etiquetado erróneo por emociones similares y difíciles de diferenciar, y un desequilibrio de clases de emoción.

En términos de generalidad, la dificultad de abarcar todos los tipos de rostros y expresiones faciales da lugar a un sesgo de las características de la población en general. Cada dataset puede contener más imágenes de personas de un determinado género, edad o etnia, lo que afecta a la capacidad de reconocer emociones raras o ausentes en el dataset de entrenamiento. Este sesgo afecta al estado del arte en reconocimiento de emociones, y el modelo con mayor precisión en un dataset determinado puede no ser siempre el mejor para reconocer emociones en el mundo real. Además, cada dataset tiene sus propias especificaciones técnicas, como el número de imágenes, el color, la resolución, el fondo, la iluminación y el tipo de archivo. Como resultado, los modelos dependen en gran medida del dataset de entrenamiento concreto, lo que limita su adaptabilidad a escenarios del mundo real en los que es necesario reconocer expresiones faciales a partir de imágenes captadas en entornos no controlados.

Un extenso dataset de alta calidad que represente la compleja heterogeneidad de las expresiones emocionales y a toda la población de forma completa y equilibrada sería el recurso ideal para entrenar un modelo de reconocimiento de buen rendimiento en aplicaciones del mundo real. Sin embargo, crear un dataset con estos requisitos es un problema enorme [62]. El objetivo de nuestro trabajo es proporcionar un dataset más amplio, diverso y general que los existentes, mediante la combinación de varios datasets conocidos del ámbito de FER. Un modelo entrenado en este dataset generalizaría más que uno específico.

La generalización de un modelo de DL es una capacidad que debe ser cuantificada. Esta métrica se convierte en un punto de referencia para un estado del arte general de rendimiento para la tarea de reconocimiento de emociones. Hasta ahora, se dispone únicamente de un rendimiento específico en el dataset de entrenamiento particular. Nuestra investigación contribuye con una primera aproximación a esta problemática mediante la generación de un *dataset evaluador*.

Para ello, el dataset combinado se divide en dos partes, una de las cuales se emplea para entrenar un modelo más genérico, y la otra, se convertirá en nuestro dataset evaluador, al cual se añaden las propiedades de equilibrio e imparcialidad mediante ingeniería de datos con técnicas de DL. Cuando el modelo se prueba con este dataset evaluador, es posible medir su capacidad de generalización, así como la de otros modelos entrenados en diferentes datasets. De este modo, proporcionamos una métrica más robusta para aplicaciones prácticas.

En concreto, planteamos las siguientes preguntas de investigación: ¿Se espera un mejor rendimiento y una mayor precisión en el reconocimiento de emociones si se mejora la calidad del dataset?, ¿Puede la combinación de varios datasets de FER mejorar la generalización de un modelo de reconocimiento de emociones?, ¿Puede un dataset combinado, equilibrado e imparcial evaluar y comparar diferentes modelos de reconocimiento de emociones, proporcionar una métrica de rendimiento más cercana al estado del arte general y útil para seleccionar el modelo más adecuado para aplicaciones del mundo real?. Una respuesta positiva a estas preguntas no resolverá por completo el problema de la falta de calidad y generalización de los datasets de FER, pero puede ayudar a establecer una base sólida para futuras investigaciones y orientar hacia una solución a largo plazo. La necesidad de hacer una contribución en este campo, con el propósito de acercar algo más la IA a la inteligencia humana en su conjunto, es la motivación detrás de este trabajo de tesis.

4.2 Datasets de FER

A nivel de emociones de tipo dimensional (continuo), existen muy pocos datasets faciales etiquetados [81], mientras que para un modelo categórico (discreto), se pueden encontrar múltiples datasets de imágenes creados para el reconocimiento automático de emociones basado en expresiones faciales. Hemos considerado FER2013 [33], AffectNet [81] y NHFI (Natural Human Face Images) [105]. Los dos primeros son muy populares y se han investigado ampliamente en el campo de FER, mientras que NHFI es menos conocido, pero más reciente y diseñado para proporcionar imágenes con una mejor anotación manual. Factores como la disponibilidad, el tamaño, el formato de imagen y las categorías de emoción hacen que estos datasets sean apropiados para nuestro trabajo. A continuación, se describen detalladamente sus características, el modo de obtenerlos y las desventajas que presentan.

4.2.1 Características

El dataset FER2013 (creado por Pierre-Luc Carrier y Aaron Courville) y AffectNet (Ali Mollahosseini, Behzad Hasani y Mohammad H. Mahoor) son estándares tomados como referencia para las competiciones [33], mientras que NHFI (Sudarshan Vaidya) es un dataset novedoso, creado con el propósito de proporcionar más datos con mejor anotación manual, lo que nos proponemos analizar en el presente estudio. La Tabla 4.1 resume las características principales de estos datasets.

Tabla 4.1: Datasets FER considerados y sus principales características.

Característica	FER2013	NHFI	AffectNet
Cantidad de imágenes	35886	5558	287401
Modelo de emoción	Discreto	Discreto	Discreto/continuo
Categorías	7	8	8
Tipo de imágenes	Imagen facial 2D	Imagen facial 2D	Imagen facial 2D
Etiquetado	Humano	Humano	Automático y humano
Equilibrado	No	No	No
Resolución (pixeles)	48x48	224x224	224x224
Color	Grayscale	Grayscale	color RGB
Formato	JPG	PNG	JPG
Peso	300 MB	50 MB	4 GB
Disponibilidad	Público	Público	Solicitud
Fuente de datos	Internet	Internet	Internet
Tamaño	Mediano	Pequeño	Grande
Entorno	In-the-wild	In-the-wild	In-the-wild
Año	2013	2020	2017
Estructura	Archivo CSV	Carpetas y archivos	Archivos de imagen y NumPy
División (%)	Entrenamiento/prueba (80/20)	Ninguna	Entrenamiento/validación (99/1)

La calidad de los datasets se ve más afectada a medida que aumenta su tamaño, por lo que seleccionamos datasets a diferentes escalas: pequeña (miles de imágenes), media (decenas de miles) y gran escala (cientos de miles). Las imágenes faciales incluidas son estáticas, no secuencias de vídeo, con una apariencia 2D o plana, en contraste con las imágenes 3D que generan una percepción de profundidad [48]. A cada imagen se le asigna una categoría de emoción, tarea realizada íntegramente por humanos, excepto en el caso de AffectNet, donde una parte se anotó manualmente y el resto automáticamente mediante una red neuronal entrenada sobre todas las muestras del conjunto de entrenamiento anotadas manualmente [81]. Los datasets no están equilibrados, es decir, no tienen el mismo número de imágenes para cada categoría, o al menos un número similar. Para examinar la influencia del color y el tamaño de la imagen en el reconocimiento, disponemos de imágenes en escala de grises y en modo RGB, así como en resolución pequeña y mediana. Con respecto a JPG y PNG, son formatos de imagen estándar fáciles de convertir entre sí. Los datasets escogidos abarcan condiciones in-the-wild, con

imágenes alejadas de un entorno controlado, más cercanas a la realidad, diferentes niveles de iluminación, edades, poses, intensidad de expresión, oclusiones, lo que convierte el reconocimiento en una tarea difícil [90].

4.2.2 Adquisición

Los datasets FER2013[1] y NHFI[2] están disponibles públicamente en Kaggle, mientras que AffectNet requiere permiso para su uso a través de un formulario de solicitud a los autores[3]. FER2013 se puede obtener en un formato de valores separados por comas (CSV) cuyas columnas representan los siguientes atributos:

- Un valor entre 0 y 6 para cada una de las 7 emociones posibles (0: enfado, 1: asco, 2: miedo, 3: feliz, 4: neutro, 5: triste, 6: sorpresa).
- Una lista de 2304 valores enteros, cada uno equivalente a un píxel de la imagen de 48x48 y el subconjunto al que pertenece (entrenamiento o prueba).

Debido a que las imágenes no son directamente visibles, utilizamos un script de Python con las librerías Pandas y NumPy para leer el archivo, almacenar los valores enteros como matrices de píxeles y convertirlos en archivos de imagen estándar. Se obtienen un total de 35886 imágenes después de transformar las matrices de píxeles a archivos de imagen en formato JPG, en escala de grises y con una resolución de 48x48 píxeles, divididas en subconjuntos de entrenamiento y prueba, 28708 y 7178 imágenes, respectivamente. Cada subconjunto incluye 7 carpetas, cada una para un tipo concreto de emoción. La descarga de NHFI es un archivo comprimido, que genera 8 carpetas, cuyos nombres son prácticamente los mismos que los del dataset anterior, sólo se excluye la categoría "desprecio" para que la comparación sea justa. Dentro de cada carpeta, hay imágenes en formato PNG. En el caso de AffectNet, el enlace proporcionado en respuesta a la solicitud permite descargar dos archivos comprimidos para el entrenamiento y la validación. Tras la extracción de cada archivo, se crea una carpeta denominada "images" que contiene los archivos JPG y otra llamada "annotations" que contiene los archivos NPY de las etiquetas correspondientes. Desarrollamos un script en Python[4] para leer la categoría de emoción del archivo NPY y mover el archivo JPG a la carpeta correspondiente. Cabe mencionar que AffectNet tiene dos versiones, utilizamos la que contiene sólo las imágenes anotadas manualmente con 8 etiquetas (pero se omite la

[1] www.kaggle.com/datasets/deadskull7/fer2013
[2] www.kaggle.com/datasets/sudarshanvaidya/random-images-for-face-emotion-recognition
[3] mohammadmahoor.com/affectnet-request-form/
[4] github.com/cimejia/FER-datasets/blob/main/createAffecnet.py

categoría de desprecio) publicada en marzo de 2021. El dataset AffectNet completo es enorme (122 GB) y es necesaria una solicitud específica [81].

4.2.3 Desventajas

La recopilación automática desde Internet como fuente y el crowdsourcing de etiquetas son las principales razones de los inconvenientes de cantidad y calidad de los datasets de FER. En cuanto a la cantidad, la mayor desventaja es el desequilibrio, incluso con categorías que superan ampliamente el número de imágenes faciales de otras categorías. Por otro lado, la calidad del contenido es muy afectada por la presencia de imágenes irrelevantes y la clasificación errónea. Estos problemas se mencionan ampliamente en la literatura y aumentan a medida que crece el tamaño del dataset [73].

4.2.3.1 Desequilibrio

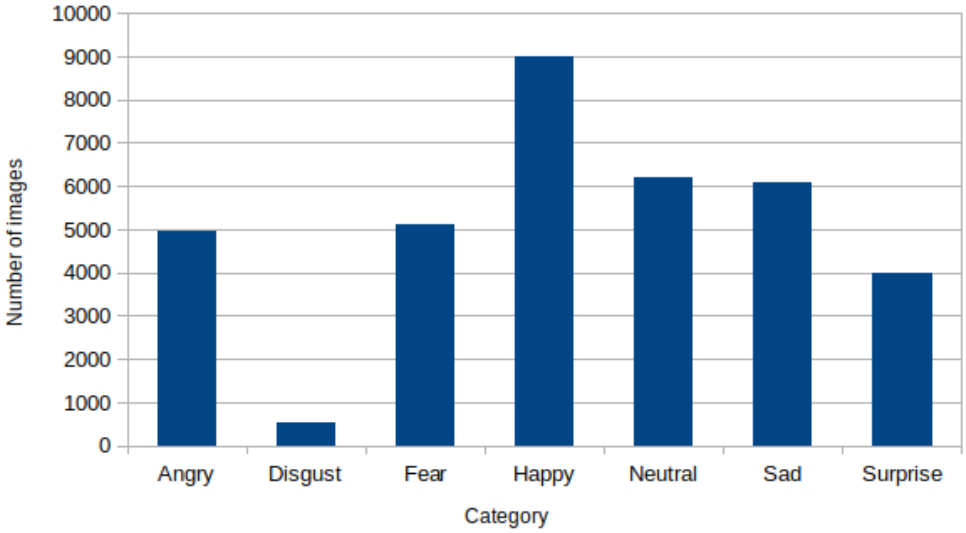
Un dataset desequilibrado podría dar lugar a un modelo de reconocimiento sesgado a favor de las clases mayoritarias. Disponer del mismo número de imágenes por categoría es una tarea difícil. Las imágenes faciales suelen proceder de Internet y se recopilan manual o automáticamente mediante *plug-ins* del navegador o scripts de programación. Las imágenes faciales generalmente son publicadas por personas que suelen mostrar caras sonrientes o felices, por lo que predomina esta categoría, en contraste con categorías como asco, enfado, o tristeza, que los usuarios no suelen publicar. La Tabla 4.2 indica el número de imágenes por categoría de emoción en cada dataset.

Tabla 4.2: Distribución de categorías y número de imágenes en los datasets FER considerados.

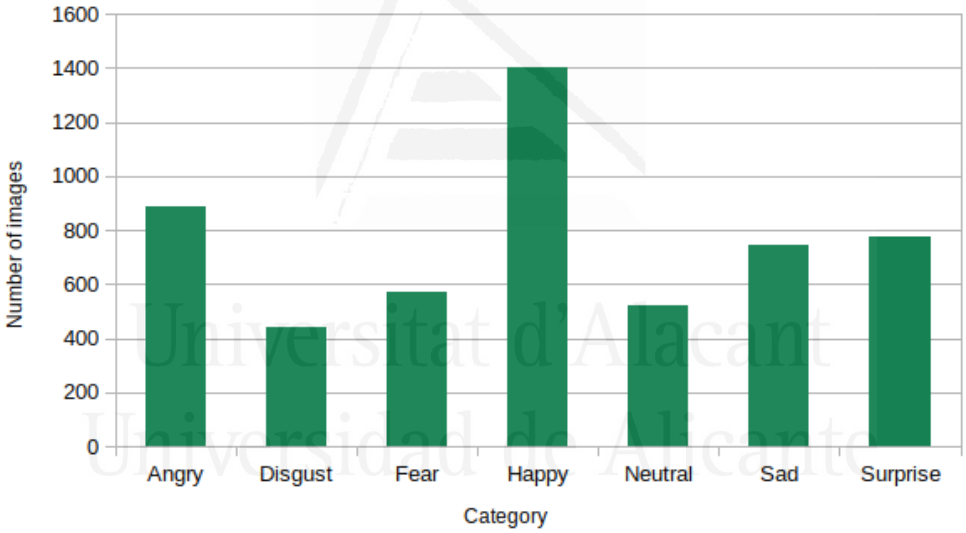
Dataset	Enfado	Asco	Miedo	Feliz	Neutral	Triste	Sorpresa	Total
FER2013	4953	547	5121	8988	6198	6077	4002	35886
NHFI	890	439	570	1406	524	746	775	5350
AffectNet	25382	4303	6878	134915	75374	25959	14590	287401

Los tres datasets muestran un desequilibrio significativo (Figura 4.1). En FER2013 (Figura 4.1a), predomina la categoría “feliz” y la categoría “enfado” tiene pocas muestras, y es aproximadamente regular para el resto de categorías. NHFI (Figura 4.1b) presenta un comportamiento similar, pero menos irregular. En AffectNet (Figura 4.1c), la diferencia en el número de imágenes entre todas las categorías es mucho más pronunciada.

La comparación de las distribuciones en la misma escala (Figura 4.1d) indica que el desequilibrio es mucho más significativo en AffectNet. Un patrón común es el mayor número de muestras para la categoría de feliz y el menor número para la categoría de



(a) Distribución de imágenes por categoría en FER2013.

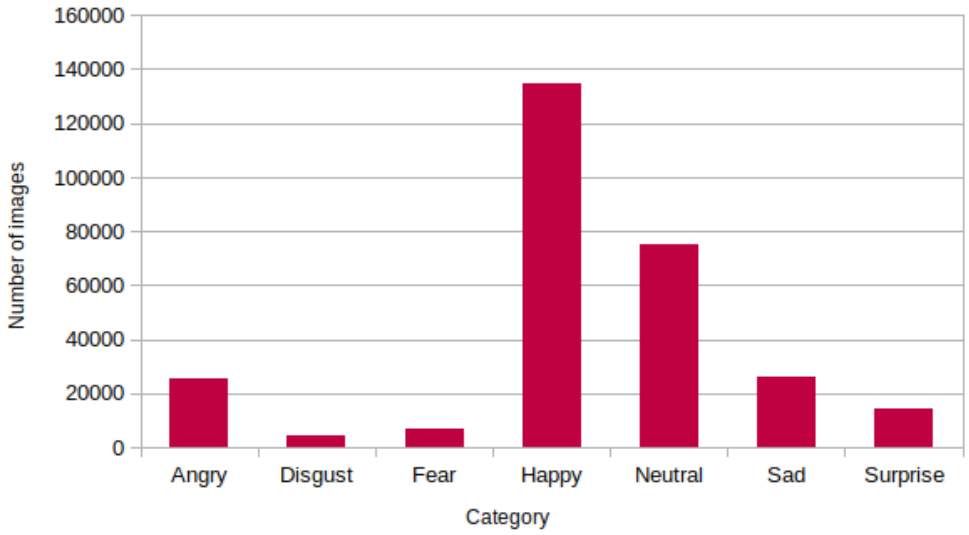


(b) Distribución de imágenes por categoría en NHFI.

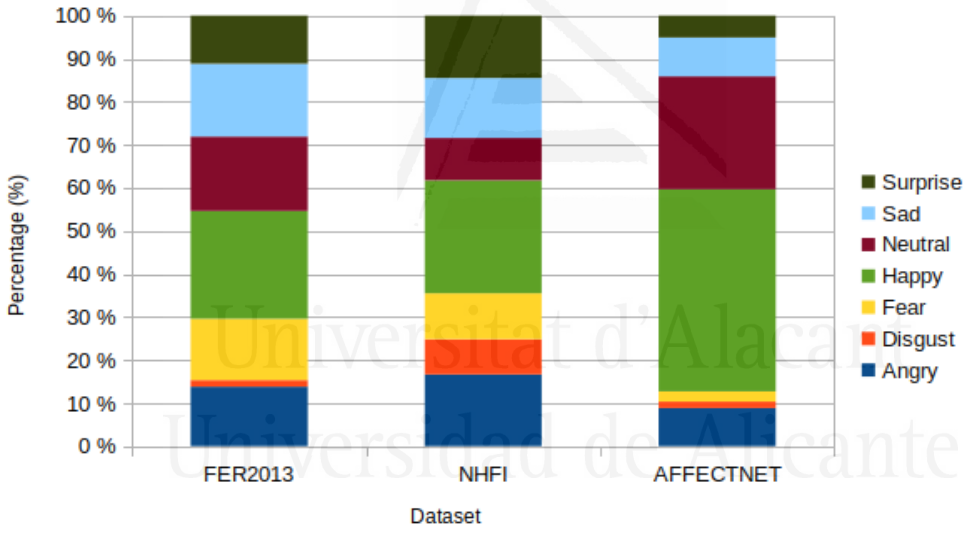
asco. Como ya se ha mencionado, esto se debe a que la gente tiende a publicar imágenes de caras felices y evita mostrar otros tipos de emoción.

4.2.3.2 Clasificación errónea e imágenes irrelevantes

Estos dos problemas se relacionan con el contenido de los datasets, así que los tratamos conjuntamente en esta sección. La clasificación o etiquetado erróneo se refiere a la colocación de imágenes faciales en las categorías equivocadas. Entre los factores que conducen a este problema se encuentran [53][104][116]:



(c) Distribución de imágenes por categoría en AffectNet.



(d) Comparación global de los 3 datasets.

Figura 4.1: Desequilibrio en (a) FER2013, (b) NHFI, (c) AffectNet, y (d) Global.

- Las emociones son subjetivas, ya que es habitual que dos personas tengan opiniones diferentes sobre la misma imagen facial.
- Existen ligeras diferencias entre determinados tipos de emoción, por ejemplo, miedo y sorpresa, asco y enfado, y desprecio y tristeza.
- El grado de expresividad varía de una persona a otra, por lo que los gestos pueden parecer exagerados en un caso e inhibidos en otros.

- Los humanos pueden sentir múltiples emociones en un instante dado, algo que es difícil de combinar en una y puede resultar confusa, por ejemplo, sonreír y lágrimas a la vez, es una emoción combinada que se confunde con tristeza.

Las imágenes irrelevantes son las que tienen marcas de agua, oclusiones, sin caras, poco visibles o muy oscuras, dibujos, caricaturas, texto o símbolos, medias caras, caras dormidas u ojos cerrados, imágenes recortadas, giradas, retocadas y duplicadas. Es importante comprobar estos inconvenientes en cada dataset. Sin embargo, una revisión manual y visual exhaustiva del gran número de imágenes resulta poco práctica. Por tal razón, diseñamos el procedimiento mostrado en la Figura 4.2 para localizar fácilmente estos errores.

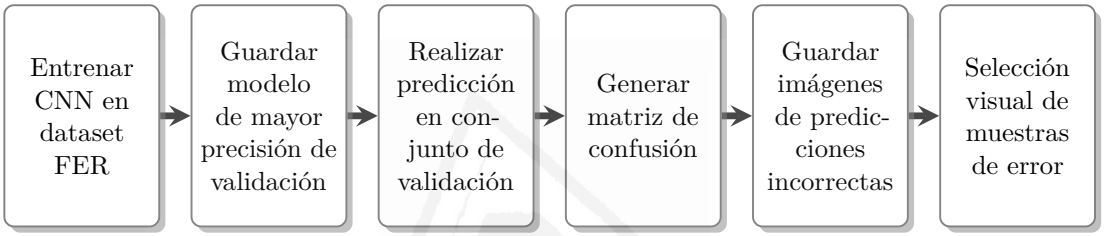


Figura 4.2: Flujo de trabajo para identificar, seleccionar y mostrar algunas imágenes irrelevantes de los datasets FER.

Reutilizamos la CNN para el reconocimiento de expresiones faciales diseñada por Akshit Bhalla [7]. Durante el entrenamiento en cada dataset, monitorizamos la precisión del conjunto de validación en cada iteración (época) para guardar los mejores parámetros del modelo. Este modelo se utiliza para realizar la predicción sobre todas las imágenes del conjunto de validación. A partir de estas predicciones se obtiene la matriz de confusión, en la que las posiciones fuera de la diagonal principal permiten identificar los fallos y sus imágenes correspondientes. Como resultado, tenemos un conjunto más pequeño de imágenes de cada clase que se almacenan en una carpeta separada. A continuación, revisamos visualmente para seleccionar ejemplos de etiquetado erróneo e imágenes irrelevantes con sus respectivos nombres de archivo (Figuras 4.3, 4.4 y 4.5).

En esta sección, examinamos los problemas de los datasets de FER, que pueden resumirse en desequilibrio de clases, la existencia de un número significativo de imágenes irrelevantes o que no corresponden a la categoría correcta. Combinados o por separado, estos problemas hacen que el rendimiento de un modelo FER se degrade considerablemente, así como que el aprendizaje esté sesgado a favor de las clases dominantes [53][59][73][119]. Por lo tanto, la búsqueda de arquitecturas y configuraciones más convenientes para los modelos de reconocimiento es una pérdida de tiempo cuando los datos utilizados son

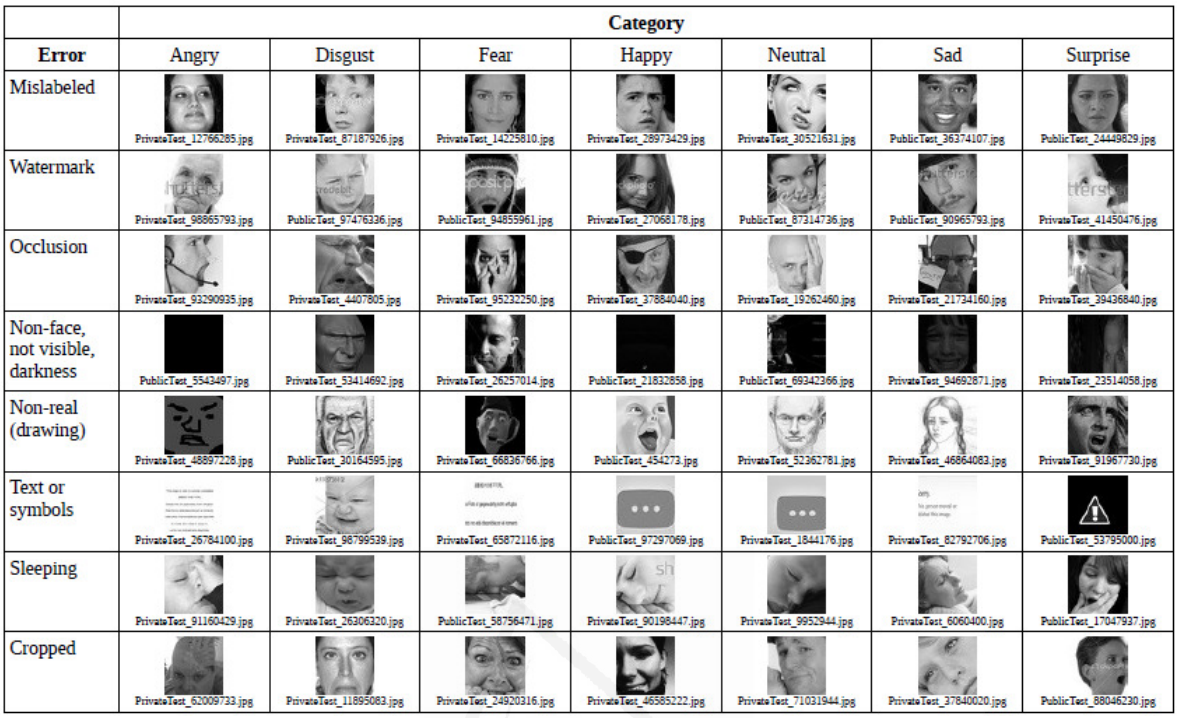


Figura 4.3: Algunos errores detectados automáticamente en el dataset FER2013.

de baja calidad. Por lo tanto, es necesario abordar estos problemas para mejorar los datasets.

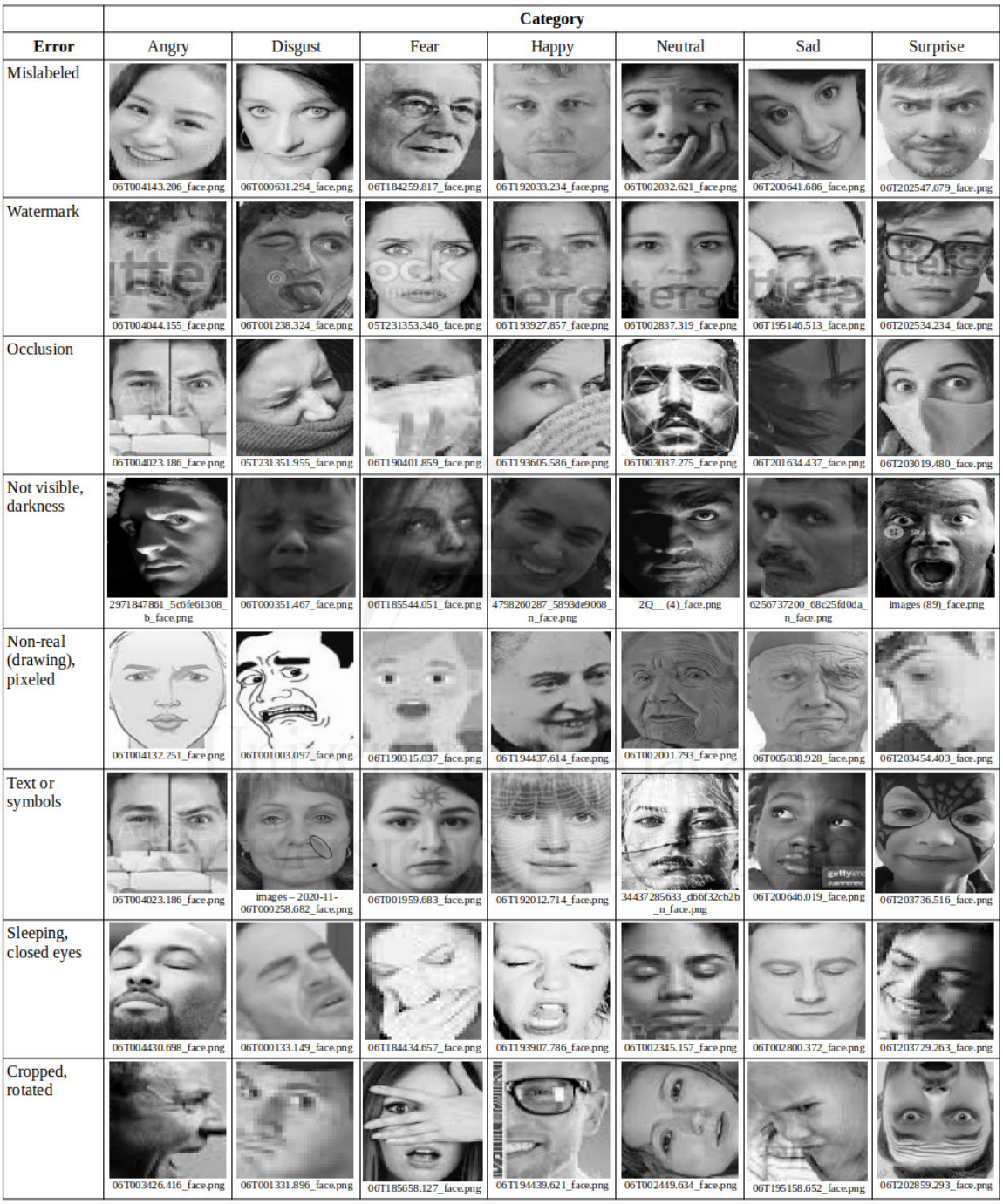


Figura 4.4: Algunos errores detectados automáticamente en el dataset NHFI.

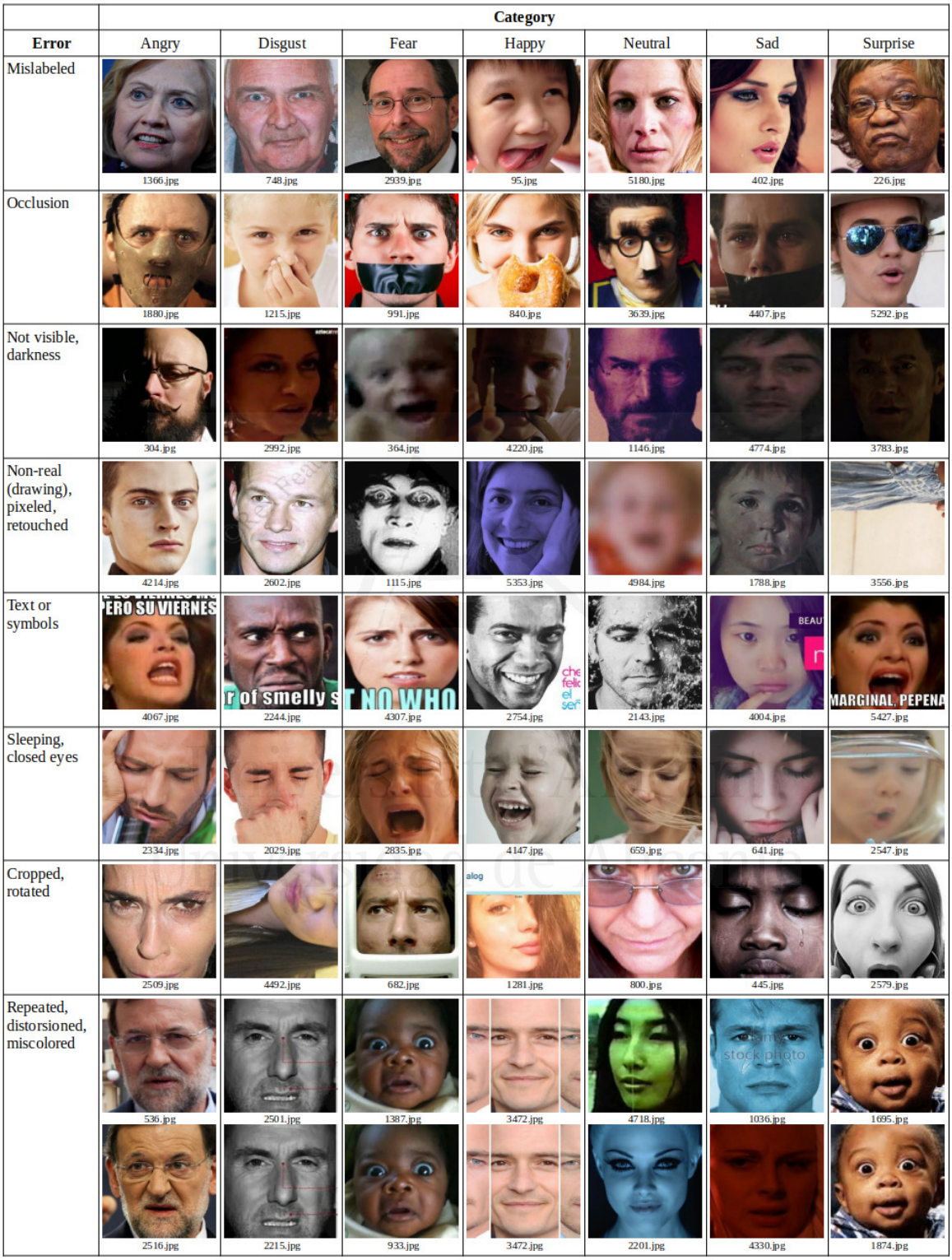


Figura 4.5: Algunos errores detectados automáticamente en el dataset AffectNet.

4.3 Metodología

Proponemos un método que se ilustra en la Figura 4.6, el cual está basado en un novedoso enfoque data-centric, que enfrenta las problemáticas de los datasets FER in-the-wild, y se apoya en el tradicional enfoque model-centric, probando arquitecturas de modelos personalizadas y de aprendizaje por transferencia, con el objetivo de mejorar el reconocimiento de emociones y emplearlo en una aplicación del mundo real.

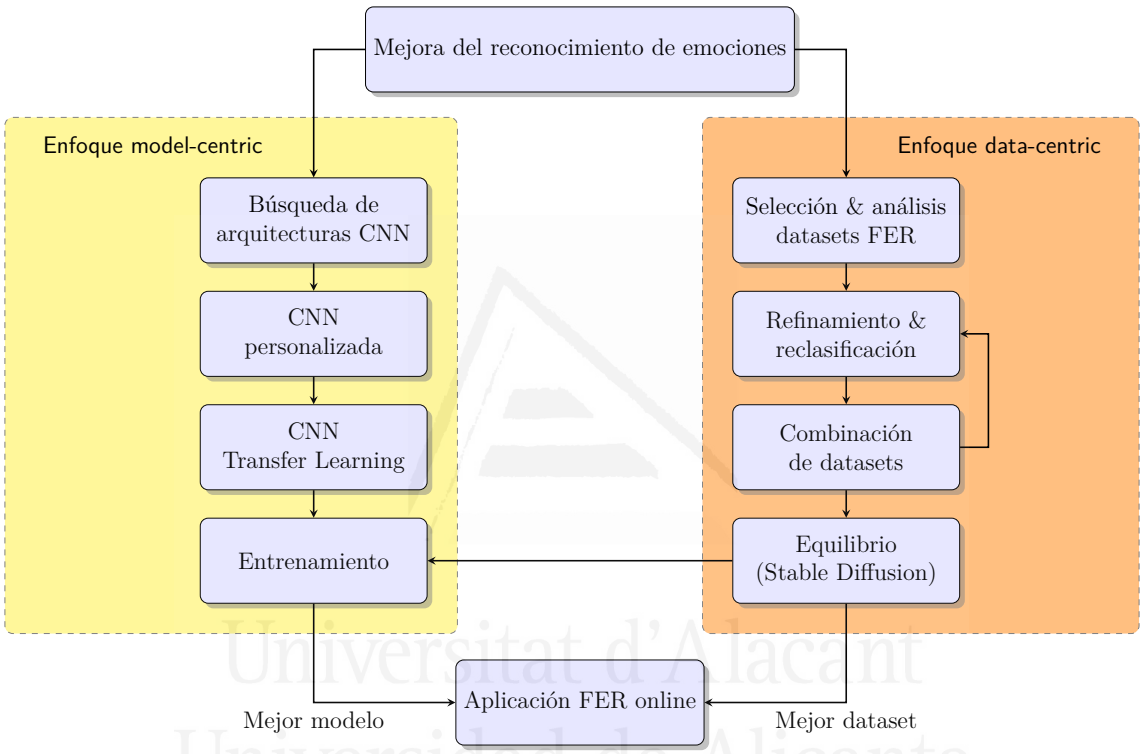


Figura 4.6: Metodología propuesta para mejorar el reconocimiento de emociones.

Nuestro enfoque data-centric diseñado para la mejora de los datasets FER se compone principalmente de un proceso de refinamiento iterativo del dataset para el reetiquetado automático de las imágenes faciales, la combinación de datasets para incrementar la cantidad, diversidad y generalización en escenarios prácticos, así como la generación de imágenes artificiales para el equilibrio. A continuación, estas estrategias son descritas en detalle.

4.3.1 Refinamiento iterativo

La clasificación errónea que suele encontrarse en los datasets de FER es probablemente el inconveniente que más influye en el menor rendimiento de los modelos de reconocimiento de emociones en escenarios reales. Dado que una inspección visual de cada imagen facial de un dataset sería una tarea extremadamente lenta y tediosa, diseñamos el método representado por el diagrama de flujo de la Figura 4.7 para reclasificar automáticamente las imágenes de un dataset y mejorar el rendimiento de un modelo FER. El método propuesto consta de los siguientes pasos:

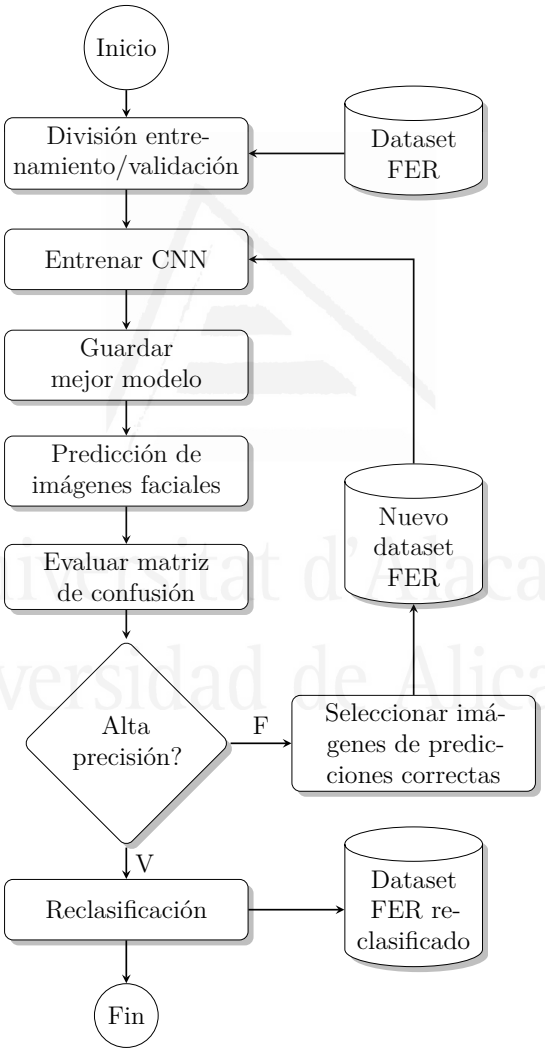


Figura 4.7: Flujo de trabajo para reclasificar automáticamente un dataset FER.

1. El dataset está organizado con una estructura de carpetas, en la que cada categoría de emoción es una carpeta que contiene los archivos de imágenes faciales correspondientes.

2. El dataset es dividido en subconjuntos de entrenamiento y validación con la misma estructura de carpetas y archivos. El subconjunto de entrenamiento es más grande y contiene las imágenes para ajustar el modelo, mientras que el subconjunto de validación se utiliza para evaluar el modelo en el momento del entrenamiento. Omitimos un subconjunto de prueba porque se necesita el mayor número posible de imágenes para el siguiente paso. Esta es la entrada que será procesada por la red convolucional.

3. Una CNN personalizada o preentrenada mediante aprendizaje por transferencia se entrena en el dataset FER. En este trabajo se muestran ambas alternativas. El entrenamiento es un proceso de optimización iterativo en el que el modelo reduce un error a medida que aprende a asociar imágenes y etiquetas de categoría.

4. El entrenamiento se supervisa para guardar en un archivo de modelo los parámetros (pesos y sesgos) correspondientes a la iteración (época) de la mejor precisión de validación.

5. El mejor modelo se utiliza para realizar la predicción de todas las imágenes faciales del dataset entero. Los resultados obtenidos permiten generar la matriz de confusión.

6. La matriz de confusión es evaluada considerando un buen dataset cuando la precisión de cada categoría supera el 90% o los números fuera de la diagonal principal son de un solo dígito. Serán necesarios varios entrenamientos sucesivos para cumplir estos criterios.

7. Las predicciones correctas en la diagonal principal de la matriz de confusión nos permiten seleccionar las imágenes faciales correspondientes, que formarán una versión más pequeña, pero mucho más fiable del dataset.

8. La nueva versión es dividida automáticamente en subconjuntos de entrenamiento y validación, y el entrenamiento se realiza con la misma CNN. El proceso se repite hasta que se alcanzan las condiciones establecidas para un buen dataset.

9. El último modelo guardado realiza la predicción de la etiqueta de emoción para todas las imágenes del dataset original. El resultado es la reclasificación automática que genera una nueva distribución del dataset con todas las imágenes faciales.

En resumen, proponemos un proceso iterativo de entrenamientos para crear versiones sucesivamente más refinadas del dataset. Cada versión es más pequeña, sólo se incluyen las predicciones correctas, pero mantiene un número significativo de imágenes. En el último entrenamiento se obtiene un dataset mucho más fiable, así como un modelo que produce un bajo número de predicciones incorrectas (valores de un solo dígito para cada clase). La red convolucional es fija en cuanto a su arquitectura e hiperparámetros a lo largo de este proceso.

La idea clave es que la extracción de características es la parte crucial de un sistema FER, y la precisión de la clasificación de emociones mejorará con una extracción eficaz de rasgos faciales [69][108]. El refinamiento progresivo del dataset produce un menor número de imágenes en cada entrenamiento, pero con una menor variabilidad de los gestos de las caras. Por lo tanto, el modelo puede captar gradualmente más rasgos distintivos de cada clase. Como consecuencia, es posible aumentar la similitud intraclase y ampliar las diferencias interclase dentro de un dataset, mejorando así la precisión del reconocimiento de emociones en escenarios del mundo real.

El dataset mejorado es el recurso esencial para entrenar un modelo de DL de clasificación de imágenes faciales en categorías de emociones. Aprovechamos las CNNs, que son las actuales herramientas del estado del arte en problemas de visión computacional. El diseño de una CNN imita el sistema visual humano, donde una parte convolucional serían los ojos de la red, mientras que una parte clasificadora sería el cerebro, que decide la clase del objeto. Las CNNs pueden ser creadas con dos enfoques principales: (1) una arquitectura personalizada, y (2) mediante la técnica de aprendizaje por transferencia. En el primer caso, definimos la estructura de las capas y los hiperparámetros de la red desde cero, mientras que en el segundo, reutilizamos un modelo ya creado, preentrenado y disponible públicamente. En este trabajo demostramos el uso de ambas alternativas, describiendo la arquitectura implementada para cada uno de los datasets seleccionados.

4.3.1.1 Arquitectura del modelo para FER2013

Reutilizamos una CNN presentada en el sitio Web de Kaggle[5], cuyo rendimiento ha mostrado buenos resultados en la tarea de reconocimiento de la expresión facial en este dataset (Figura 4.8).

La imagen de entrada de 48x48 píxeles en escala de grises es procesada por 4 capas convolucionales, cada una de las cuales aplica una serie de filtros (kernels) para generar mapas de características que incluyen patrones detectados jerárquicamente, desde los más simples a los más complejos. Aquí se aplican 64, 128, 512, y 512 filtros de tamaño

[5] https://www.kaggle.com/bhallaakshit/facial-expression-recognition

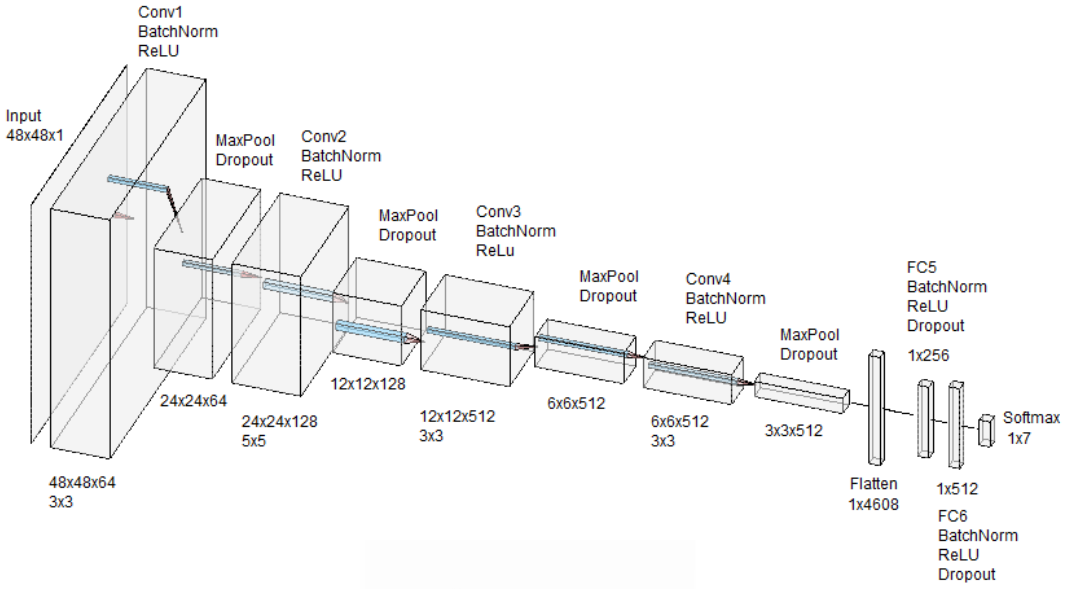


Figura 4.8: Arquitectura de la CNN personalizada para el dataset FER2013.

3x3, 5x5, 3x3, y 3x3 píxeles, respectivamente. A continuación, una función de activación ReLU convierte los valores negativos en cero y mantiene los positivos. Luego, una operación de maxpooling reduce las dimensiones de la imagen a la mitad, pero conserva las características detectadas. La normalización por lotes estabiliza el resultado de una convolución, mientras que el dropout permite la participación activa de todas las neuronas en el proceso de aprendizaje. Ambas son técnicas de regularización recomendadas para evitar posibles sobreajustes. La operación de aplanamiento convierte los mapas de características en un vector de valores como entrada para el clasificador, que es una red neuronal tradicional con una capa de entrada que recibe las características en forma de vector, dos capas ocultas de 256 y 512 neuronas, y una capa de salida con una función de activación Softmax para 7 valores de probabilidad, uno por cada clase de emoción.

4.3.1.2 Arquitectura del modelo para NHFI

La misma arquitecura de CNN es probada con este dataset. Sin embargo, los resultados tras el primer filtrado indicaron un aumento insignificante de la precisión (aprox. 1.5 %) como se muestra en la Tabla 4.3.

Tabla 4.3: Refinamiento del dataset NHFI utilizando la CNN personalizada.

Entrenamiento	Imágenes (entrenamiento)	Imágenes (validación)	Total	Precisión
1	4278	1072	5350	0.5732
2	3211	616	3827	0.5885

Por ende, buscamos otras arquitecturas para lograr una mayor precisión. Un modelo que utiliza aprendizaje por transferencia mostró el mejor rendimiento para este dataset. En el primer filtrado, la precisión mejoró de 0.5597 a 0.8367, un incremento de 27.7% frente al 1.5% con la CNN personalizada. Es demostrado que el método propuesto funciona para ambos casos (modelos preentrenados y personalizados). Mediante aprendizaje por transferencia, la fase de entrenamiento es mucho más rápida, pues sólo entrenamos los parámetros del clasificador manteniendo fija la base convolucional que ya habría aprendido características útiles para la mayoría de los problemas de visión por computadora. La estructura se presenta en la Figura 4.9.

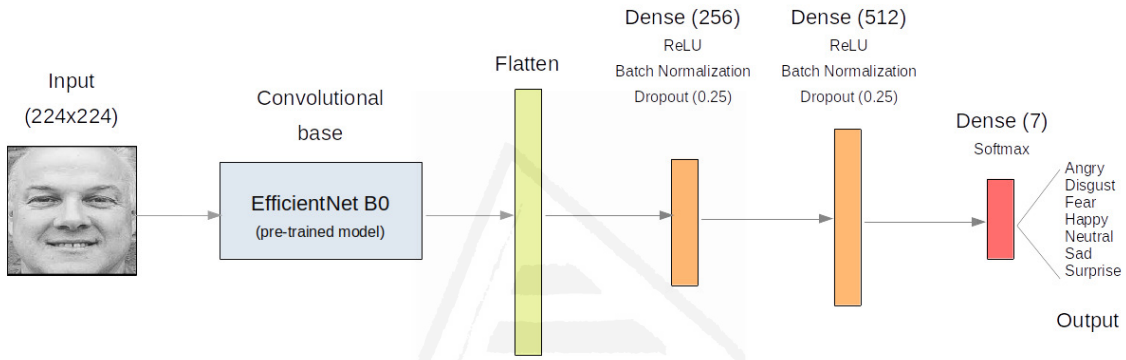


Figura 4.9: Arquitectura de la CNN con aprendizaje por transferencia para el dataset NHFI.

La arquitectura está basada en *EfficientNet*, una CNN muy popular preentrenada en el dataset ImageNet [102]. Utilizamos la versión *B0*, cuya base convolucional se mantiene para la extracción de características. La ventaja es que se acepta como entrada la imagen con la resolución original de 224x224 píxeles. El clasificador recibe las características en forma de vector aplanado para decidir la clase a la que pertenece la imagen de entrada mediante una red neuronal completamente conectada formada por dos capas densas de 256 y 512 neuronas, a las que se aplica la función de activación ReLU más las técnicas de regularización de normalización por lotes y dropout para reducir el posible sobreajuste. La función Softmax de la última capa densa genera una distribución de probabilidades correspondiente a cada una de las 7 categorías de emoción.

4.3.1.3 Arquitectura del modelo para AffectNet

Realizamos varios intentos con diferentes arquitecturas para determinar la CNN más adecuada para este dataset. El mejor resultado se obtuvo con la CNN utilizada para FER2013 (Figura 4.8). Sólo es necesario cambiar la resolución y el modo de color de las imágenes de AffectNet de 224x224 píxeles en RGB a 48x48 píxeles en escala de

grises. Esta conversión se realiza automáticamente utilizando el generador de imágenes de Python.

4.3.2 Combinación de datasets de FER

En esta sección, se explica el trabajo desarrollado para crear el mayor dataset combinado de FER in-the-wild, según lo que conocemos. Este dataset se divide en subconjuntos de entrenamiento y de prueba, este último combinado, equilibrado, insesgado y bien etiquetado, diseñado para medir la calidad de generalización de los modelos y de los datasets en los que se entrenan. Además, presentamos el primer y mayor dataset de imágenes faciales totalmente artificial, creado con Stable Diffusion, una herramienta que se ha convertido en el estado del arte para los modelos generativos. Estos productos pueden ser útiles para la investigación y el desarrollo en el reconocimiento de emociones. El flujo de trabajo se representa gráficamente en la Figura 4.10. Seguidamente, describimos cada una de sus etapas.

Figura 4.10: Metodología para la mejora de la generalización en FER.

4.3.2.1 Creación del dataset combinado

Dado que la recopilación de datos es una tarea costosa, abordamos el problema de la falta de generalidad combinando datasets existentes para ampliar la gama de rostros, variaciones de expresión, y muchos individuos diferentes. Para ello, utilizamos las versiones reclasificadas que nos permitieron mejorar la precisión de los modelos categóricos de reconocimiento de emociones.

Tabla 4.4: Distribución de las categorías de emociones y el número de imágenes faciales de los datasets FER considerados.

Dataset	Enfado	Asco	Miedo	Feliz	Neutral	Triste	Sorpresa	Total
FER2013	4953	547	5121	8988	6198	6077	4002	35886
NHFI	890	439	570	1406	524	746	775	5350
AffectNet (balanceado)	4300	4300	4300	4300	4300	4300	4300	30100
Total	10143	5286	9991	14694	11022	11123	9077	71336
FER2013 (reclasificado)	4817	532	3842	9202	7074	6090	4329	35886
NHFI (reclasificado)	336	514	383	1585	1042	962	528	5350
AffectNet (reclasificado)	4394	3893	5004	4594	4224	4071	3920	30100
Combinado	9547	4939	9229	15381	12340	11123	8777	71336

Las categorías de emoción y el número de imágenes faciales totales y por categoría se presentan en la Tabla 4.4, tanto para los datasets originales como para sus versiones reclasificadas, que se utilizan para formar el dataset FER combinado. Aunque el AffectNet original contiene 287401 imágenes, aplicamos *downsampling* considerando la categoría con menor número de imágenes (asco) para obtener una versión reducida, pero equilibrada, de 4300 imágenes seleccionadas aleatoriamente por categoría. Así evitamos el gran desequilibrio que caracteriza a este dataset, que podría causar un sesgo en el entrenamiento del modelo en el dataset combinado.

Cada dataset se distingue por el número y las características de sus propias imágenes faciales. Por lo tanto, la tarea de combinación implica la mezcla de diferentes atributos, como se muestra en la Tabla 4.1. La diversidad de estas características nos permitirá evaluar la capacidad de generalización de los modelos y datasets analizados.

Procedemos a juntar las versiones reclasificadas de FER2013, NHFI y AffectNet en uno solo, que se convierte en el mayor dataset combinado en el ámbito de FER. Para tal fin, organizamos y creamos la estructura de directorios y subdirectorios, donde la carpeta nombrada como “0” corresponde a la categoría “enfado”, “1” a “asco”, “2” a “miedo”, “3” a “feliz”, “4” a “neutral”, “5” a “triste”, y “6” a “sorpresa”. Esto permite inferir automáticamente la lista de clases, donde cada subdirectorio será tratado como una clase diferente. El orden de las clases se asignará a los identificadores de las etiquetas. Luego, movemos los archivos de imágenes faciales de cada dataset reclasificado a las categorías correspondientes. Es necesario asegurarse de que no hay nombres de archivo repetidos para evitar sobrescribirlos y modificar el número total de archivos. De esta manera logramos un dataset más amplio, representativo y diverso compuesto por 71336 imágenes faciales, un número significativo para el campo de las emociones. La mezcla de datasets, trabajando como uno solo, es útil para entrenar un modelo que puede lograr una mayor generalización en el reconocimiento de emociones, debida a la mayor variedad de datos, así como reducir la dependencia de un dataset específico. A partir del dataset combinado, creamos un conjunto de entrenamiento y un conjunto de prueba, este último diseñado para ser una propuesta de conjunto evaluador que permita medir y comparar el rendimiento de diferentes modelos de reconocimiento de emociones. A continuación, describimos el procedimiento para generar este dataset, el cual debe satisfacer las propiedades de ser equilibrado e insesgado.

4.3.2.2 Creación del dataset evaluador

La imparcialidad y el equilibrio son requisitos fundamentales que debe cumplir un dataset evaluador de lo bueno y genérico que es un modelo de reconocimiento de emociones. Para

crear un dataset que satisfaga estas exigencias, seguimos la metodología de trabajo que se muestra en la Figura 4.11.

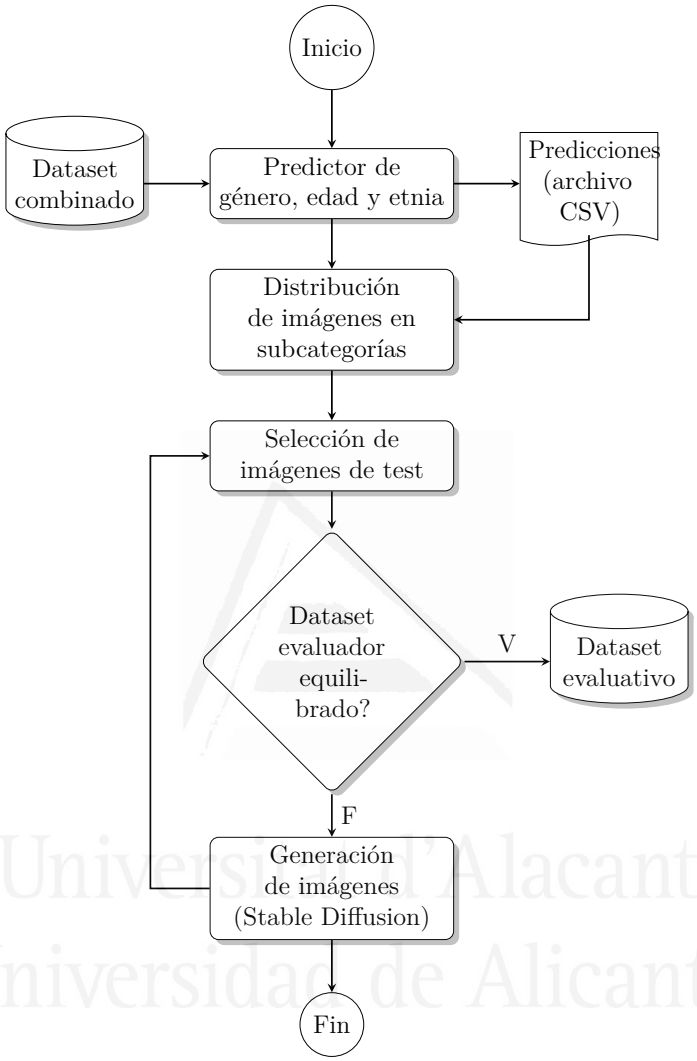


Figura 4.11: Metodología para la creación del dataset evaluador.

Predicción de género, edad y etnia. El dataset evaluador debe ser representativo de la población general. Entre las características más relevantes para formar grupos de personas con fines sociales y comerciales se encuentran el género, la edad y la etnia [101]. Estos atributos pueden analizarse en una imagen facial y son los que suelen introducir sesgos en los datasets de FER. Por lo tanto, es deseable conseguir un dataset evaluador no sesgado. Precisamente, el reconocimiento automático de estas características se ha convertido en uno de los campos más reconocidos del DL por sus aplicaciones en redes sociales, videovigilancia, análisis biométrico, entre otros usos [86].

Aprovechamos el predictor disponible públicamente en GitHub[6]. El autor utiliza una arquitectura de red convolucional simple entrenada en el dataset in-the-wild de acceso abierto UTKFace[7], que proporciona más de 20 mil imágenes faciales, con una sola cara en cada imagen, proporciona las imágenes faciales alineadas y recortadas, con sus respectivas etiquetas de género, edad y etnia. Aunque la implementación de la CNN está disponible en el repositorio de GitHub, el modelo resultante con sus respectivos pesos no está disponible fácilmente para su reutilización. Por tal motivo, entrenamos la red convolucional ejecutando el script de Python publicado en el mismo repositorio[8], el cual modificamos ligeramente, especialmente para guardar el modelo en formato h5 [9], obteniendo las gráficas de rendimiento mostradas en la Figura 4.12.

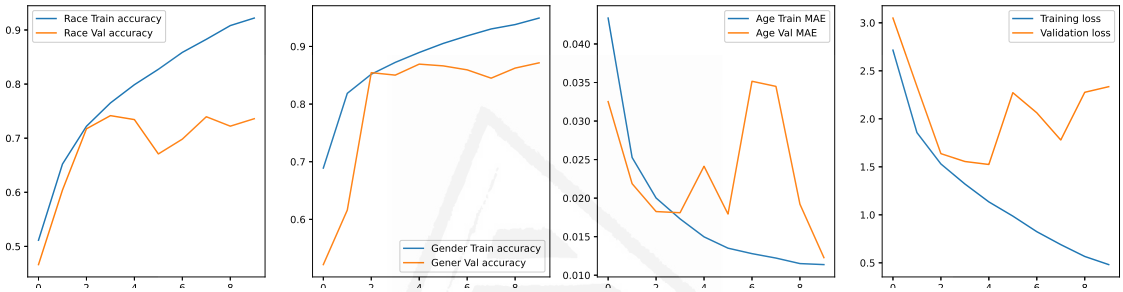


Figura 4.12: Curvas de aprendizaje del modelo de predicción de edad, género y etnia.

Las curvas de aprendizaje del modelo se comportan de forma bastante aceptable. Las variables categóricas de etnia y género presentan las curvas de precisión de entrenamiento y validación. Ambas alcanzan una valor significativo, lo que indica una buena precisión. Para la variable cuantitativa de edad, el gráfico corresponde a la métrica del error medio absoluto (Mean Absolute Error, MAE), cuya curva está muy próxima a cero en la última época, tanto para el entrenamiento como para la validación. Realizamos algunas predicciones de ejemplo utilizando las imágenes de caras del dataset UTKFace, cuyos resultados se presentan en la Figura 4.13. En la parte superior de cada imagen se muestran los valores de predicción, mientras que en la parte inferior aparecen las etiquetas de verdad correspondientes. Los resultados presentan una buena aproximación en la mayoría de los casos a las respuestas verdaderas. Esta precisión del modelo es suficiente para evitar una revisión manual exhaustiva de cada una de las imágenes de cada dataset considerado en este trabajo.

[6] https://github.com/Sobika2531/Age-Gender-And-Race-Detection-Using-CNN
[7] https://susanqq.github.io/UTKFace/
[8] https://github.com/cimejia/novel-FER-datasets/blob/main/AGR-prediction/AGRprediction-V2.py
[9] https://github.com/cimejia/novel-FER-datasets/blob/main/AGR-prediction/best_model_agr.h5

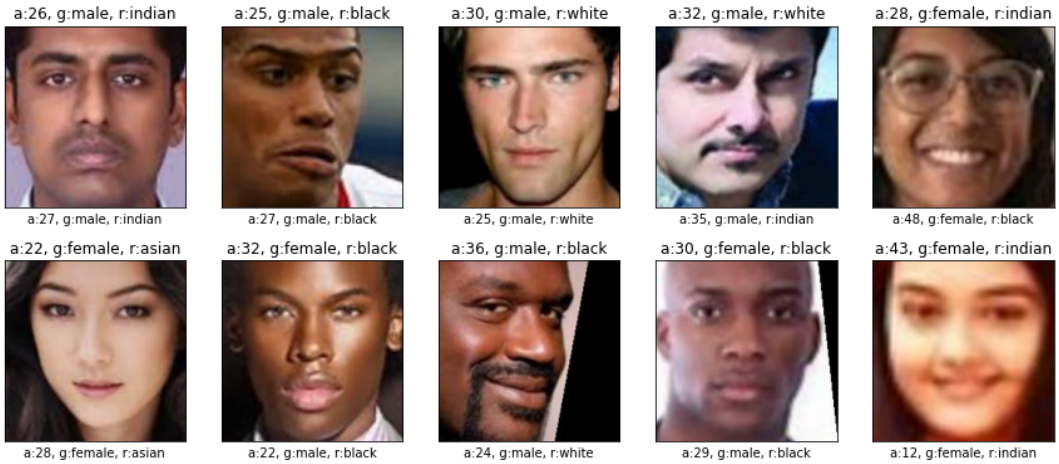


Figura 4.13: Resultados de la predicción de género, edad y etnia para algunas imágenes del dataset UTKFace.

El modelo obtenido permite predecir el género, la edad y la etnia de todas las imágenes faciales de cada dataset. Esto facilita la selección de las imágenes que formarán el dataset evaluador. En un nuevo script[10], cargamos el modelo en formato h5, se leen todas las imágenes de cada una de las categorías de los datasets y utilizamos el modelo para realizar la predicción de características, obteniendo los tres valores que se almacenan junto al nombre de la imagen en un archivo CSV. Hay 7 archivos en total, uno por cada categoría. Este procedimiento se aplica para la versión reclasificada de NHFI, FER2013 y AffectNet. En la Figura 4.14 se presenta una muestra de dicho archivo.

IMAGE	AGE	RACE	GENDER
nhfi_101images_-_2020-11-06T003453.521_face.png	adult	white	male
nhfi_102images_-_2020-11-06T003454.433_face.png	adult	white	male
nhfi_104images_-_2020-11-06T003457.012_face.png	young	white	male
nhfi_1067images-2020-11-05T231437.192_face.png	adult	white	male
nhfi_1085images-2020-11-06T000139.006_face.png	young	white	male
nhfi_108images_-_2020-11-06T003505.662_face.png	adult	indian	male
nhfi_1092images-2020-11-06T000146.076_face.png	adult	white	male
nhfi_1096images-2020-11-06T000154.107_face.png	adult	white	female
nhfi_110images_-_2020-11-06T003508.295_face.png	adult	white	male
nhfi_111images_-_2020-11-06T003509.556_face.png	adult	white	male
nhfi_112Q___5_face.png	young	white	female
nhfi_1169images-2020-11-06T000951.890_face.png	adult	black	male
nhfi_1187images-2020-11-06T001008.819_face.png	adult	white	male
nhfi_118images_-_2020-11-06T003523.437_face.png	child	white	male
nhfi_1210images-2020-11-06T001040.744_face.png	young	white	female
nhfi_121images_-_2020-11-06T003528.161_face.png	young	white	male
nhfi_122Q___face.png	adult	white	male
nhfi_1243images-2020-11-06T001241.445_face.png	young	asian	male

Figura 4.14: Extracto del archivo de predicción de género, edad y etnia.

[10] https://github.com/cimejia/novel-FER-datasets/blob/main/AGR-prediction/AGRprediction-test-V2.py

La predicción de la edad es un valor numérico, a diferencia del género y la etnia. Para convertirla en un valor cualitativo, definimos las siguientes reglas: edad inferior a 15 años, se asigna la clase "niño"; edad igual o superior a 15 años e inferior a 30, la clase es "joven"; edad igual o superior a 30 años e inferior a 65 es "adulto"; y edad igual o superior a 65 es la clase "anciano".

Además de simplificar la selección de imágenes faciales, la predicción del género, la edad y la etnia nos da una aproximación del desequilibrio de estas variables en cada dataset. Reunimos los 7 archivos de predicción en uno solo, lo importamos a una hoja de cálculo y elaboramos los gráficos de distribución que se presentan en la Figura 4.15.

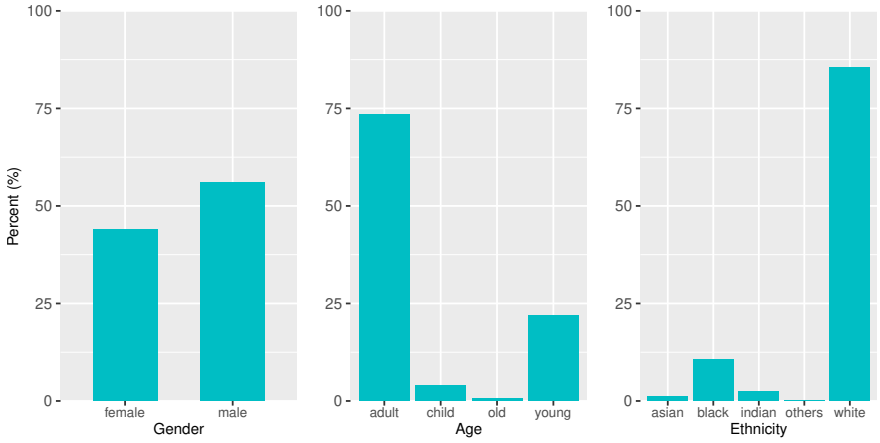
El elevado desequilibrio entre grupos de cada variable es evidente. El género masculino, la edad adulta y la etnia blanca predominan en los tres datasets. Los modelos entrenados con estos datasets mostrarán un sesgo favorable a estos grupos. Siguen, por un amplio margen, la edad joven y la etnia negra. Por último, hay poca presencia de niños y ancianos, así como de asiáticos, indios y otras etnias no blancas ni negras. Estos resultados confirman la conocida falta de representatividad de los datasets de FER in-the-wild, con sobrerrepresentación de ciertos grupos de personas y subrrepresentación de otros. Este es uno de los principales problemas que reducen el rendimiento del reconocimiento de emociones en diferentes datasets y entornos reales. Por lo tanto, nuestro dataset evaluador incluye todos los grupos definidos de género, edad y etnia por igual, es decir, el mismo número de imágenes faciales dentro de cada clase, lo que se explica a continuación.

Distribución de imágenes faciales. En este paso creamos las carpetas y subcarpetas según la jerarquía que se muestra en la Figura 4.16. La carpeta principal corresponde a cada categoría de emoción, mientras que las subcarpetas corresponden a todas las combinaciones posibles de género, edad y etnia. Como resultado, cada tipo de emoción se compone de 40 subcategorías ($2x4x5$) que nos permiten seleccionar las imágenes faciales de forma equilibrada e imparcial.

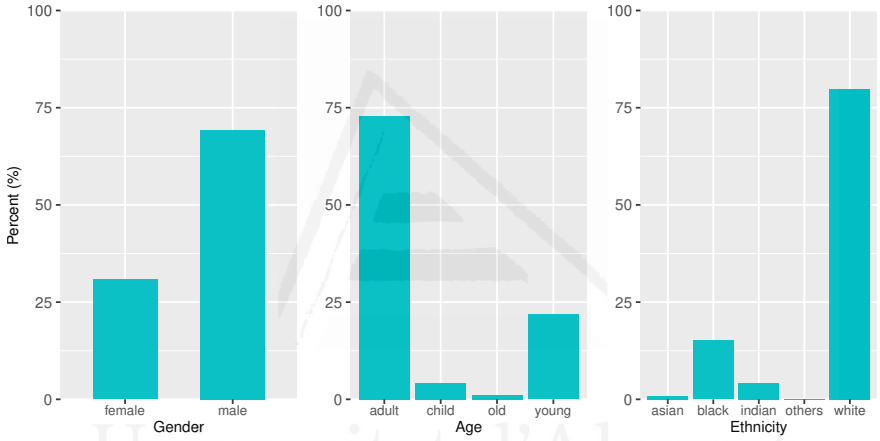
La nomenclatura utilizada para cada subcarpeta es clave para la posterior localización de las imágenes. El nombre se forma juntando el género, la edad y la etnia, por ejemplo, en la categoría "enfado" existe una carpeta denominada *female-adult-asian*, que almacena las imágenes faciales detectadas como mujeres de edad adulta de origen asiático con expresión de enfado.

Las imágenes faciales de cada dataset se distribuyen automáticamente en las 40 carpetas mediante un script[11] con el apoyo de las librerías *csv* y *shutil*. Este script lee el

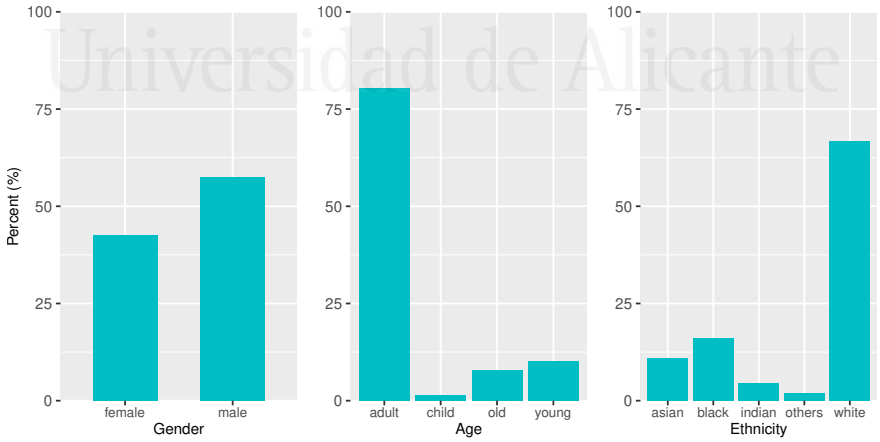
[11] https://github.com/cimejia/novel-FER-datasets/blob/main/AGR-prediction/folders-distribution.py



(a) FER2013



(b) NHFI



(c) AffectNet

Figura 4.15: Distribución de las imágenes faciales según el predictor de género, edad y etnia para los datasets: (a) FER2013, (b) NHFI, y (c) AffectNet.

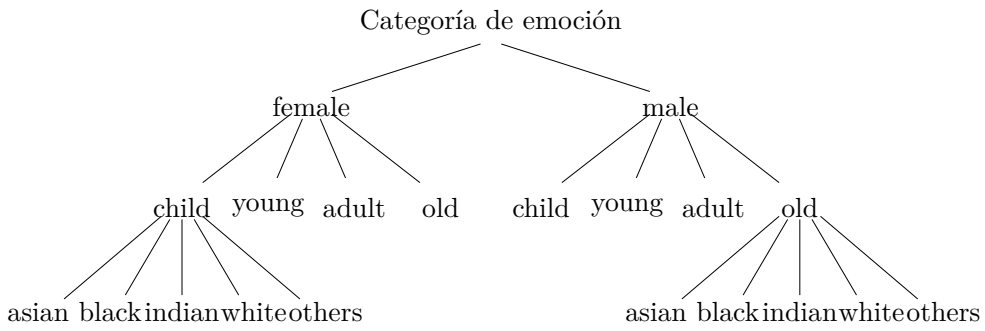


Figura 4.16: Generación de subcategorías para el dataset evaluador.

archivo de predicción en formato de texto cuyas líneas contienen los valores del nombre del archivo de imagen, género, edad y etnia delimitados por una coma. Iterativamente, se lee cada uno de los registros y se forma el nombre de la carpeta correspondiente con estos valores. Por último, se mueve el archivo de imagen a su carpeta respectiva. Este proceso se realiza para cada tipo de emoción, obteniendo una nueva distribución del dataset estructurada en 7 categorías de emociones y 40 subcategorías, una para cada combinación de género, edad y etnia. Esta distribución automática permite agilizar en gran medida la selección equitativa e imparcial de las imágenes faciales, frente a una forma totalmente manual y exhaustiva.

Selección de imágenes faciales de prueba. En primer lugar, es necesario determinar el número de imágenes que deben seleccionarse de cada subcategoría de género, edad y etnia, y para cada categoría de emoción. Este número resulta de dividir el tamaño del dataset evaluador por las 40 subcategorías. El total de imágenes está condicionado por el subconjunto de prueba más pequeño de los tres datasets considerados para la combinación. La Tabla 4.5 muestra la división de los datasets en tres subconjuntos: *entrenamiento*, *validación* y *prueba*, aplicando un 80%, 10% y 10%, respectivamente. Esta proporción es una de las más recomendadas para el aprendizaje automático. En consecuencia, el tamaño del subconjunto de prueba NHFI (535) determina el número de imágenes que deben seleccionarse de cada dataset.

Tabla 4.5: Distribución de subconjuntos de entrenamiento, validación y prueba para cada dataset FER.

Dataset	Entrenamiento (80%)	Validación (10%)	Prueba (10%)
FER2013 (reclasificado)	28708	3589	3589
NHFI (reclasificado)	4280	535	535
AffectNet (reclasificado)	4394	3893	5004

Esta cantidad no es perfectamente divisible por el número total de subcategorías. Seleccionando una imagen por cada una de las 40 subcategorías y por cada una de las 7 categorías de emoción, obtenemos 280 (40x7) imágenes en total, lo que equivale al 5.23% del dataset, es decir, por debajo de la estimación. Seleccionando 3, el resultado es de 840 (3x40x7), lo que equivale a un 15.7%, es decir, superior a la estimación. Por lo tanto, seleccionamos por inspección visual 2 imágenes faciales para cada una de las 40 subcategorías, un total de 560 imágenes faciales correspondientes al 10.47%, un porcentaje similar al recomendado. La Tabla 4.6 presenta el número de imágenes faciales seleccionadas en total, por dataset y categoría dentro de cada dataset, que deberían ser 1680, 560 y 80, respectivamente.

Tabla 4.6: Número de imágenes seleccionadas por cada dataset y categoría de emoción para conformar el dataset evaluador.

Dataset	Enfado	Asco	Miedo	Feliz	Neutral	Triste	Sorpresa	Total
FER2013	67	50	70	80	80	77	64	488
NHFI	62	62	59	79	75	69	55	551
AffectNet	77	70	70	80	75	72	75	519
Total	206	182	199	239	230	218	194	1468

Durante la tarea de selección, en algunos casos se presentó sólo una imagen facial o ninguna. Esto ocurre sobre todo en las clases asociadas a niños y ancianos, así como a personas de origen asiático, indio y latino. Por lo tanto, los datasets originales tienen un número insuficiente de imágenes para satisfacer los criterios de género, edad y etnia, lo que nos impide completar el dataset evaluador. Por lo tanto, debemos generar las imágenes faciales que faltan para lograr el equilibrio. La Tabla 4.7 muestra el número de imágenes faciales que son necesarias para equilibrar cada uno de los datasets, así como cada una de las categorías de emoción. En total, hay 212 imágenes faciales que no se encontraron en los datasets siguiendo los criterios de género, edad y etnia.

Tabla 4.7: Número de imágenes faciales faltantes para equilibrar el dataset evaluador.

Dataset	Enfado	Asco	Miedo	Feliz	Neutral	Triste	Sorpresa	Total
FER2013	13	30	10	0	0	3	16	72
NHFI	18	18	21	1	5	11	25	99
AffectNet	3	10	10	0	5	8	5	41
Total	34	58	41	1	10	22	46	212

Generación de imágenes faciales artificiales. La generación de imágenes sintéticas es nuestra estrategia para obtener las 212 imágenes de expresión facial restantes. En esta

sección, presentamos el uso de GAN y Stable Diffusion, y comparamos los resultados para seleccionar la técnica más adecuada.

GAN [34]. Como se revisó en el Capítulo 2, las imágenes sintéticas para aumentar los datasets faciales se generan especialmente con GANs. Normalmente esto ocurre para la etapa de entrenamiento. En nuestro caso, se requiere para la etapa de evaluación para lograr un dataset de prueba perfectamente equilibrado e imparcial. Reutilizamos una red GAN ya diseñada [12], pero la adaptamos a nuestro problema de interés[12]. Implementamos una arquitectura que combina un modelo generador y otro discriminador en uno solo de mayor tamaño (Figura 4.17).

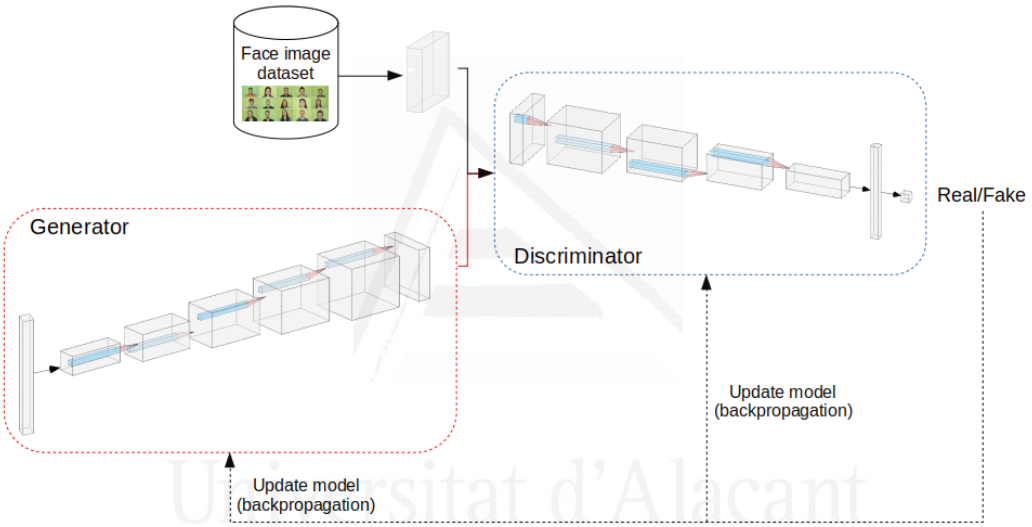


Figura 4.17: Arquitectura de la GAN.

La tarea del generador es producir imágenes sintéticas (falsas) a partir de vectores latentes aleatorios, mientras que el discriminador se encarga de decidir si estas imágenes pasan por reales. El nombre de toda la red se debe a que ambos modelos trabajan de forma adversaria. El entrenamiento del discriminador mejora la diferenciación de las imágenes, el resultado se utiliza para entrenar al generador con el fin de obtener imágenes progresivamente más parecidas al dataset original. Este proceso se realiza hasta que el discriminador asigna una imagen sintética como real.

Tras el entrenamiento de la red GAN utilizando las imágenes faciales de la categoría "asco", tenemos una muestra de los resultados obtenidos en la Figura 4.18. No todas las imágenes faciales son útiles. Aunque la apariencia es la de un rostro, ciertos rasgos

[12] https://github.com/cimejia/novel-FER-datasets/tree/main/FER-artificial-dataset/GAN

no se han formado completamente, por lo que debemos realizar un proceso de selección exhaustivo de las mejores muestras. Cabe destacar que el único mecanismo para controlar la expresión de las personas en las imágenes sintéticas es el contenido del dataset de entrada, por lo que las tareas de entrenamiento y selección de imágenes deben realizarse para cada categoría de emoción. Se trata de un proceso difícil y que requiere mucho tiempo (horas o días en función de la capacidad del hardware). Además, la calidad de las imágenes sintéticas resultantes está condicionada por las imágenes de entrada, mismas que son de baja resolución.

Figura 4.18: Muestra de imágenes faciales generadas por el modelo GAN para la categoría de asco.

Stable Diffusion. Los inconvenientes detectados en GAN motivan la búsqueda de una alternativa para crear imágenes artificiales. Recientemente, la *IA generativa* se ha desarrollado rápidamente y ha alcanzado gran relevancia con resultados impresionantes. En particular, los modelos de difusión se han convertido en el estado del arte en síntesis y superresolución de imágenes [95]. Entre las herramientas generativas más destacadas, *Stable Diffusion* permite generar imágenes digitales de alta calidad a partir de descrip-

ciones en lenguaje natural. Las imágenes generadas son similares a las utilizadas para entrenar el modelo.

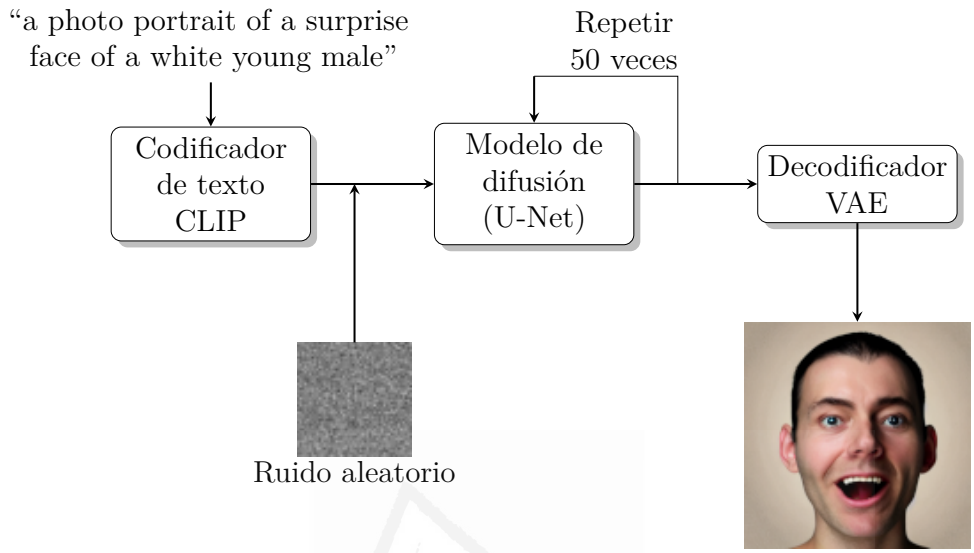


Figura 4.19: Arquitectura de Stable Diffusion.

La arquitectura de Stable Diffusion que se muestra en la Figura 4.19 está inspirada en aquella presentada en el sitio de Keras[13] y consta de un VAE (Variational AutoEncoder), un modelo de difusión basado en una red U-Net y un codificador de texto CLIP (Contrastive Language Image Pretraining). Cada componente dispone de su propia red neuronal. El funcionamiento puede explicarse en dos etapas: entrenamiento e inferencia. Ambas etapas utilizan representaciones latentes aprendidas por el VAE para codificar y decodificar las imágenes. Esta representación es una versión comprimida y probabilística que permite optimizar el proceso y generar variaciones de la imagen original. Durante el entrenamiento, la difusión añade progresivamente ruido a las imágenes de entrada para generar versiones ruidosas. Las imágenes originales y sus pares ruidosos constituyen el dataset para entrenar un modelo U-Net que asigne imágenes ruidosas a imágenes de alta calidad. En la inferencia, un ruido gaussiano aleatorio y una descripción textual (prompt) son las entradas del modelo. Antes, el prompt pasa por el codificador CLIP para generar el texto incrustado que condiciona el contenido visual. El modelo U-Net se encarga de predecir y eliminar el ruido (denoising), generando la salida basada en el prompt mediante capas de atención cruzada. Este proceso se repite un número determinado de veces (50 por defecto), en el que el ruido añadido se reduce gradualmente. Por último, la representación latente pasa por el decodificador VAE para obtener la imagen final.

[13] keras.io/examples/generative/random_walks_with_stable_diffusion

Stable Diffusion es desarrollado por *Stability AI* como código abierto y gratuito. A través de la empresa *Hugging Face*[14], están disponibles el código y los pesos[15], una versión Web[16], un cuaderno de programación[17] y la librería *diffusers* para su descarga e instalación. Utilizamos esta última soportada por la librería *Torch* e importamos *StableDiffusionPipeline* para instanciar la versión 1.4 del modelo. Se trata de un tipo de modelo de difusión preentrenado para visión y utilizado como herramienta de inferencia. No volvemos a entrenar el modelo en los datasets de FER, ya que fue entrenado originalmente con imágenes 512x512 de un subconjunto de LAION-5B[18], un dataset de 5,850 millones de pares imagen-texto [99]. El modelo se ejecuta en modo de inferencia con unas pocas líneas de código y una frase de texto (*prompt*) como argumento.

La generación de imágenes sintéticas adecuadas depende de la calidad de la indicación. Se trata de una indicación textual formada por varios tokens interpretados por la IA para convertirla en una imagen acorde con nuestras necesidades. Como no es una tarea trivial, existen herramientas de apoyo, por ejemplo, *Lexica*[19] es un buscador de imágenes generadas por Stable Diffusion que permite visualizar el prompt utilizado para dichas imágenes. Tras muchos experimentos con esta herramienta, sugerimos la siguiente estructura para conseguir buenos prompts.

Estilo + tópico principal + adjetivos + refuerzo + datos técnicos

1. El estilo es el tipo de imagen deseado, por ejemplo, ilustración, arte digital, fotografía, lienzo, caricatura, dibujo, pintura o retrato.

2. El tema principal debe describirse específicamente con los objetos que se observarán en la imagen. Esto se refiere a un paisaje, primer plano o vista general, y lo que debe ir en el centro como una persona, animal, planta o cualquier cosa.

3. Uno o varios adjetivos que definen con precisión las propiedades o atributos de los objetos.

4. Incluir palabras para expresar una acción, complementar o reforzar elementos con términos sinónimos o alternativos.

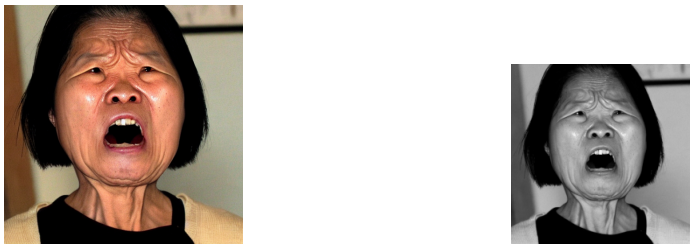
[14] https://huggingface.co/
[15] https://huggingface.co/CompVis/stable-diffusion
[16] https://huggingface.co/spaces/stabilityai/stable-diffusion
[17] https://colab.research.google.com/github/huggingface/notebooks/blob/main/diffusers/stable_diffusion.ipynb
[18] https://laion.ai/blog/laion-5b
[19] https://lexica.art/

5. Datos técnicos sobre la imagen y el grado de detalle para conseguir imágenes más realistas, por ejemplo, alta calidad, 4K o hd.

Las partes indicadas deben estar bien enlazadas, converger en un mismo concepto, sin incoherencias ni contradicciones y evitando divergencias. El orden influye en la importancia que el modelo da a estas partes, separándolas con una coma en lugar de con una frase larga que las agrupe a todas.

Utilizando las pautas sugeridas, creamos imágenes faciales con una emoción específica que se acercan más al resultado deseado. La Tabla 4.8 muestra las indicaciones utilizadas para producir las imágenes restantes para equilibrar nuestro dataset evaluador. Estas imágenes se obtienen sustituyendo la palabra “person” por la frase que especifica el género, la edad y la etnia, por ejemplo, la subcategoría "*female-old-asian*" de la categoría “enfado” del dataset NHFI no tenía imágenes faciales asociadas, por lo que las generamos especificando la frase "A detailed photographic portrait of a perfect face of an asian old female, feeling an extreme rage, expressing a very angry face, features well-defined, facing the camera, realistic, 4K, hd".

La Figura 4.20a muestra la imagen resultante que tiene el tamaño por defecto de 512x512 píxeles y color RGB. Convertimos a escala de grises 224x224 (Figura 4.20b) de acuerdo con NHFI. Así, se obtienen 212 imágenes faciales para equilibrar el dataset evaluador. Sin embargo, es necesario generar un mayor número de imágenes porque no todas ellas muestran claramente el tipo de emoción solicitada. Por este motivo, en la Tabla 4.8 se incluye la eficacia de los prompts, ya que algunas emociones son más complejas que otras. Las categorías de feliz y triste son las más fáciles de producir, 9 de cada 10 imágenes están bien generadas. Sigue la categoría neutral, con 8 de 10, mientras que las categorías de asco y enfado presentan la mayor complejidad, ya que sólo 3 y 4 imágenes de 10 son adecuadas, respectivamente.



(a) **(b)**

Figura 4.20: Imagen facial generada por Stable Diffusion para la categoría de enfado y la subcategoría de female-old-asian.

Tabla 4.8: Prompts utilizados para la generación de imágenes faciales de emociones con Stable Diffusion.

Categoría	Prompt	Eficacia
Enfado	"A detailed photographic portrait of a perfect face of a <person>, feeling an extreme rage, expressing a very angry face, features well-defined, facing the camera, realistic, 4K, hd"	40%
Asco	"A detailed photographic portrait of a perfect face of a <person>, feeling angry, with expression of very disgusted, forehead wrinkler, brow lowerer, cheek raiser, narrowed eyes, nose wrinkler, upper lip raiser, chin raiser, lip part, facing the camera, realistic, 4K, hd"	30%
Miedo	"A detailed photographic portrait of a perfect face of a <person>, expressing great dread, with facial gestures of fearful, fright, in panic, facing the camera, realistic, 4K, hd"	60%
Feliz	"A detailed photo portrait of a perfect whole front face of a <person>, expressing very happiness, facial gestures of happy, smiling, facial features well-defined, facing the camera, background any, ultra realistic, 4K, hd"	90%
Neutral	"A detailed photo portrait of a perfect face of a <person>, expressing neutrality, facial gestures of neutral, facial features well-defined, facing the camera, background any, ultra realistic, 4K, hd"	80%
Triste	"A detailed photo portrait of a perfect whole front face of a <person>, very sad face with tears, expressing extreme frustration, crying, facial features well-defined, facing the camera, background any, ultra realistic, 4K, hd"	90%
Sorpresa	"A detailed photographic portrait of a perfect face of a <person>, expressive face of surprise with the mouth open, extremely amazed, eyebrows very raised, upper eyelid raised, lips parted, jaw dropped, facial features well-defined, facing the camera, background any, realistic, 4K, hd"	50%

La Figura 4.21 muestra todas las imágenes generadas artificialmente para equilibrar el subconjunto de prueba para la categoría de enfado de NHFI. Son especialmente necesarios los rostros femeninos, de origen indio, ancianos y niños. Del mismo modo, generamos las imágenes en función del género, la edad y la etnia para los subconjuntos de prueba de FER2013 y AffectNet. Los tres subconjuntos de prueba se combinan para obtener un único dataset equilibrado e imparcial (Tabla 4.9), que resulta útil como punto de referencia para evaluar la capacidad de generalización de un modelo de reconocimiento de emociones en aplicaciones del mundo real.

Figura 4.21: Todas las imágenes faciales generadas por Stable Diffusion para la categoría de enfado.

Dataset	Enfado	Asco	Miedo	Feliz	neutral	Triste	Sorpresa	Total
NHFI	80	80	80	80	80	80	80	560
FER2013	80	80	80	80	80	80	80	560
AffectNet	80	80	80	80	80	80	80	560
Total	240	240	240	240	240	240	240	1680

Tabla 4.9: Distribución del dataset evaluador combinado, equilibrado e insesgado.

4.3.2.3 Creación del dataset de FER artificial

El buen rendimiento de Stable Diffusion en cuanto a calidad de las imágenes sintéticas y tiempo de generación nos motiva a contribuir con un dataset de FER totalmente artificial. El etiquetado es automático y está controlado por el prompt, que determina la categoría. Sin embargo, el equilibrio y el sesgo resultan difíciles de controlar debido a la eficacia variable de los prompts. Para las categorías de asco y sorpresa, es necesario incluir las unidades de acción correspondientes en sus prompts para mejorar la precisión de las imágenes sintéticas. Utilizamos el modelo de difusión para generar artificialmente miles de imágenes faciales considerando todas las categorías de emoción y los criterios de género, edad y etnia. La Tabla 4.10 presenta la distribución del dataset de FER artificial tras el proceso de selección de las mejores imágenes faciales.

La generación de cada imagen sintética implica el proceso de eliminación de ruido en 50 pasos, que es relativamente lento y consume mucha memoria, ya que tarda unos 11

Tabla 4.10: Distribución del dataset FER artificial generado con Stable Diffusion.

Dataset	Enfado	Asco	Miedo	Feliz	Neutral	Triste	Sorpresa	Total
Entrenamiento	580	944	990	1047	1109	1278	669	6617
Prueba	240	240	240	240	240	240	240	1680
Total	812	1129	1165	1253	1142	1353	768	8297

segundos con el hardware disponible, es decir, 5 por minuto y 300 por hora. Aunque las imágenes generadas se inspeccionan visualmente para seleccionar las más adecuadas y evitar las que son caricaturescas, demasiado distorsionadas, totalmente oscuras o con más de una persona, el uso de Stable Diffusion es mucho más eficiente, sencillo de utilizar y las imágenes son de alta resolución en comparación con GAN.

4.3.2.4 Selección del modelo

En las secciones anteriores, hemos preparado datasets de FER individuales y combinados mejorados para entrenar y evaluar modelos de DL de reconocimiento de emociones. Esto incluye también el dataset artificial. El objetivo es conseguir un modelo de reconocimiento de emociones con una mejor generalización a aplicaciones del mundo real. Para conocer si la generalización mejora, realizamos el entrenamiento y la evaluación de los modelos de DL en cada uno de los datasets analizados aquí. En esta sección, se describen las arquitecturas de modelos que serán entrenados en las versiones mejoradas de los datasets FER2013, NHFI, AffectNet, combinado y artificial. Una misma arquitectura puede tener un rendimiento diferente en función del dataset de entrenamiento. Esta dependencia significa que el rendimiento de un modelo no es necesariamente el mismo para diferentes datasets.

Por lo tanto, debemos identificar el mejor modelo para cada dataset. Basándonos en el método de refinamiento iterativo expuesto en la Sección 4.3.1, la CNN personalizada presentó el mejor rendimiento para FER2013 y AffectNet, mientras que la técnica de aprendizaje por transferencia basada en la versión B0 de EfficientNet [102] es la mejor alternativa para NHFI. Estas arquitecturas ya han sido representadas y analizadas en el mencionado apartado. Sólo resta el dataset artificial, para el cual la red MobileNetV2 [46] resultó ser la de mejor rendimiento luego de una serie de pruebas con varias arquitecturas, configuraciones e hiperparámetros. Se trata de un tipo de red convolucional de última generación reconocida en su momento por su nivel de velocidad y optimización. La Figura 4.22 presenta la arquitectura de aprendizaje por transferencia basada en esta red. Una ventaja es que se acepta como entrada la imagen con el tamaño original de 224x224 píxeles, que es procesada por la parte convolucional del modelo preentrenado para la

extracción de características. Las características extraídas son recibidas por el clasificador en forma de vector aplanado, que es la entrada para una red neuronal tradicional con dos capas densas de 256 y 512 neuronas, a las que se aplica la función de activación ReLU, además de la normalización por lotes y dropout para reducir el posible sobreajuste. La función Softmax de la última capa densa produce una distribución de probabilidad correspondiente a cada una de las 7 categorías de emoción a la que pertenece la imagen de entrada.

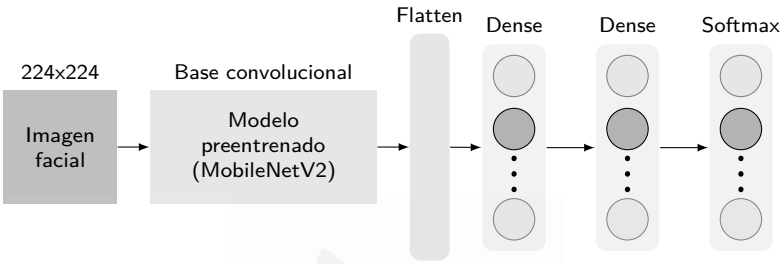


Figura 4.22: Arquitectura de la CNN basada en MobileNetV2 para el entrenamiento del dataset artificial.

4.4 Experimentación y resultados

Esta sección describe la parte experimental y consta de tres apartados. Comienza con la especificación del hardware y software. A continuación, se detallan los experimentos para mejorar los datasets individualmente. Se aplica un refinamiento iterativo mediante entrenamientos sucesivos del modelo de DL hasta lograr una versión reclasificada de cada dataset. Los resultados se evalúan utilizando curvas de aprendizaje, matrices de confusión y la precisión de validación como métrica de comparación del dataset original y su nueva versión. Por último, realizamos una serie de experimentos single- y cross-dataset. Explicamos cómo se organizan las versiones mejoradas de los datasets, su combinación y el dataset artificial para entrenar sus respectivos modelos de DL. La evaluación basada en los datos de prueba del mismo dataset y de los demás, nos permite conocer la calidad y la generalización de los distintos datasets en el reconocimiento de emociones.

4.4.1 Herramientas tecnológicas

Las principales características de la plataforma computacional utilizada son: procesador Intel(R) Core(TM) i9-7920X a 2.90GHz, memoria RAM de 64 GB, GPU NVIDIA GeForce RTX2080 con 12 GB de RAM y sistema operativo Linux Ubuntu 18.04.5 LTS.

Realizamos la implementación con el lenguaje de programación *Python* versión 2.7.17, con el soporte de las librerías *TensorFlow* y *Keras* para la creación de las CNNs y *PyTorch* para la ejecución del modelo de difusión. *Sklearn* para las métricas y matriz de confusión. Librerías estándar como *OS*, *NumPy* y*Matplotlib*, para la gestión de directorios y archivos, arreglos numéricos y visualización de gráficos, respectivamente. La utilidad *ImageDataGenerator* para el preprocesamiento de imágenes, la división automática en conjuntos de entrenamiento y validación, y la normalización de los píxeles.

4.4.2 Refinamiento iterativo

4.4.2.1 Entrenamiento

Este proceso tiene como objetivo que el modelo aprenda a asociar las imágenes faciales y sus categorías de emoción. Los hiperparámetros deben ser definidos explícitamente antes del entrenamiento. No hay reglas fijas para determinar estos valores, son el resultado de varias pruebas para encontrar los más convenientes. La Tabla 4.11 muestra estos valores para cada modelo y dataset. Cabe indicar que los hiperparámetros se mantienen para todos los experimentos.

Tabla 4.11: Hiperparámetros utilizados en el proceso de refinamiento de cada dataset.

Hiperparámetro	**FER2013**	**NHFI**	**AffectNet**
Entrada	48, 48, 1	224, 224, 3	48, 48, 3
Entrenamiento-validación (%)	80-20	80-20	80-20
Tamaño de lote	64	64	64
Tasa de aprendizaje	0.01 to 0.00001	0.01 to 0.00001	0.001 to 0.00001
Optimizador	Adam	Adam	Adam
Función de pérdida	categorical_crossentropy	categorical_crossentropy	categorical_crossentropy
Métricas	loss & accuracy	loss & accuracy	loss & accuracy
Número de clases	7	7	7
Épocas	100	50	50
Aumento de datos	Yes	No	No
Cantidad de entrenamientos	5	5	5

El método de refinamiento demandó cinco entrenamientos sucesivos para cada uno de los datasets hasta cumplir los criterios definidos de calidad. En cada entrenamiento, la red convolucional recibe como entrada las imágenes faciales del subconjunto de entrenamiento de cada dataset en lotes de 64 imágenes. Utilizamos un generador de imágenes por lotes como *ImageDataGenerator* de Keras debido a que el gran número y tamaño de las imágenes puede provocar un problema de memoria si las proporcionamos en una sola estructura de almacenamiento. También nos permite pasar las imágenes desde los directorios etiquetando automáticamente la imagen con la categoría respectiva y realizar el aumento de datos en caso de ser necesario. Para cada lote, se comparan las etiquetas

predichas y las reales, obteniendo un error o *pérdida* y una *precisión* mediante la función *categorical_crossentropy*. Se aplican los algoritmos de retropropagación y *Adam*, ambos basados en el descenso de gradiente, para actualizar los pesos del modelo en función del valor de la *tasa de aprendizaje*. Cuando se completan todos los lotes, se realiza una *época*, es decir, una iteración de todas las imágenes de entrenamiento. Los valores de precisión y pérdida se miden después de cada época utilizando las imágenes del subconjunto de validación. Para conocer el máximo de precisión, se han ejecutado 100 épocas ara FER2013, mientras que para NHFI y AffectNet, 50 épocas han sido suficientes, ya que por encima de estos valores, el comportamiento del modelo permanece prácticamente estable y no es apreciable una mejora. Se aprovecha la utilidad *callback* para realizar ciertas acciones durante el entrenamiento como establecer un punto de control y reducir la tasa de aprendizaje. El modelo sólo se guarda si la precisión de validación en la época actual es mayor que la de la última época. Por otro lado, la tasa de aprendizaje nos dice cuánto se actualizarán los pesos cada vez, y suele estar entre 0 y 1. Disminuirá desde un valor inicial hasta un mínimo si la pérdida no se reduce después de un cierto número de épocas, lo que resulta en un mejor entrenamiento.

4.4.2.2 Evaluación

Los resultados de la experimentación se presentan gráficamente mediante curvas de aprendizaje y matrices de confusión, mientras que la métrica numérica utilizada para la comparación es la precisión de validación. Estas herramientas permiten evaluar el rendimiento del modelo y la mejora del dataset. Durante el entrenamiento y la validación de cada modelo, se han recogido los valores de pérdida y precisión, respectivamente. Esto genera las denominadas *curvas de aprendizaje*, donde el eje horizontal representa el número de épocas mientras que el eje vertical puede representar la precisión o el error. La *matriz de confusión*, también conocida como *matriz de error*, es una tabla para visualizar el rendimiento del modelo, ya que presenta información sobre las clasificaciones reales y predichas realizadas por un modelo clasificador. Las filas representan las instancias de clases reales, mientras que las columnas representan las instancias que el clasificador predice [39]. A partir de esta matriz, se pueden obtener varias métricas de rendimiento, sin embargo, nos centramos en la *precisión*, que compara el número de predicciones correctas (en la diagonal) dividido por el número total. A continuación, se presentan los resultados obtenidos para cada uno de los datasets analizados.

4.4.2.3 FER2013

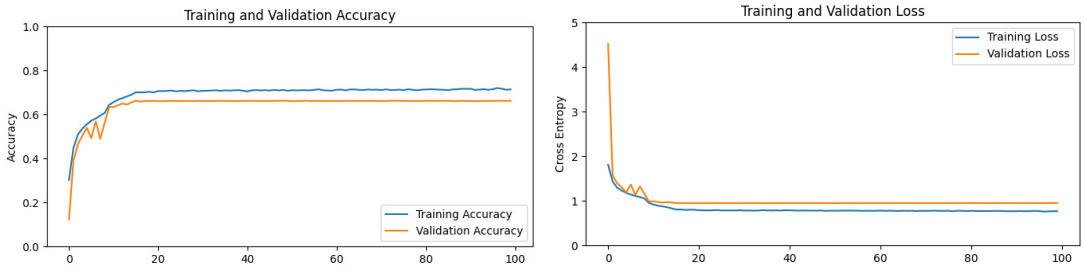
Las curvas de aprendizaje y la matriz de confusión de cada uno de los cinco entrenamientos en este dataset se muestran en la Figura 4.23. Para cada entrenamiento (incluida la validación) se presentan: las curvas de precisión (izquierda), las curvas de pérdida (centro) y la matriz de confusión correspondiente (derecha). A medida que se realizan más entrenamientos, las curvas de precisión (entrenamiento y validación) alcanzan valores más altos, mientras que las curvas de pérdida van disminuyendo en altura y se acercan a cero. Además, los pares de curvas están muy próximos entre sí. Por lo tanto, la precisión del modelo es mayor, el error es menor y no hay sobreajuste. Este comportamiento es ideal y se debe al filtrado sucesivo del dataset. Las matrices de confusión incluyen las predicciones del tipo de emoción para todas las imágenes del dataset. Los entrenamientos progresivos provocan el efecto deseado en cada matriz, es decir, reducir los valores fuera de la diagonal principal y aumentar los valores en esta diagonal. El modelo es cada vez más preciso porque las predicciones erróneas se descartan en los entrenamientos posteriores. Como resultado, se capturan más características distintivas de cada clase. De este modo, disminuye la variabilidad intraclase de las imágenes faciales y aumenta la variabilidad interclase.

El proceso de refinamiento del dataset se resume en la Tabla 4.12. Fueron necesarios cinco entrenamientos (cuatro operaciones de filtrado) en FER2013 para alcanzar la métrica de rendimiento esperada (precisión de validación). Otro entrenamiento no se considera porque no produjo una mejora significativa de la precisión. El número de imágenes disminuye gradualmente, pero sigue siendo considerable en cada entrenamiento. El modelo con mayor precisión (97.7%) ha captado los rasgos más distintivos de cada categoría de emoción y es conveniente para reclasificar todas las imágenes del dataset. El método *predict()* se utiliza para asignar la categoría de cada imagen facial del dataset original, generando una nueva distribución de FER2013. La comparación se presenta en la Tabla 4.13.

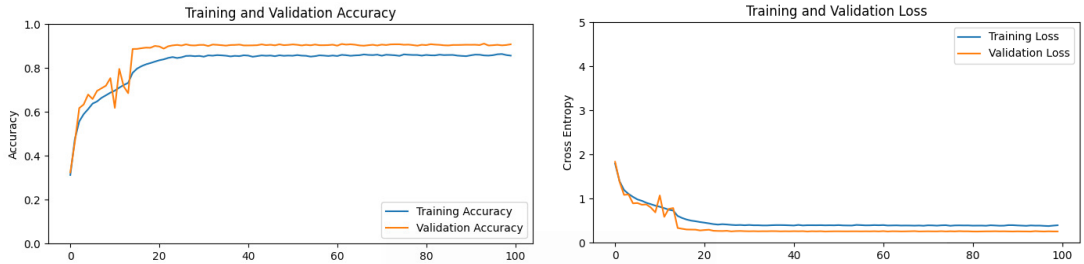
Tabla 4.12: Resumen de los resultados experimentales del dataset FER2013.

Entrenamiento	Imágenes (entrenamiento)	Imágenes (validación)	Total	Precisión
1	28708	7178	35886	0.6702
2	25415	4810	30225	0.9089
3	24001	4379	28380	0.9582
4	23654	4179	27833	0.9761
5	23488	4079	27567	0.9770

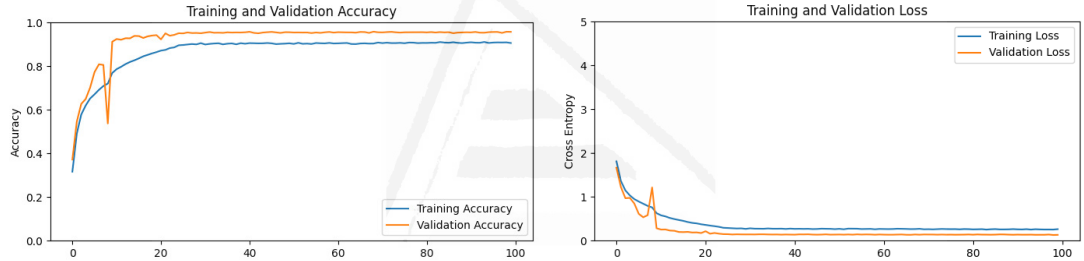
La Figura 4.25 muestra que las categorías “asco” y “triste” tienen una variación mínima, las de “enfado”, “feliz” y “sorpresa” varían moderadamente, y las categorías



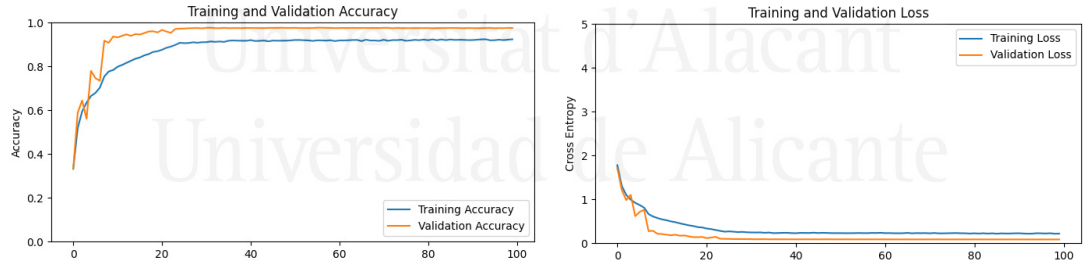
(a) Entrenamiento #1



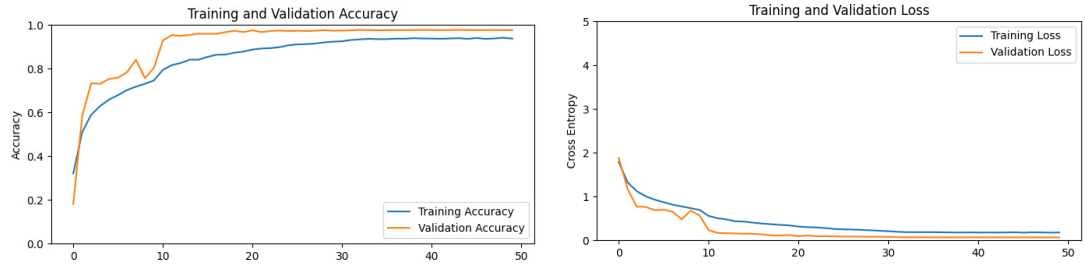
(b) Entrenamiento #2



(c) Entrenamiento #3

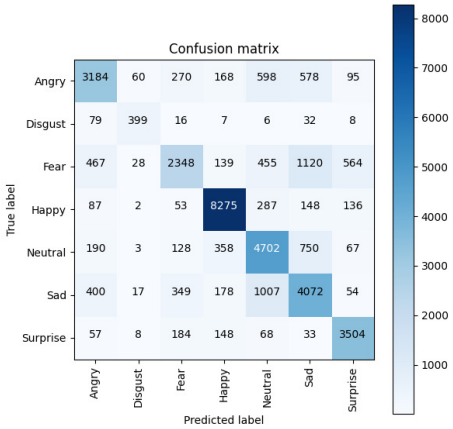


(d) Entrenamiento #4

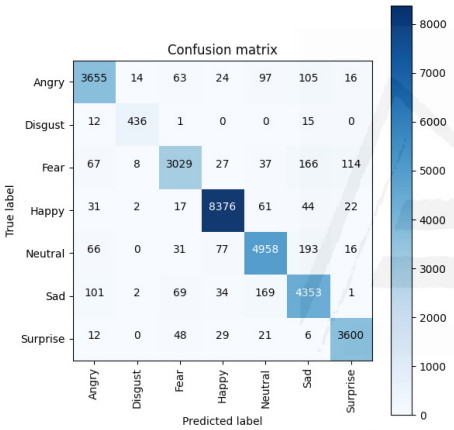


(e) Entrenamiento #5

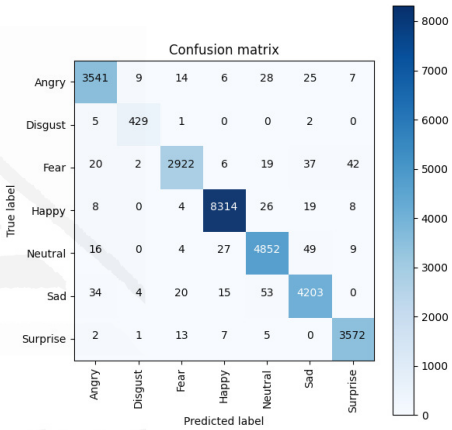
Figura 4.23: Curvas de aprendizaje de cinco entrenamientos sucesivos del dataset FER2013.



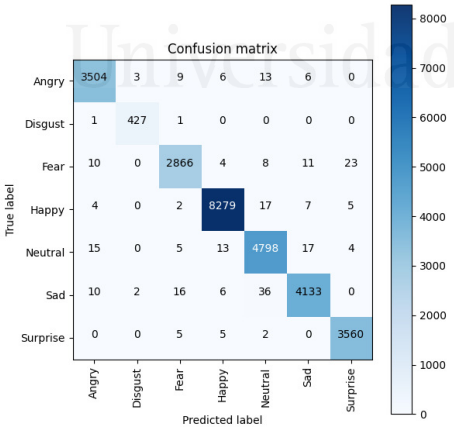
(a) Entrenamiento #1



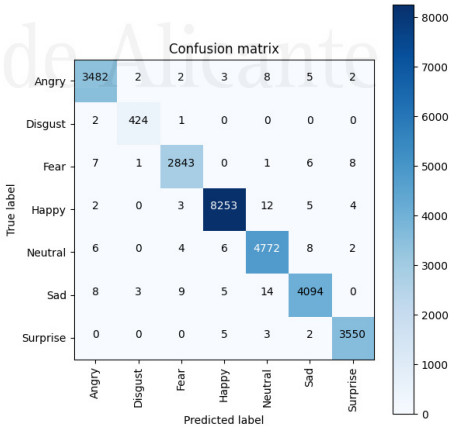
(b) Entrenamiento #2



(c) Entrenamiento #3



(d) Entrenamiento #4



(e) Entrenamiento #5

Figura 4.24: Matrices de confusión de cinco entrenamientos sucesivos del dataset FER2013.

más afectadas son "miedo" (decreciente) y "neutral" (creciente), lo que indica que el dataset FER2013 original adolece de imágenes faciales mal clasificadas, especialmente entre estas dos categorías.

Tabla 4.13: Distribución del dataset FER2013 original y reclasificado.

Dataset	Enfado	Asco	Miedo	Feliz	Neutral	Triste	Sorpresa	Total
FER2013 (original)	4953	547	5121	8988	6198	6077	4002	35886
FER2013 (reclasificado)	4817	532	3842	9202	7074	6090	4329	35886

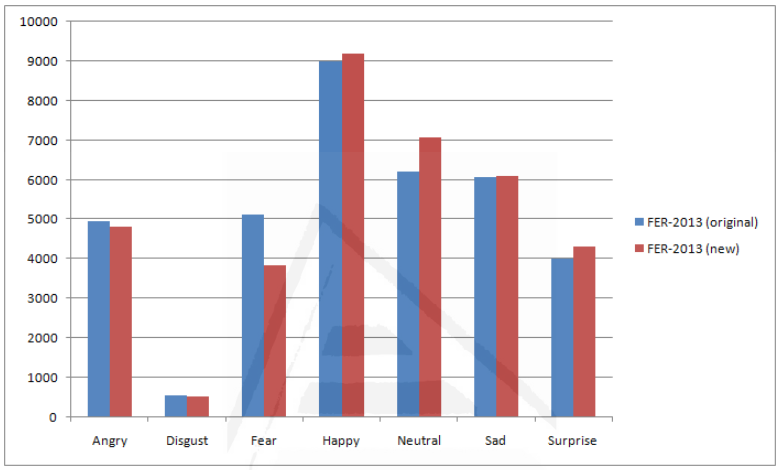


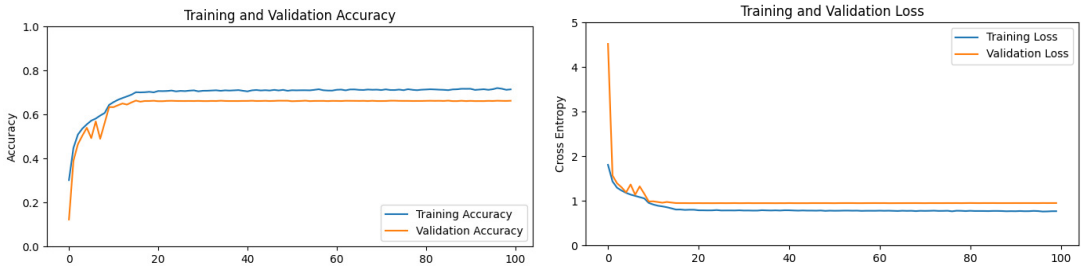
Figura 4.25: Comparación gráfica de ambas distribuciones.

La prueba decisiva de la eficacia de nuestro método consiste en entrenar la misma CNN en el dataset reclasificado de FER2013. La Figura 4.26 muestra que se obtienen mejores curvas de aprendizaje y la matriz de confusión indica más predicciones correctas y menos incorrectas. Una mayor precisión y una menor pérdida se verifican en la Tabla 4.14.

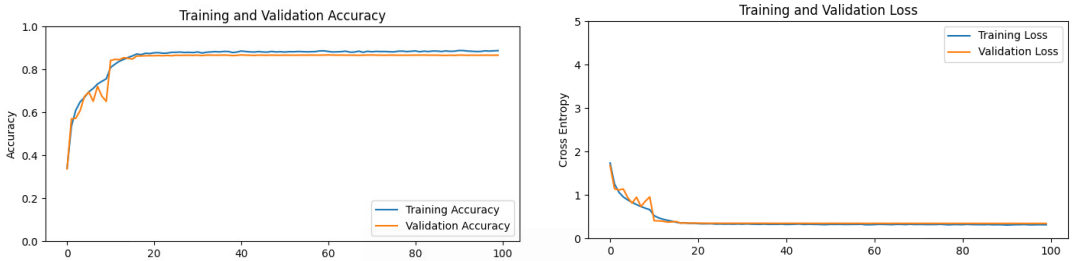
Tabla 4.14: Comparación de los resultados de entrenamiento de los datasets FER2013 original y reclasificado.

Dataset	Imágenes (entrenamiento)	Imágenes (validación)	Total	Precisión
FER2013 (original)	28708	7178	35886	0.6626
FER2013 (reclasificado)	28708	7178	35886	0.8671

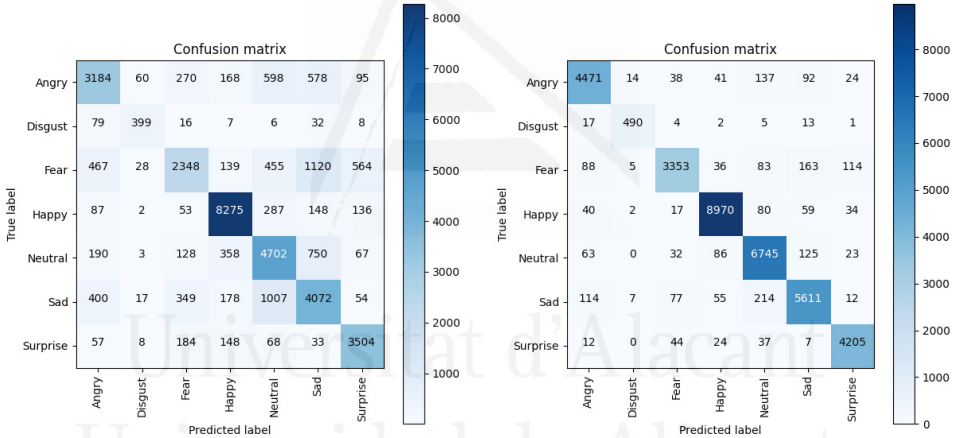
Los resultados confirman un nuevo dataset más fiable que mantiene el número de imágenes. La versión reclasificada de FER2013 permite un aumento muy significativo de la precisión de validación del modelo en un 20.45% y la pérdida es mucho menor (0.34). La precisión de entrenamiento es aceptable (88.76%), muy cercana a la precisión de validación y la pérdida es menor. No hay sobreajuste ni diferencias significativas



(a) Curvas de aprendizaje de FER2013 original.



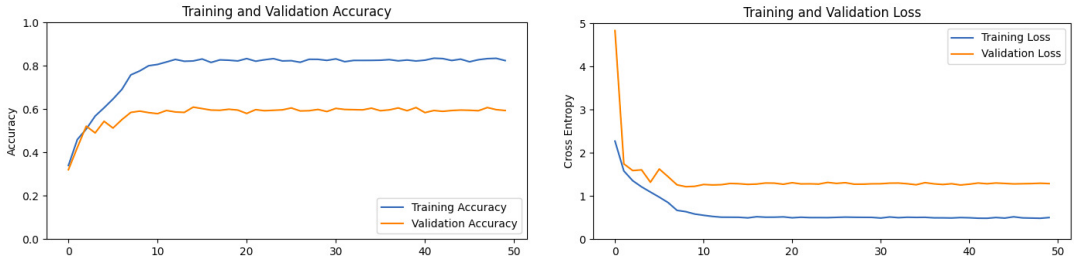
(b) Curvas de aprendizaje de FER2013 reclasificado.



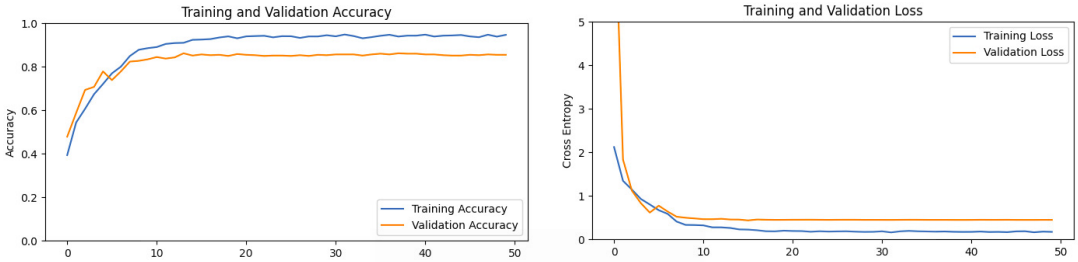
(c) Matrices de confusión: FER2013 original (izq.) y FER2013 reclasificado (der.).

Figura 4.26: Comparación entre el dataset FER2013 original y el reclasificado.

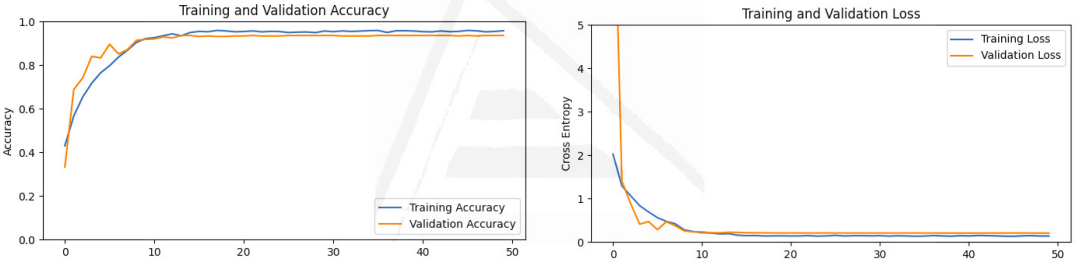
entre la pérdida de entrenamiento y la de validación. Todas las categorías muestran una mejora de la precisión, en particular, hay una mejora notable para “enfado” (un aumento del 26%), “miedo” (un aumento del 38%) y “triste” (un aumento del 25%), es decir, las que mostraban mayor solapamiento o confusión. Según los experimentos, sólo 40 épocas en cada entrenamiento serían suficientes, ya que el comportamiento permanece prácticamente estable más allá de este número.



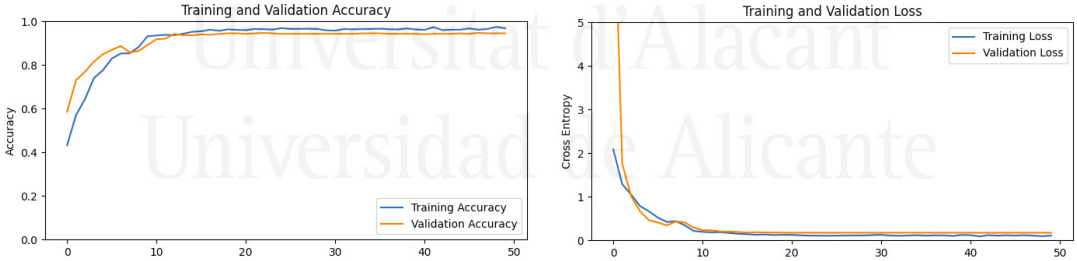
(a) Entrenamiento #1



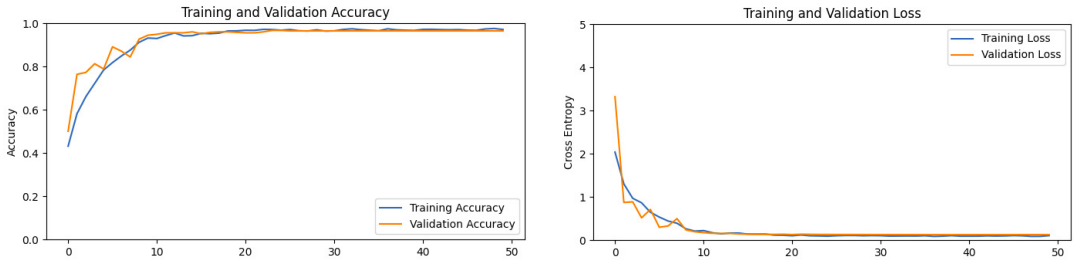
(b) Entrenamiento #2



(c) Entrenamiento #3

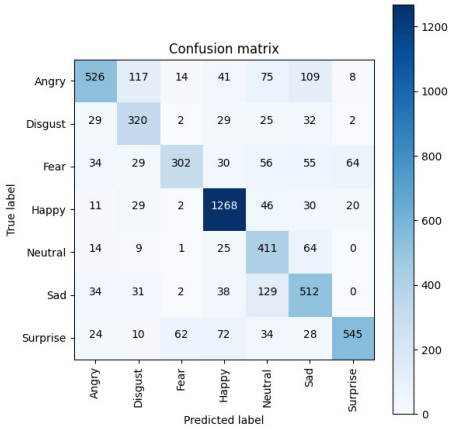


(d) Entrenamiento #4

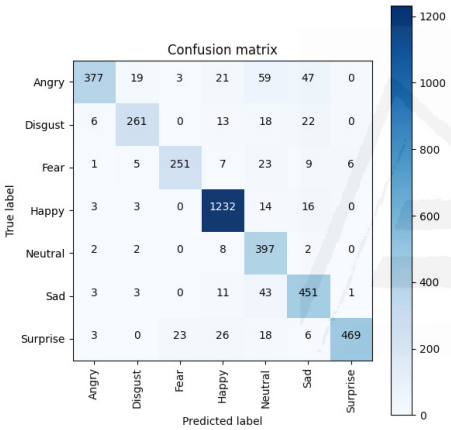


(e) Entrenamiento #5

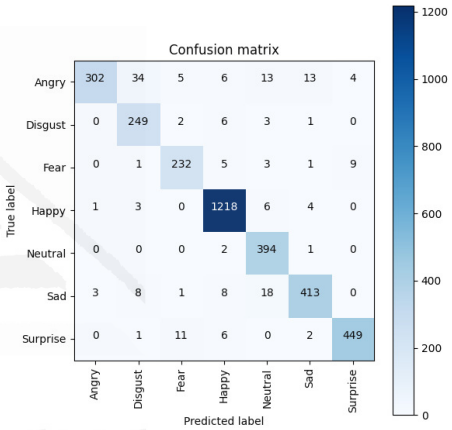
Figura 4.27: Curvas de aprendizaje de cinco entrenamientos sucesivos del dataset NHFI.



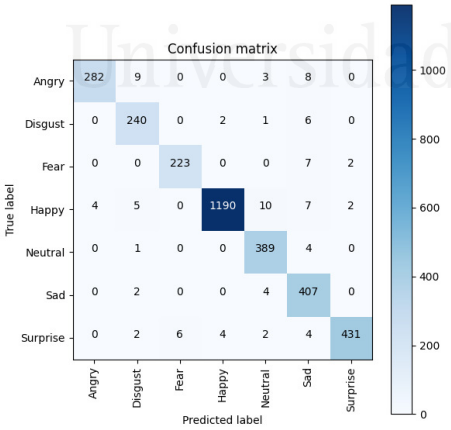
(a) Entrenamiento #1



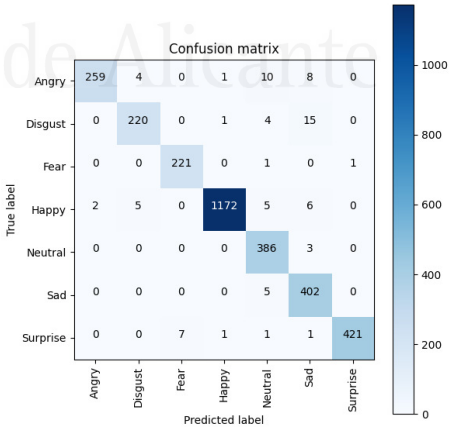
(b) Entrenamiento #2



(c) Entrenamiento #3



(d) Entrenamiento #4



(e) Entrenamiento #5

Figura 4.28: Matrices de confusión de cinco entrenamientos sucesivos del dataset NHFI.

4.4.2.4 NHFI

Las curvas de precisión del dataset NHFI (izquierda de la Figura 4.27) comienzan bastante separadas entre sí, lo que evidencia la presencia de sobreajuste, pero a medida que se realizan los entrenamientos, las curvas se acercan y alcanzan una precisión elevada. Lo mismo ocurre con las curvas de pérdida, pero en sentido contrario, acercándose cada vez más al eje horizontal. Las matrices de confusión muestran valores más altos en la diagonal principal y valores más bajos fuera de esta diagonal, lo que indica la mejora progresiva de la precisión del modelo, así como la calidad del dataset utilizado en cada entrenamiento. A pesar del sucesivo descarte de predicciones incorrectas, el número de imágenes es significativo respecto a la cantidad original. La Tabla 4.15 muestra la evolución de los entrenamientos sobre el dataset NHFI.

Tabla 4.15: Resumen de los resultados experimentales del dataset NHFI.

Entrenamiento	Imágenes (entrenamiento)	Imágenes (validación)	Total	Precisión
1	4278	1072	5350	0.5597
2	3284	600	3884	0.8367
3	2936	502	3438	0.9382
4	2786	471	3257	0.9533
5	2713	449	3162	0.9666

La reclasificación del dataset NHFI original se realiza con el modelo de mayor precisión (96.66%). Se genera una nueva distribución que se muestra en la Tabla 4.16. La Figura 4.29 evidencia que las categorías de enfado y neutral tuvieron los mayores cambios, lo que indica que estas categorías tienen la mayor variabilidad intraclase en el dataset original.

Tabla 4.16: Distribución del dataset NHFI original y reclasificado.

Dataset	Enfado	Asco	Miedo	Feliz	Neutral	Triste	Sorpresa	Total
NHFI (original)	890	439	570	1406	524	746	775	5350
NHFI (reclasificado)	336	514	383	1585	1042	962	528	5350

Para demostrar la mejora del modelo en el reconocimiento, se entrena la misma CNN en el dataset NHFI reclasificado y se compara el resultado con el dataset original (Figura 4.30). El sobreajuste no se redujo, pero la precisión es mayor, tanto en el subconjunto de entrenamiento como en el de validación. La pérdida disminuye para el dataset NHFI reclasificado, así como los valores fuera de la diagonal de la matriz de confusión.

Los resultados de rendimiento para las distribuciones original y reclasificada de NHFI se presentan en la Tabla 4.17. Hemos conseguido aumentar significativamente la precisión tanto en el conjunto de entrenamiento como en el de validación, en un 18.74% y un 14.47%, respectivamente. Exceptuando las categorías “enfado” y “feliz”, la precisión de

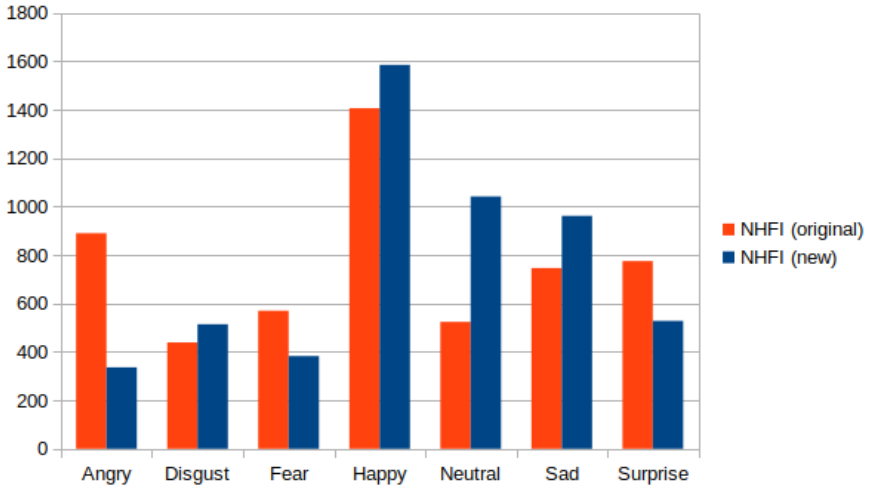


Figura 4.29: Comparación gráfica de ambas distribuciones.

Tabla 4.17: Comparación entre el dataset NHFI original y el reclasificado.

Dataset	Imágenes (entrenamiento)	Imágenes (validación)	Total	Precisión (entrenamiento)	Precisión (validación)
NHFI (original)	4278	1072	5350	0.7676	0.5597
NHFI (reclasificado)	4278	1072	5350	0.9550	0.7044

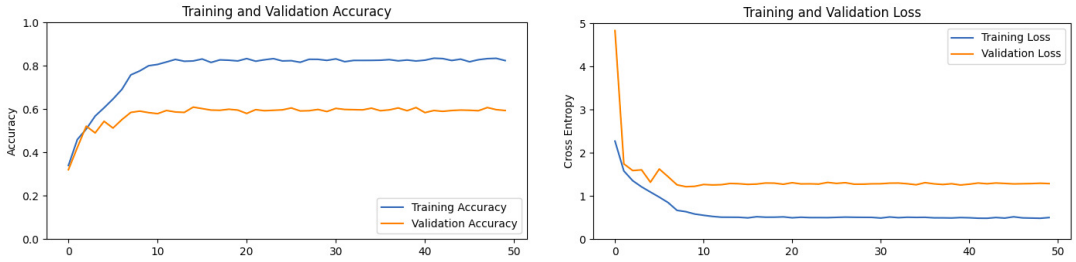
validación es altamente incrementada en el resto de categorías, particularmente en la categoría "triste" del 49 al 85%. La metodología basada en el filtrado sucesivo con un modelo de transfer learning se ha aplicado con éxito en un dataset distinto a FER2013.

4.4.2.5 AffectNet

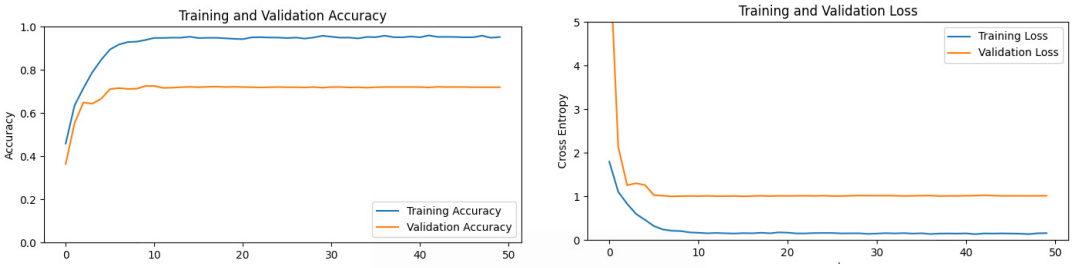
La versión de AffectNet que seleccionamos contiene 287401 imágenes con un gran desequilibrio entre las categorías (Figura 4.1c). El entrenamiento en este dataset puede dar lugar a sesgos y a una evaluación errónea de la precisión del modelo. Por lo tanto, aplicamos *downsampling* para equilibrar todas las categorías considerando la que tiene el menor número de imágenes. La categoría "asco" limitó a las demás con 4300 imágenes, de las cuales 3800 se han seleccionado aleatoriamente para el entrenamiento utilizando la librería *splitfolders*, mientras que 500 imágenes por defecto vienen con el dataset para la validación. La versión equilibrada se muestra en la Tabla 4.18.

Tabla 4.18: Distribución equilibrada del dataset AffectNet.

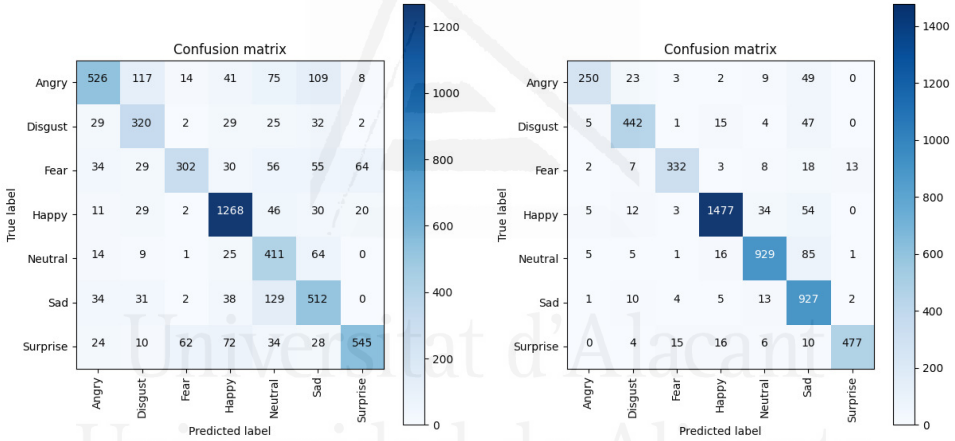
Conjunto	Enfado	Asco	Miedo	Feliz	Neutral	Triste	Sorpresa	Total
Entrenamiento	3800	3800	3800	3800	3800	3800	3800	26600
Validación	500	500	500	500	500	500	500	3500
Total	4300	4300	4300	4300	4300	4300	4300	30100



(a) Curvas de aprendizaje de NHFI original.



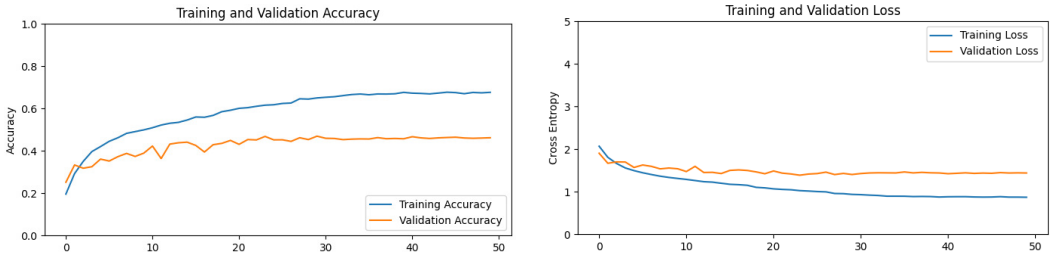
(b) Curvas de aprendizaje de NHFI reclasificado.



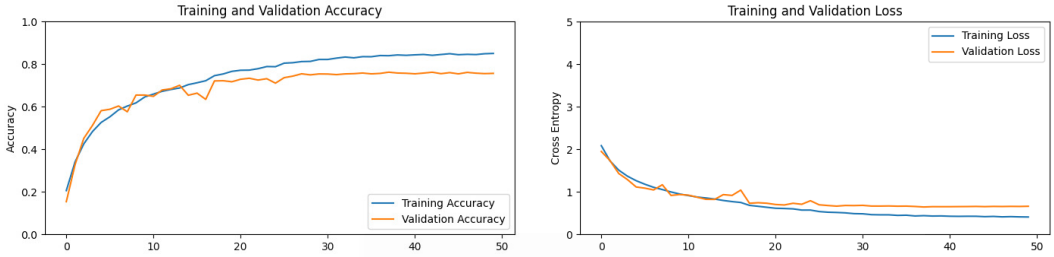
(c) Matrices de confusión: NHFI original (izq.) y NHFI reclasificado (der.).

Figura 4.30: Comparación entre el dataset NHFI original y el reclasificado.

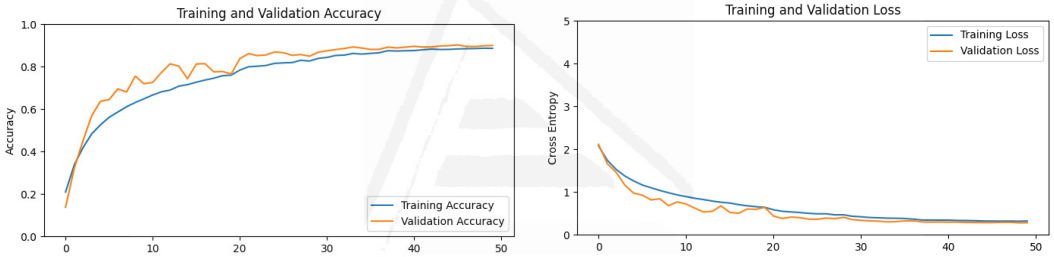
El proceso de refinamiento se realiza en esta versión equilibrada de AffectNet. Las curvas de precisión para los conjuntos de entrenamiento y validación (Figura 4.31), comienzan con una pequeña separación, que va disminuyendo a medida que se realizan sucesivos entrenamientos, incluso la curva de validación termina superando en precisión a la de entrenamiento. El mismo comportamiento, pero en sentido contrario, se presenta para las curvas de pérdidas. Los valores de la diagonal principal de la matriz de confusión aumentan con cada entrenamiento y disminuyen fuera de esta diagonal, indicando una mayor precisión del modelo debido a un mejor dataset. La evolución de los sucesivos entrenamientos se resume en la Tabla 4.19.



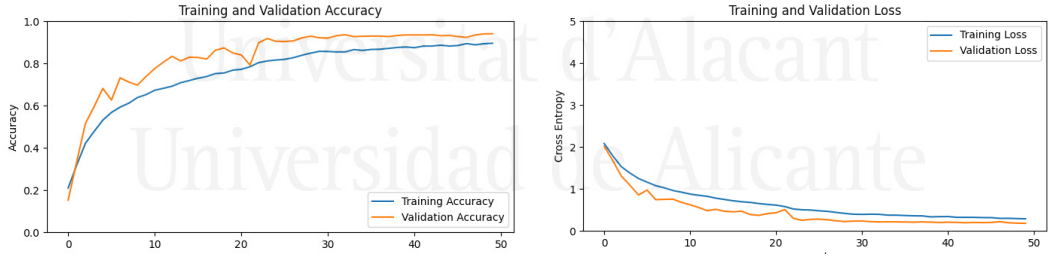
(a) Training #1



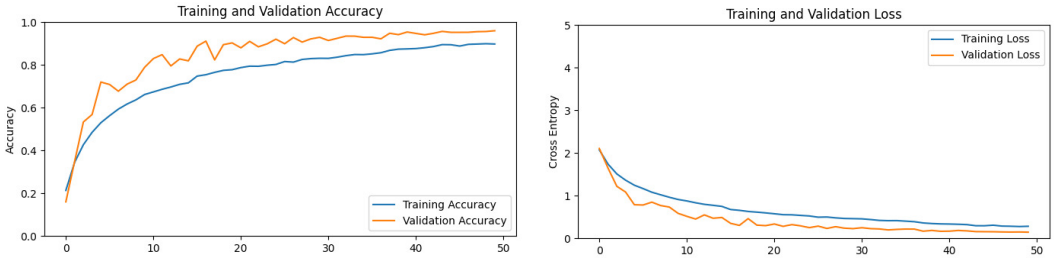
(b) Training #2



(c) Training #3

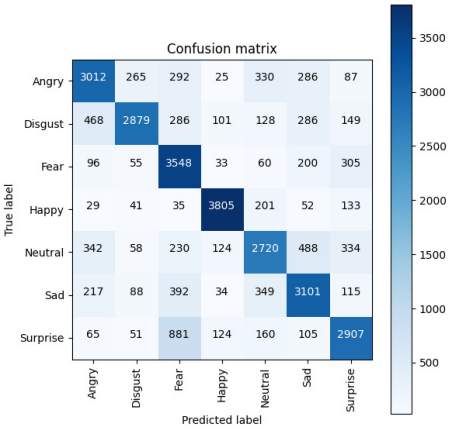


(d) Training #4

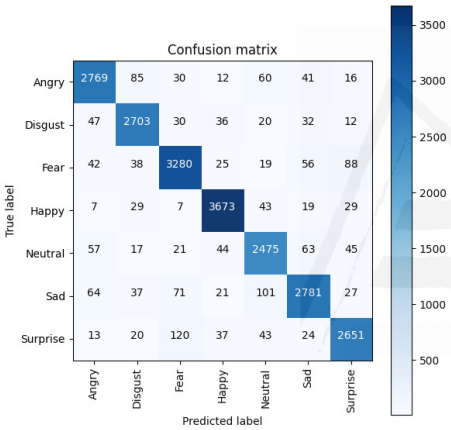


(e) Training #5

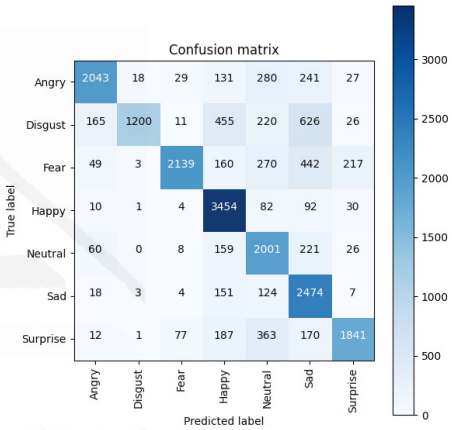
Figura 4.31: Curvas de aprendizaje de cinco entrenamientos sucesivos de AffectNet.



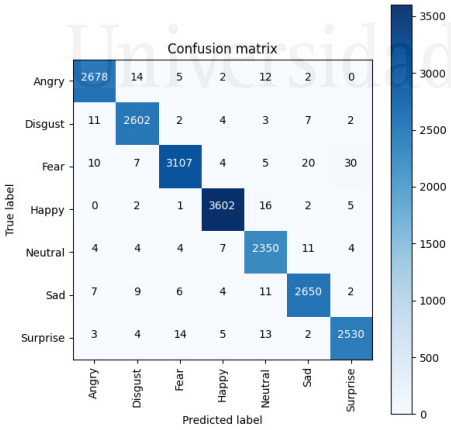
(a) Entrenamiento #1



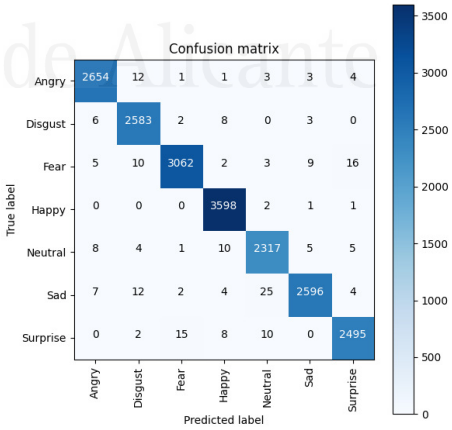
(b) Entrenamiento #2



(c) Entrenamiento #3



(d) Entrenamiento #4



(e) Entrenamiento #5

Figura 4.32: Matrices de confusión de cinco entrenamientos sucesivos del dataset AffectNet.

Tabla 4.19: Resumen de los resultados del dataset equilibrado de AffectNet.

Entrenamiento	Imágenes (entrenamiento)	Imágenes (validación)	Total	Precisión
1	26600	3500	30100	0.4686
2	17587	4393	21980	0.7612
3	16268	4064	20332	0.9016
4	15843	3956	19799	0.9401
5	15617	3902	19519	0.9590

El modelo del último entrenamiento alcanza la mayor precisión de validación (95.9%), lo que nos permite reclasificar el dataset equilibrado. Se genera una nueva distribución de las 30100 imágenes que se presenta en la Tabla 4.20.

Tabla 4.20: Distribución del dataset AffectNet equilibrado y reclasificado.

Dataset	Enfado	Asco	Miedo	Feliz	Neutral	Triste	Sorpresa	Total
AffectNet (equilibrado)	4300	4300	4300	4300	4300	4300	4300	30100
AffectNet (reclasificado)	4394	3893	5004	4594	4224	4071	3920	30100

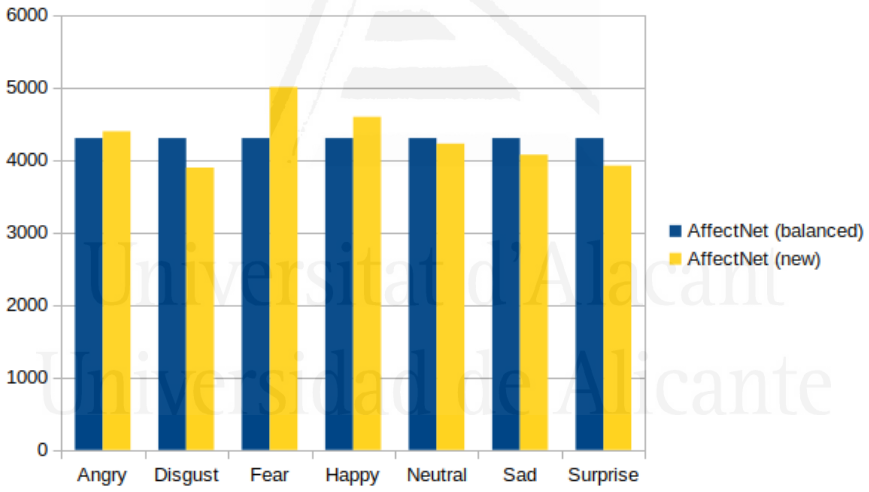
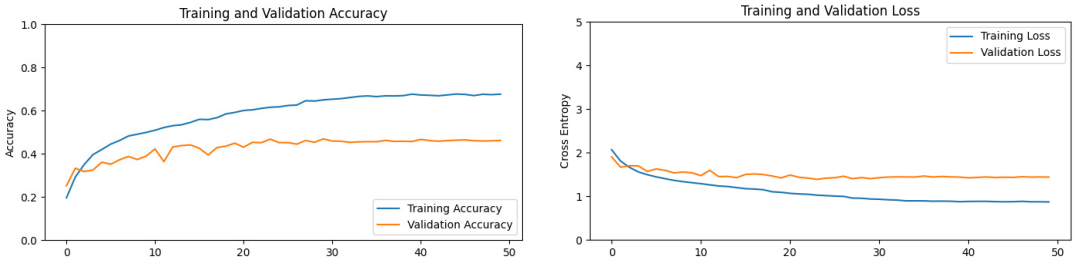


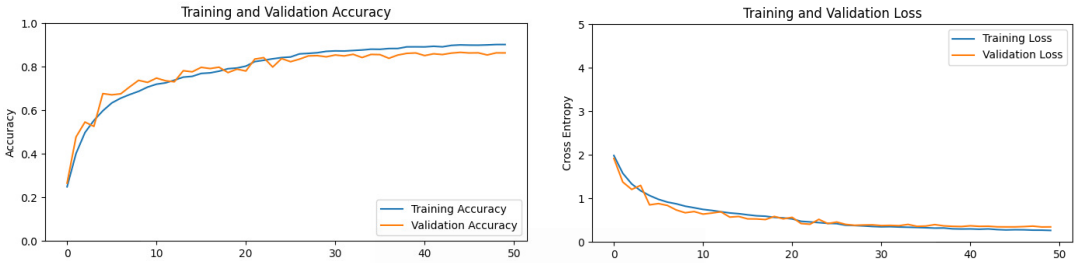
Figura 4.33: Comparación gráfica de ambas distribuciones.

Tras la reclasificación del dataset equilibrado, la nueva distribución se vuelve desequilibrada (Figura 4.33). Las categorías de feliz y miedo han aumentado significativamente, mientras que la categoría de enfado ha aumentado ligeramente. En el resto de los casos, se produce una disminución, principalmente en las categorías de asco y sorpresa.

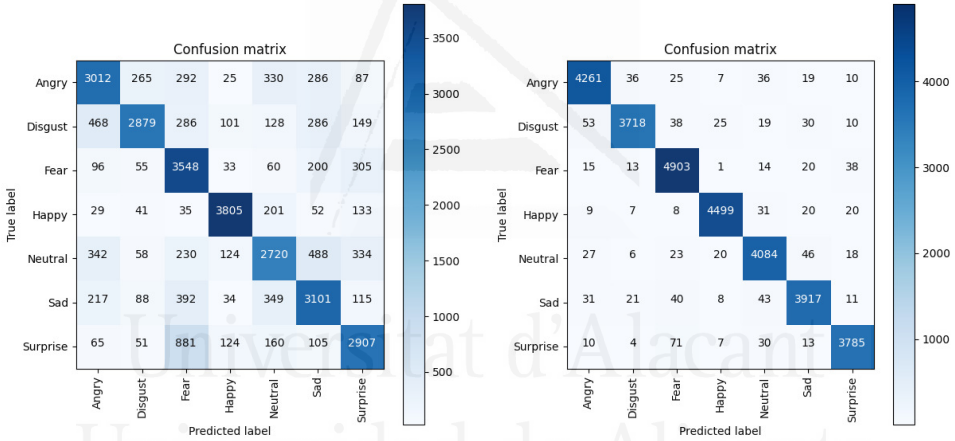
A continuación, el modelo basado en CNN se entrena en la nueva versión de AffectNet para verificar que nuestro método es eficaz. La Figura 4.34 presenta las curvas de aprendizaje de ambas versiones del dataset, donde la nueva AffectNet (Figura 4.34b) permite alcanzar un mejor rendimiento del modelo y mayor precisión.



(a) Curvas de aprendizaje de AffectNet equilibrado.



(b) Curvas de aprendizaje de AffectNet reclasificado.



(c) Matrices de confusión de AffectNet equilibrado (izq.) y reclasificado (der.)

Figura 4.34: Comparación entre el dataset AffectNet equilibrado y reclasificado.

Tabla 4.21: Comparación de los resultados de entrenamiento de los datasets AffectNet equilibrado y reclasificado.

Dataset	Imágenes (entrenamiento)	Imágenes (validación)	Total	Precisión (entrenamiento)	Precisión (validación)
AffectNet (equilibrado)	26600	3500	30100	0.6763	0.4686
AffectNet (reclasificado)	24084	6016	30100	0.9013	0.8652

Se observa una notable mejora de la precisión en comparación con el primer entrenamiento del dataset equilibrado (Tabla 4.21). Debido a la reducción del muestreo, la proporción de división es del 88% y el 12%, para los conjuntos de entrenamiento y

validación, respectivamente. Para la nueva versión del dataset, la proporción es del 80% y el 20%, y al disponer de más imágenes de validación, el porcentaje de precisión casi se duplica (un incremento de 39.66%).

Nuestro método ha sido aplicado con éxito a una versión más pequeña y equilibrada del dataset AffectNet. El objetivo es mejorar el dataset AffectNet original, que es más grande y está desequilibrado. Para ello, se utiliza el último modelo entrenado para reclasificar las imágenes faciales en la versión completa de AffectNet. La nueva distribución se presenta en la Tabla 4.22.

Tabla 4.22: Distribución del dataset AffectNet original y nuevo.

Dataset	Enfado	Asco	Miedo	Feliz	Neutral	Triste	Sorpresa	Total
AffectNet (original)	25382	4303	6878	134915	75374	25959	14590	287401
AffectNet (nuevo)	30827	17475	19145	114275	52760	29160	23759	287401

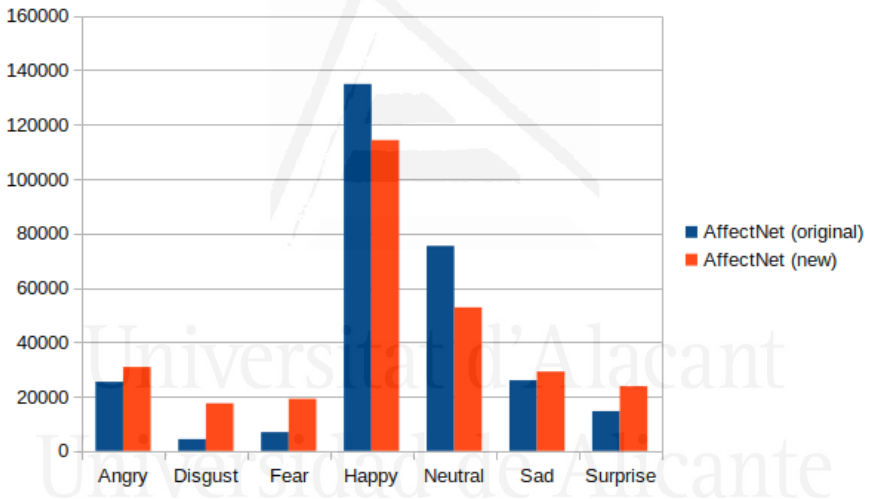
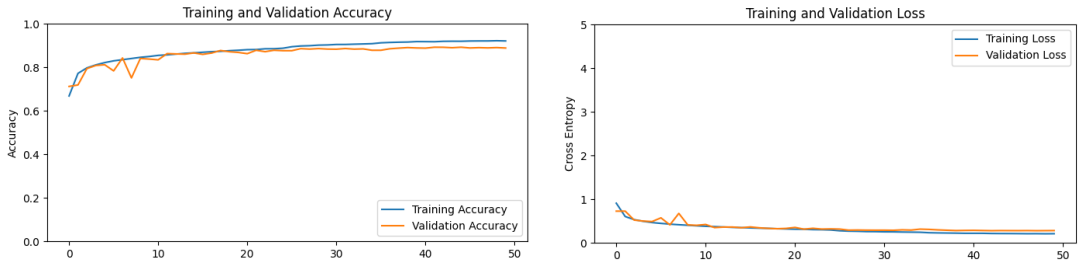


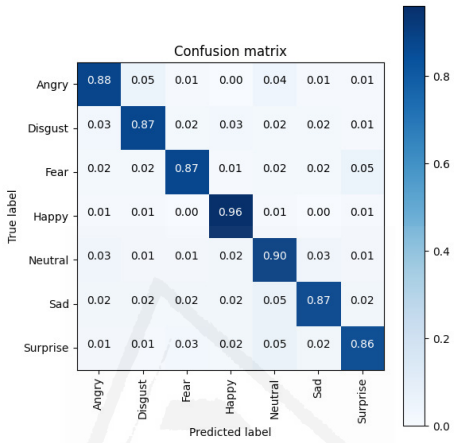
Figura 4.35: Comparación gráfica de ambas distribuciones.

El diagrama de barras de la Figura 4.35 muestra que la forma de la distribución de la nueva AffectNet reclasificada es algo parecida, sin embargo, hay un claro aumento de imágenes en las categorías de miedo, asco y sorpresa. Esto sugiere que muchas imágenes faciales de estas categorías se clasificaron erróneamente como felices o neutras.

A continuación, se demuestra la mejora del rendimiento en el reconocimiento de emociones. La versión reclasificada de AffectNet se utiliza para entrenar el mismo modelo basado en una CNN, lo que da como resultado las curvas de aprendizaje y la matriz de confusión que se muestran en la Figura 4.36. Las curvas de precisión de los conjuntos de entrenamiento y validación aumentan desde la primera época y alcanzan un nivel muy alto, cercano al 90%. Además, ambas curvas se mantienen muy próximas entre sí.



(a) Curvas de aprendizaje de la reclasificación de AffectNet total.



(b) Matriz de confusión de AffectNet total.

Figura 4.36: Curvas de aprendizaje y matriz de confusión del dataset reclasificado de AffectNet.

Las curvas de error disminuyen juntas hasta niveles cercanos a cero, lo cual es deseable. Utilizando los resultados de validación sugeridos por los creadores del dataset, calculamos la precisión con el método *evaluate()* y generamos la matriz de confusión normalizada. La precisión en el conjunto de validación reclasificado es del 89.17%. Para cada categoría de emoción, la precisión fluctúa entre el 86% y el 96%, lo que demuestra una alta tasa de reconocimiento, sin un sesgo notorio para ninguna de las categorías en comparación con el dataset original. Este comportamiento confirma un mejor rendimiento de FER en el dataset reclasificado de AffectNet.

Tabla 4.23: Comparación del rendimiento del estado del arte en los datasets FER considerados.

Dataset	Trabajo	Modelo	Precisión
FER2013	[53]	VGG fine tuning	73.28%
FER2013 (reclasificado)	Nuestro	CNN personalizada	86.71%
NHFI	Nuestro	EfficientNet-B0 transfer learning	55.97%
NHFI (reclasificado)	Nuestro	EfficientNet-B0 transfer learning	70.44%
AffectNet	[111]	CNN-Mecanismo de atención	65.69%
AffectNet (reclasificado)	Nuestro	CNN personalizada	89.17%

Por último, en la Tabla 4.23 se comparan los resultados del método propuesto con el rendimiento del estado del arte en los mismos datasets utilizados en este trabajo. Se trata de arquitecturas *single-network* que no utilizaron imágenes adicionales a las existentes en los datasets. En todos los casos, nuestras versiones reclasificadas de los datasets permiten obtener los valores de precisión más altos, tanto para el modelo personalizado como para el aprendizaje por transferencia. Para el nuevo dataset NHFI, no existe ningún informe formal sobre la precisión de la clasificación, por lo que establecemos como referencia la precisión alcanzada por el modelo de aprendizaje por transferencia en el dataset original, y la contrastamos con la versión reclasificada. Estos resultados demuestran la eficacia de nuestro método de refinamiento, ya que mejora el rendimiento de los modelos de FER incluso alcanzando valores de precisión del estado del arte.

4.4.3 Combinación de datasets

Una vez que se ha verificado la mejora de la calidad de los datasets por separado, éstos son combinados formando un único dataset, el cual es útil para entrenar un modelo de DL que se espera sea más genérico. Para comprobarlo, utilizamos cada uno de los datasets (incluso el artificial) para entrenar su respectivo modelo de DL y evaluamos en el conjunto de prueba de los distintos datasets.

4.4.3.1 Organización de datasets

El recurso fundamental para el entrenamiento es el dataset. A lo largo de las secciones anteriores, hemos preparado en total 5 datasets que se utilizan para nuestra experimentación (Tabla 4.24). Cada dataset está organizado en dos carpetas: "TRAIN" que contiene las imágenes faciales para ajustar el modelo, y "TEST" para las imágenes del subconjunto de prueba equilibrado e insesgado (sólo el artificial no es insesgado) que permite evaluar el modelo. En el momento del entrenamiento, el conjunto de entrenamiento se subdivide automáticamente en una proporción de 80:20 en los subconjuntos de entrenamiento y validación, respectivamente.

Tabla 4.24: División de todos los datasets para el entrenamiento y prueba.

Dataset	**TRAIN**		**TEST**	**Total**
	Entrenamiento (80%)	**Validación (20%)**		
FER2013	28321	7077	560	35958
NHFI	3915	974	560	5449
AffectNet	23669	5912	560	30141
Combinado	55896	13972	1680	71548
Artificial	5284	1321	1680	8285

4.4.3.2 Entrenamiento

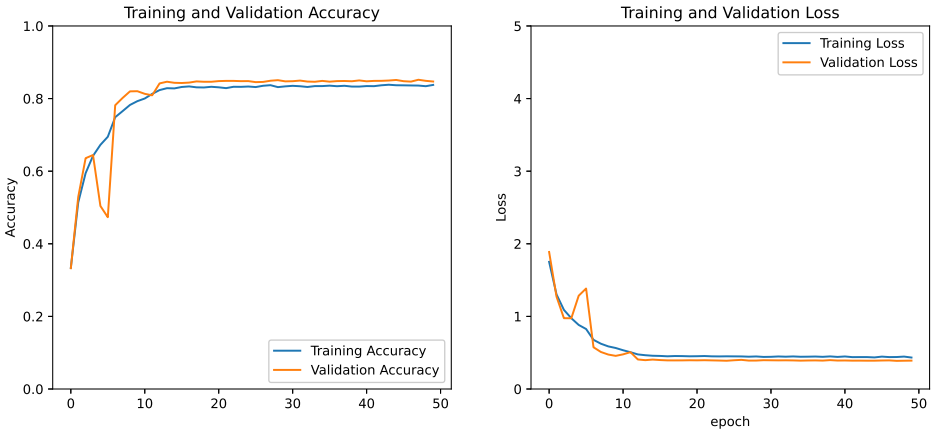
Realizamos un entrenamiento en cada conjunto de datos utilizando los hiperparámetros enumerados en la Tabla 4.25 y la arquitectura de red más adecuada, es decir, la CNN personalizada para FER2013, AffectNet y combinado, mientras que EfficienteNetB0 y MobileNetV2 para NHFI y el dataset artificial, respectivamente.

Tabla 4.25: Hiperparámetros de entrenamiento para cada dataset. Estos valores son los más convenientes luego de varias pruebas.

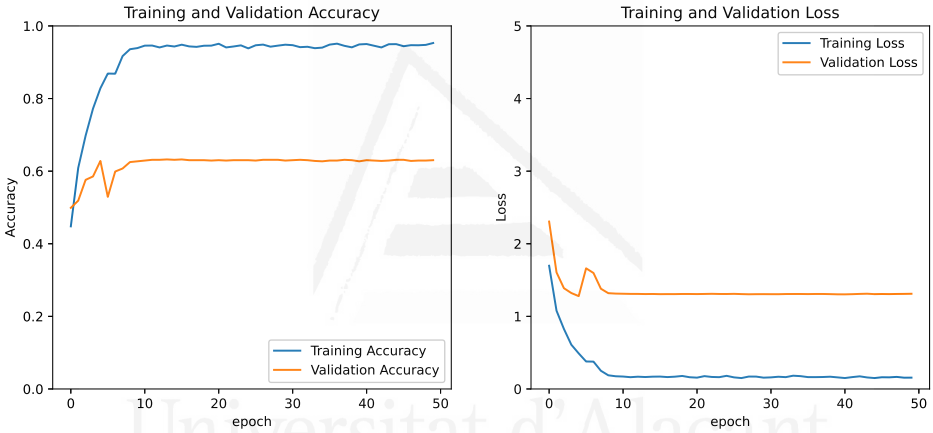
Hiperparámetro	FER2013	NHFI	AffectNet	Combinado	Artificial
Entrada	48,48,1	224,224,3	48,48,3	48,48,3	224,224,3
Tamaño de lote	64	64	64	64	64
Normalización	1/255	No	1/255	1/255	1/255
Optimizador	Adam	Adam	Adam	Adam	Adam
Tasa aprendizaje	0.01 to 0.00001	0.01 to 0.00001	0.0003 to 0.00001	0.0003 to 0.00001	0.01 to 0.00001
Pérdida	categorical_crossentropy	categorical_crossentropy	categorical_crossentropy	categorical_crossentropy	categorical_crossentropy
Métricas	loss & accuracy	loss & accuracy	loss & accuracy	loss & accuracy	loss & accuracy
Clases	7	7	7	7	7
Épocas	50	50	50	50	50
Aumento de datos	Sí	No	Sí	Sí	No

Durante el entrenamiento, el modelo aprende a asociar imágenes faciales con etiquetas de emoción. Utilizando *ImageDataGenerator* de Keras, las imágenes faciales se pasan al modelo en lotes de 64 imágenes y se etiquetan automáticamente con la categoría respectiva. Esta utilidad también se encarga del aumento de datos, la normalización de píxeles y la unificación de las diferentes resoluciones y colores del dataset combinado. Para cada lote, las etiquetas predichas y reales se comparan mediante la función *categorical_crossentropy*, obteniendo pérdidas y precisión para los subconjuntos de entrenamiento y validación. Los algoritmos backpropagation y *Adam* (basados en el descenso de gradiente) reducen el error actualizando los pesos en función de la *tasa de aprendizaje*. Este valor disminuye desde un valor inicial hasta un mínimo si la pérdida no mejora tras unas pocas épocas. Las métricas de pérdida y precisión para las 50 épocas se representan como curvas de aprendizaje con Matplotlib.

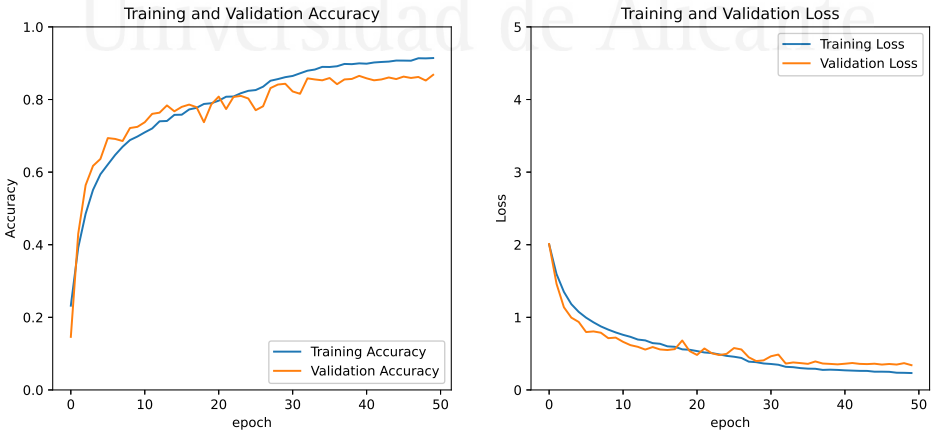
La Figura 4.37 muestra las curvas de aprendizaje de la red convolucional en cada dataset. Se presentan las curvas de precisión (izquierda) y pérdida (derecha) de los subconjuntos de entrenamiento y validación. Idealmente, las curvas de precisión deberían aumentar en altura a medida que avanzan las épocas (eje horizontal), mientras que las curvas de error deberían aproximarse a cero. Además, las curvas de entrenamiento y validación deben estar muy próximas entre sí para evitar el sobreajuste o el subajuste. El mejor rendimiento del modelo corresponde a los datasets AffectNet, FER2013 y combinado, en este orden, mientras que en el caso de los datasets NHFI y artificial el sobreajuste es marcado con una gran separación de las curvas de entrenamiento y validación, tanto para la precisión como para la pérdida. Esto indica que el modelo se



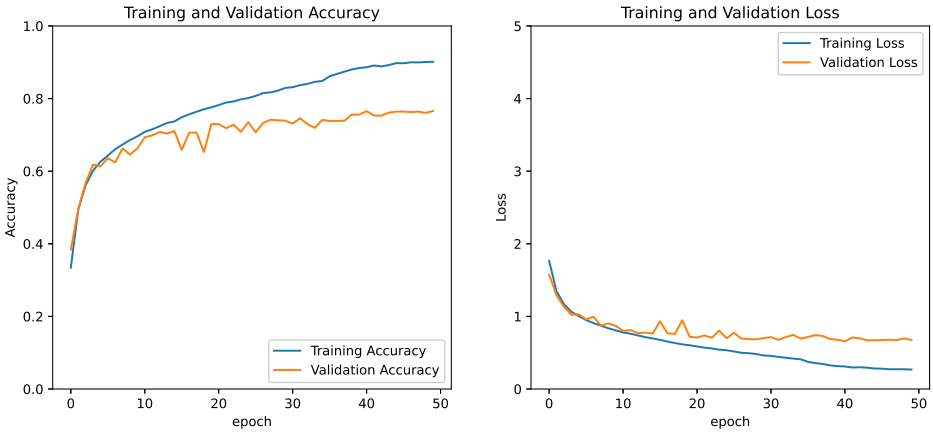
(a) CNN personalizada entrenada en FER2013.



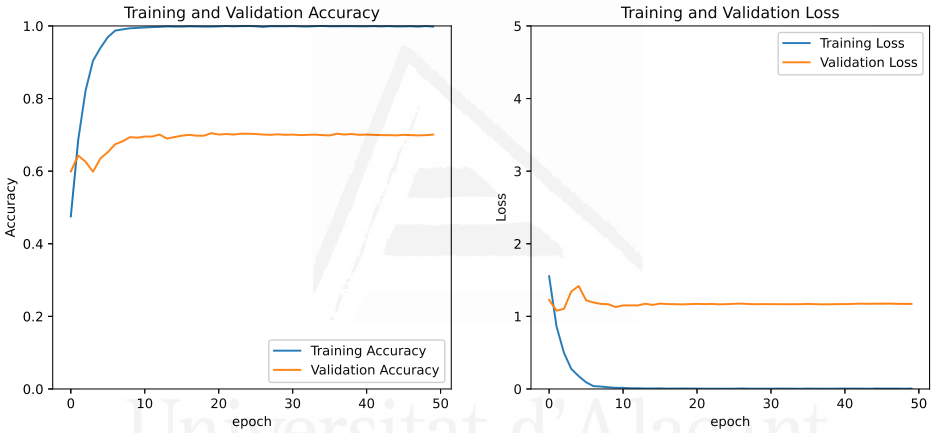
(b) CNN basada en EfficientNetB0 entrenada en NHFI.



(c) CNN personalizada entrenada en AffectNet.



(d) CNN personalizada entrenada en el dataset combinado.



(e) CNN basada en MobileNetV2 entrenada en el dataset artificial.

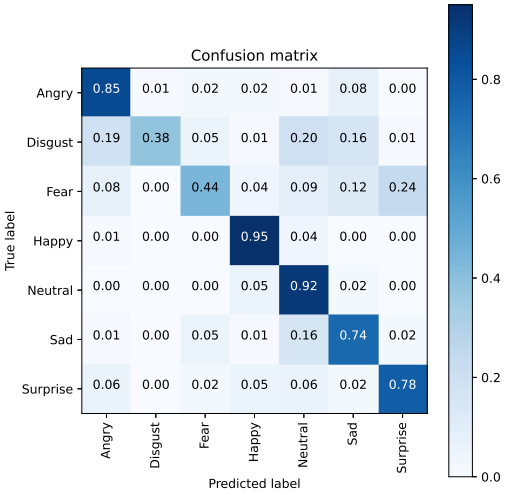
Figura 4.37: Curvas de aprendizaje de las fases de entrenamiento y validación de las redes convolucionales para los datasets considerados.

ajusta bastante bien a los datos de entrenamiento, pero muestra un rendimiento regular en las imágenes de validación. Nos interesa especialmente el modelo entrenado con el dataset combinado. Los niveles alcanzados por las curvas de precisión están por encima del 90% para el entrenamiento y del 75% para la validación, mientras que la curva de pérdida está muy cerca de cero para el entrenamiento y por debajo de 1 para la validación. Se trata de un rendimiento aceptable teniendo en cuenta el tamaño y la variabilidad del dataset. Aunque hay una separación entre las curvas de entrenamiento y validación, la distancia no es grande. Esto sugiere que el modelo puede tener un buen rendimiento en el dataset evaluador.

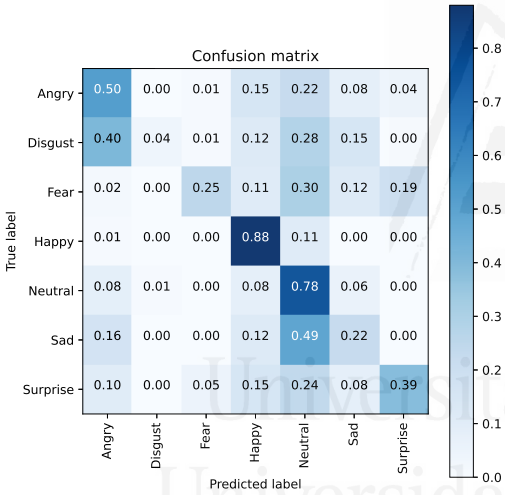
4.4.3.3 Evaluación

El uso práctico de un modelo de DL requiere no solo saber lo bueno que es en el entrenamiento, sino también cómo rinde con imágenes nuevas. Llevamos a cabo evaluaciones single- y cross-dataset para medir el rendimiento en el mismo dataset y también en el resto para determinar la capacidad de generalización. Para ello, utilizamos los modelos obtenidos en la fase de entrenamiento y los subconjuntos de prueba de FER2103, NHFI, AffectNet, combinado y artificial. Es esencial que las imágenes faciales del subconjunto de prueba se oculten al modelo hasta la evaluación. En el enfoque single-dataset, cada modelo se evalúa con el subconjunto de prueba del mismo dataset de entrenamiento. El rendimiento es específico y limitado a un dataset concreto, lo que no es muy útil para contextos diversos como el mundo real. En el enfoque cross-dataset, el dataset de prueba es diferente del dataset de entrenamiento. Normalmente, se utiliza un único dataset de prueba, pero nosotros ampliamos la evaluación a más datasets que combinados funcionan como uno solo. Nos fijamos especialmente en el rendimiento utilizando el dataset evaluador, que proporciona una medida más general para las aplicaciones prácticas. Los resultados se presentan visual y cuantitativamente mediante matrices de confusión y la métrica de precisión resumida en una tabla comparativa.

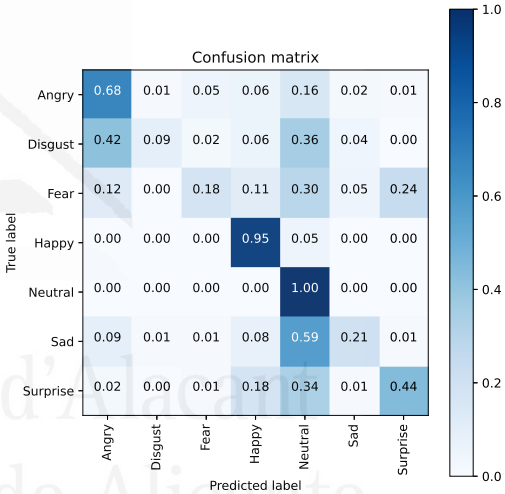
La *matriz de confusión* muestra el número de predicciones verdaderas positivas, verdaderas negativas, falsas positivas y falsas negativas realizadas por un modelo de clasificación. En este caso, utilizamos los valores normalizados entre 0 y 1. Las Figuras 4.38, 4.39, 4.40, 4.41 y 4.42 son las matrices de confusión para los subconjuntos de prueba FER2013, NHFI, AffectNet, combinado y artificial, respectivamente. Cada figura contiene una matriz single-dataset y cuatro matrices cross-dataset, ya que hay cinco subconjuntos de prueba en total. Todas las matrices utilizan el modo de mapa de calor, en el que cada celda se colorea en función de su valor, es decir, cuanto mayor es el valor de la celda o el número de predicciones, mayor es la intensidad del color. La matriz va acompañada de una escala que muestra la gama de colores y valores. Un rendimiento ideal del modelo resaltaría la diagonal principal de la matriz con la mayor intensidad de color, mientras que las celdas restantes deberían mostrar la menor intensidad de color.



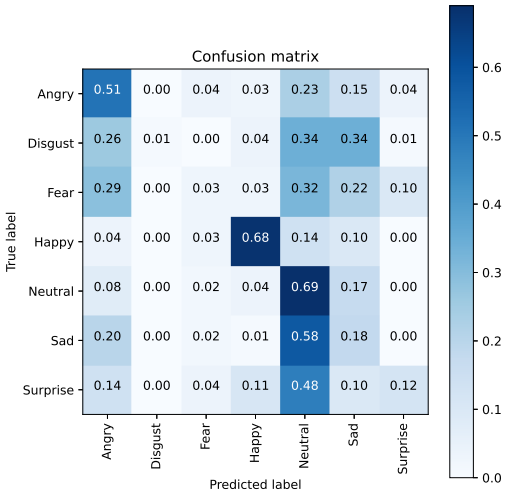
(a) Modelo entrenado y evaluado en FER2013.



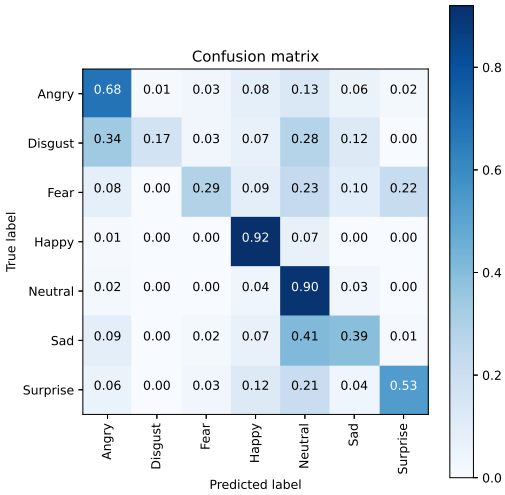
(b) Evaluado en NHFI.



(c) Evaluado en AffectNet.

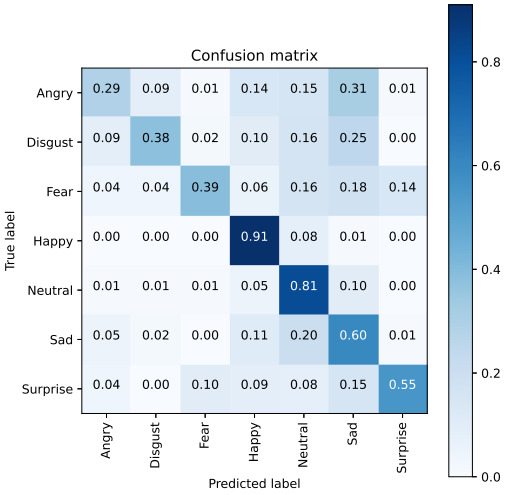


(d) Evaluado en el dataset artificial.

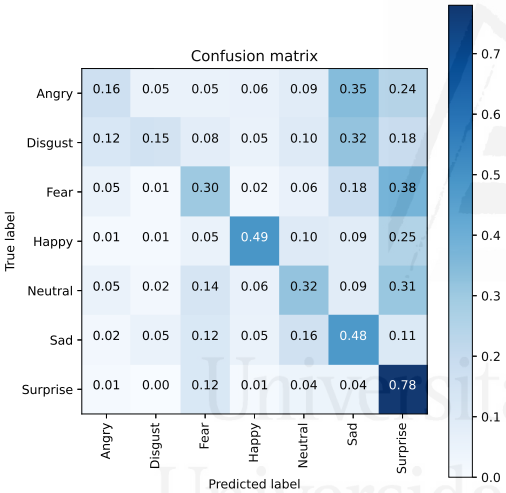


(e) Evaluado en el dataset combinado.

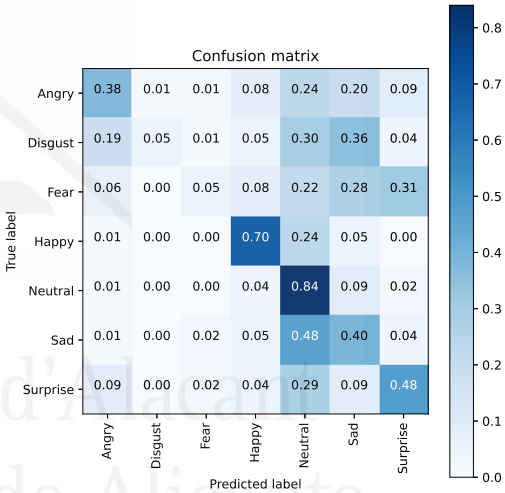
Figura 4.38: Matrices de confusión single- y cross-dataset para el modelo CNN personalizado entrenado en FER2013 y evaluado en cada subconjunto de prueba.



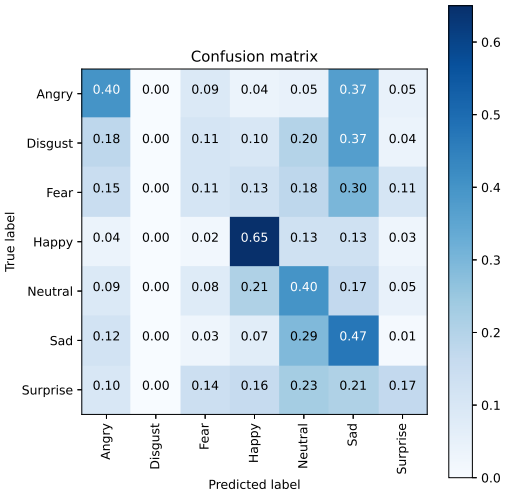
(a) Modelo entrenado y evaluado en NHFI.



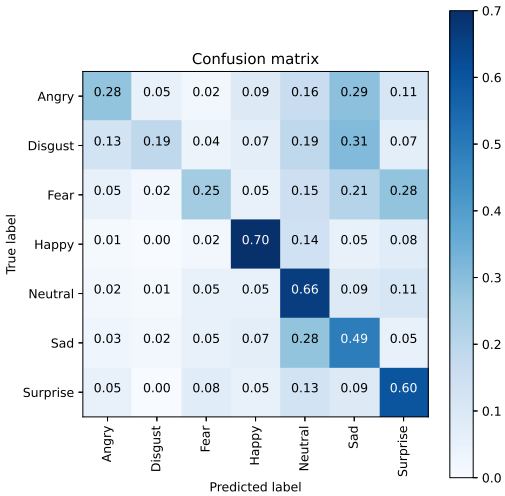
(b) Evaluado en FER2013.



(c) Evaluado en AffectNet.

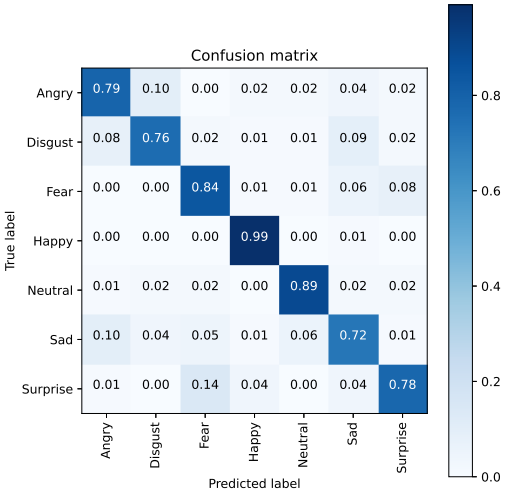


(d) Evaluado en el dataset artificial.

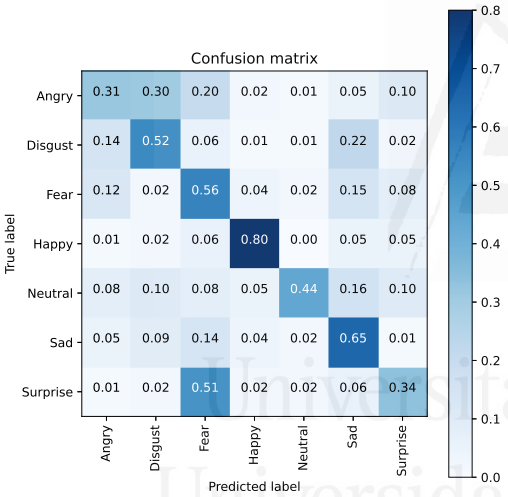


(e) Evaluado en el dataset combinado.

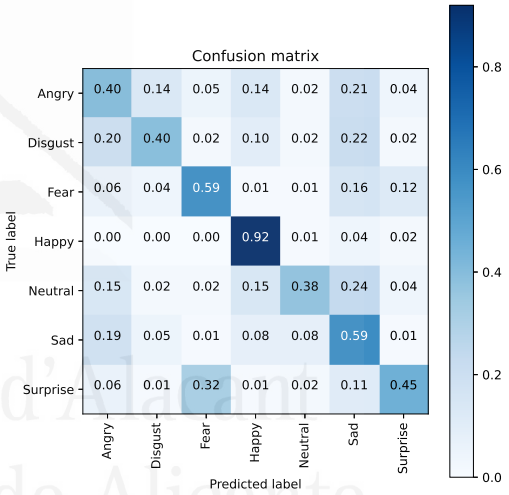
Figura 4.39: Matrices de confusión single- y cross-dataset para el modelo basado en EfficienteNetB0 entrenado en NHFI y evaluado en cada subconjunto de prueba.



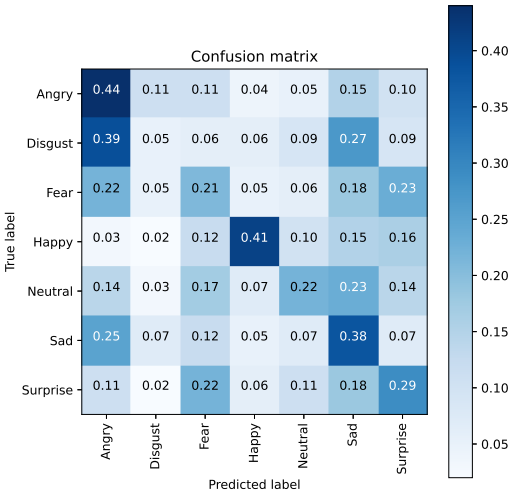
(a) Modelo entrenado y evaluado en AffectNet



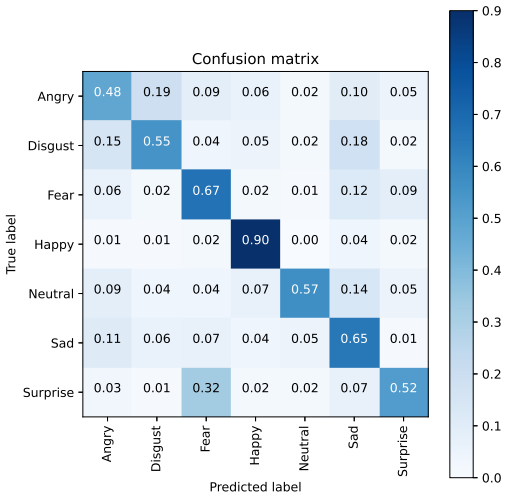
(b) Evaluado en FER2013.



(c) Evaluado en NHFI.

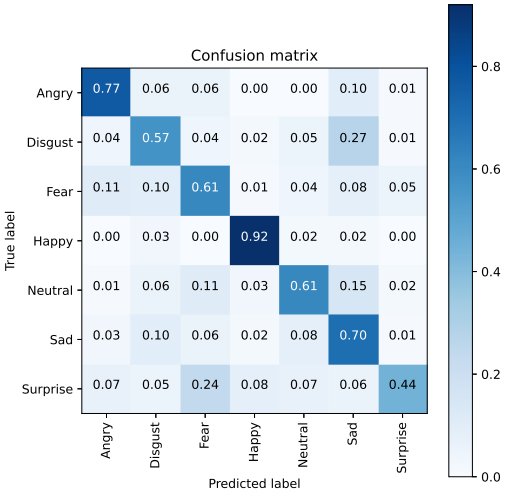


(d) Evaluado en el dataset artificial.

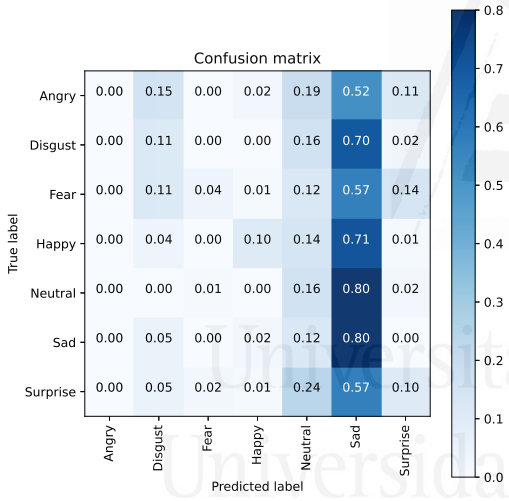


(e) Evaluado en el dataset combinado.

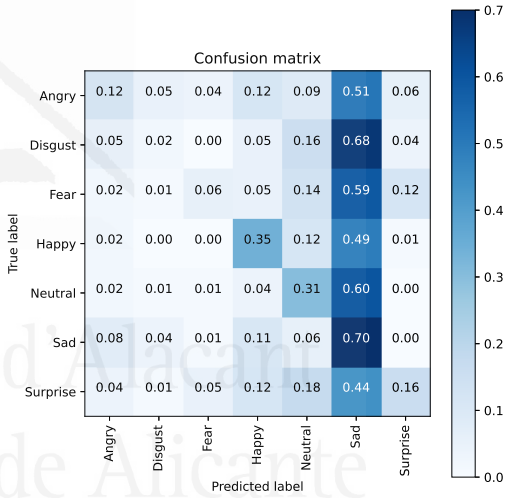
Figura 4.40: Matrices de confusión single- y cross-dataset para el modelo CNN personalizado entrenado en AffectNet y evaluado en cada subconjunto de prueba.



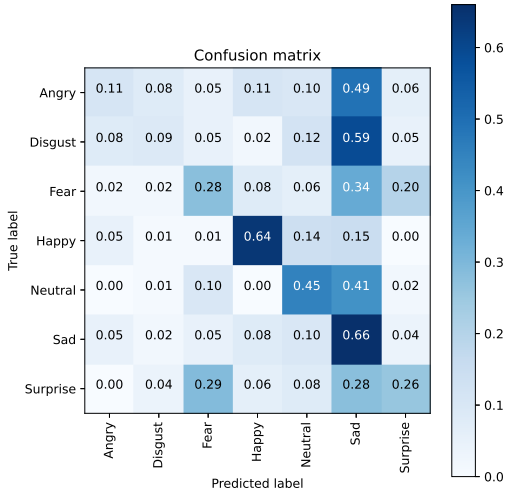
(a) Modelo entrenado y evaluado en el dataset artificial.



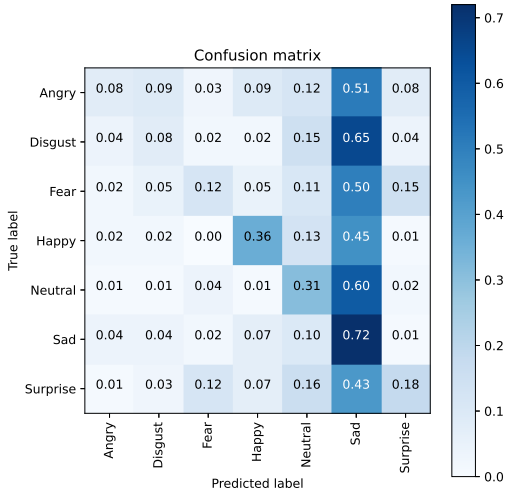
(b) Evaluado en FER2013.



(c) Evaluado en NHFI.

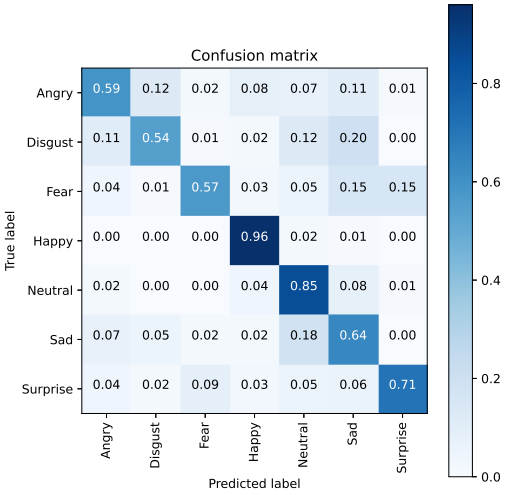


(d) Evaluado en AffectNet.

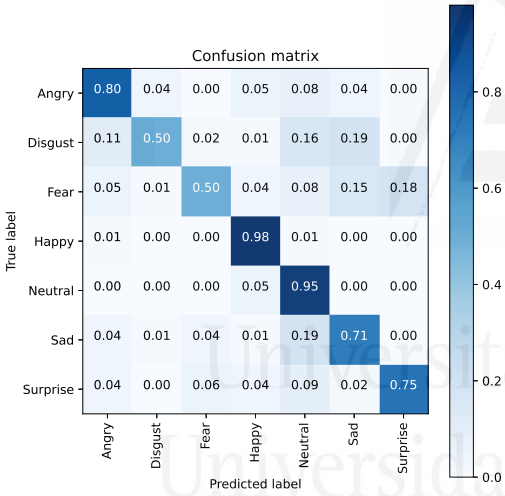


(e) Evaluado en el combinado.

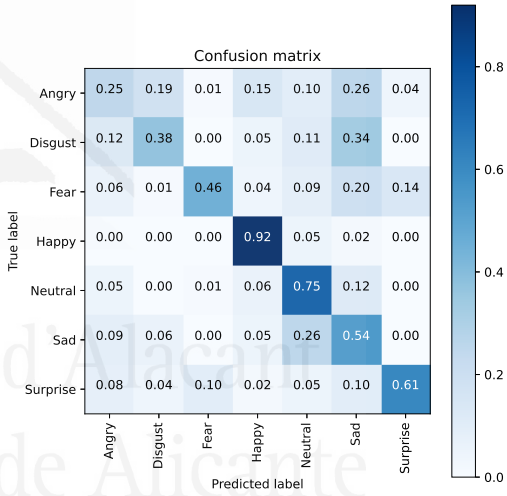
Figura 4.41: Matrices de confusión single- y cross-dataset para el modelo basado en MobileNetV2 entrenado en el dataset artificial y evaluado en cada subconjunto de prueba.



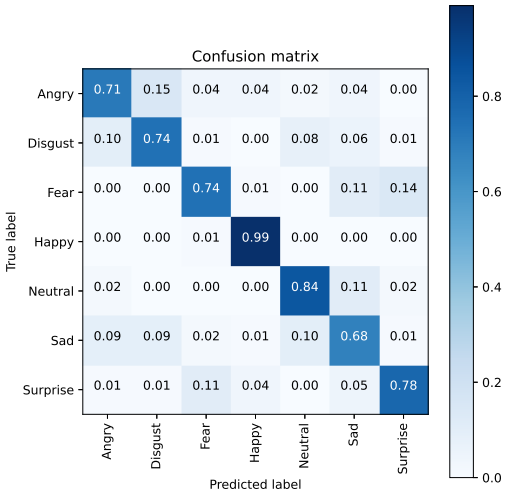
(a) Modelo entrenado y evaluado en el dataset combinado.



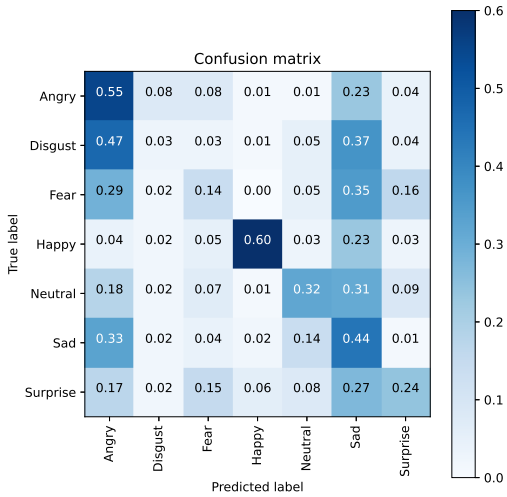
(b) Evaluado en FER2013.



(c) Evaluado en NHFI.



(d) Evaluado en AffectNet.



(e) Evaluado en el artificial.

Figura 4.42: Matrices de confusión single y cross-dataset para el modelo CNN personalizado entrenado en el dataset combinado y evaluado en cada subconjunto de prueba.

Para todos los datasets de prueba, la matriz de confusión que más se aproxima a este comportamiento es el enfoque single-dataset (Figura 4.38a, 4.39a, 4.40a, 4.41a y 4.42a), por lo que se espera que el rendimiento del reconocimiento sea aceptable para imágenes del mismo dataset. Una situación similar ocurre con el modelo entrenado en el dataset combinado y probado con FER2013 y AffectNet (Figura 4.42b y 4.42d), que no son estrictamente casos cross-dataset, ya que las imágenes de prueba proceden de los datasets más grandes utilizados para la combinación. Por el contrario, el peor rendimiento, reflejado por una distribución caótica de los colores, corresponde a la evaluación en el subconjunto de prueba artificial (Figura 4.38d, 4.39d, 4.40d y 4.42e). Esto sugiere una disimilitud significativa entre los datasets reales y el dataset sintético. Lo mismo ocurre en la dirección opuesta, es decir, cuando el modelo entrenado con el dataset artificial se evalúa en subconjuntos de prueba de datasets reales (Figura 4.41b, 4.41c y 4.41d). Las evaluaciones single- y cross-dataset del modelo entrenado en el dataset combinado muestran el mejor rendimiento global (Figura 4.42). A excepción de la matriz de confusión del subconjunto artificial (Figura 4.42e), las demás son muy similares, mostrando la diagonal principal con los colores más oscuros y el resto de las celdas con los colores más claros. El dataset combinado es el único caso capaz de obtener un rendimiento homogéneo consigo mismo y con los demás.

El gradiente de colores es una herramienta visual útil para hacerse rápidamente una idea del rendimiento global del modelo e identificar patrones o tendencias en los datos. También nos informa sobre la precisión en cada clase individualmente. El mejor reconocimiento en los datasets reales corresponde principalmente a las categorías feliz y neutro, con raras excepciones, mientras que las categorías de asco y miedo son las peor reconocidas. Aunque el dataset artificial tiene un rendimiento aceptable en single-dataset, las evaluaciones cross-dataset indican un sesgo notable para la categoría de triste, que se explica por el entrenamiento del modelo con un mayor número de imágenes de esta categoría.

Para una evaluación cuantitativa, hay muchos valores en la matriz de confusión de las 7 categorías de emoción, lo que puede resultar difícil de analizar. Por lo tanto, se necesita un indicador que mida la calidad general de cada modelo y dataset para facilitar la comparación. Utilizamos la *precisión* como métrica de rendimiento, que se calcula sumando el número de instancias correctas en todas las clases y dividiéndolo por el número total de instancias en el dataset. La Tabla 4.26 resume los resultados de precisión obtenidos por cada modelo en cada subconjunto de prueba.

Esta tabla es útil para comparar el rendimiento del modelo con el propio dataset y con el resto, por ejemplo, 0.7214 es la precisión single-dataset lograda por el modelo

Tabla 4.26: Resumen de las precisiones obtenidas en los experimentos single- y cross-dataset.

		Conjunto de prueba				
Mejor modelo	**Entrenamiento**	FER2013	NHFI	AffectNet	Artificial	Combinado
CNN personalizada	FER2013	0.7214	0.4357	0.5054	0.3185	0.5542
EfficientNetB0	NHFI	0.3821	0.5607	0.4125	0.3143	0.4518
CNN personalizada	AffectNet	0.5179	0.5321	0.8125	0.2845	0.6190
MobileNetV2	Artificial	0.1875	0.2482	0.3554	0.6613	0.2637
CNN personalizada	Combinado	0.7411	0.5589	0.7804	0.3310	0.6935

CNN personalizado para FER2013, mientras que 0.5179 es la precisión cross-dataset lograda por el modelo CNN personalizado para FER2013, pero entrenado en AffectNet. Por lo tanto, se obtiene una precisión diferente para el mismo dataset en función del dataset de entrenamiento, aunque la arquitectura del modelo sea la misma. De este cuadro comparativo se desprenden varios resultados:

- La diagonal principal (naranja) contiene los valores de precisión obtenidos en el mismo dataset. Esta precisión single-dataset supera la precisión cross-dataset para cada modelo y dataset. Esto corrobora la fuerte dependencia del modelo del dataset utilizado para el entrenamiento.

- A excepción de NHFI, el resto de los datasets muestra un rendimiento single-dataset entre bueno y muy bueno, especialmente AffectNet y FER2013. Sin embargo, esto puede dar una falsa idea del rendimiento en el reconocimiento de emociones. No es un indicador de cómo funcionará el modelo con imágenes que no proceden del dataset de entrenamiento. Esta precisión puede ser aceptable para imágenes de prueba del propio dataset, pero cuando el modelo se prueba con datasets diferentes, los resultados son muy bajos, lo que indica falta de generalización.

- La última columna (verde) muestra los valores de precisión proporcionados por nuestro dataset evaluador, que abarca imágenes de prueba y características de diferentes datasets. Se trata de una métrica más robusta y genérica para comparar modelos y datasets, ya que indica lo bueno que es el modelo para la generalización a más datasets. Proponemos este punto de referencia para un estado del arte más general en la tarea de reconocimiento de emociones.

- Basándonos en la métrica propuesta, podemos esperar mejores resultados de reconocimiento de emociones en situaciones reales utilizando el modelo CNN personalizado entrenado en el dataset combinado, que alcanza la mayor precisión (rojo) y supera a FER2013, NHFI y AffectNet en capacidad de generalización en un 13.93%, 24.17% y 7.45%, respectivamente. Esto valida que combinar varios datasets

in-the-wild para entrenar una red convolucional permite obtener un modelo con una generalización mejorada, menos dependiente de un dataset específico y útil para aplicaciones prácticas.

- La precisión cross-dataset considerando sólo un dataset de prueba (celdas sin color) no es una buena medida de la generalización, ya que se extiende a un único dataset diferente del de entrenamiento. Como era de esperar, estas precisiones son muy bajas, lo que sugiere la gran disparidad de imágenes y propiedades entre los datasets de FER, por ejemplo, el modelo CNN personalizado se entrena en FER2013 para imágenes JPG de dimensiones (48,48,1), pero cuando se evalúa en NHFI, las imágenes son PNG de dimensiones (224,224,3). Aunque la utilidad ImageDataGenerator realiza la conversión de formato, la disimilitud subyacente degrada el rendimiento.

- El análisis de las celdas sin color por filas sugiere que los valores pueden reflejar la similitud entre datasets, por ejemplo, FER2013 es más similar a AffectNet, mientras que AffectNet es más similar a NHFI. En cambio, todos los datasets reales son menos similares al dataset artificial. Esto puede deberse a la baja variación intraclase en las imágenes faciales sintéticas y a la brecha de dominio entre las imágenes faciales sintéticas y reales. Por lo tanto, los datasets de imágenes faciales reales no pueden sustituirse completamente por los sintéticos, pero es posible complementarlos para tareas relacionadas con los rostros.

4.5 Conclusiones

El reconocimiento de emociones humanas mediante DL a partir de imágenes faciales es un gran reto, que no puede abordarse con un enfoque model-centric sin antes trabajar en la mejora de la materia prima que son los datos. Los actuales datasets de imágenes faciales creados para el reconocimiento y la clasificación de emociones adolecen de diversas problemáticas, que se han demostrado y tratado en el presente trabajo.

El enfoque de ingeniería de datos que proponemos, apoyado en técnicas y herramientas de DL, ha dado una respuesta positiva a cada una de las tres preguntas planteadas en esta investigación. Primero, un rendimiento del estado del arte en el reconocimiento de emociones ha sido alcanzado utilizando las versiones refinadas de los datasets FER2013, NHFI y AffectNet. Segundo, la combinación de estas versiones mejoradas ha permitido entrenar un modelo de mayor capacidad de generalización, lo cual es demandado por las aplicaciones del mundo real. En tercer lugar, las estrategias diseñadas para equilibrar y

hacer más representativo e imparcial un dataset de imágenes faciales, nos permitieron crear un conjunto evaluador que proporciona una métrica de rendimiento más cercana al estado del arte general y útil para comparar y seleccionar el modelo más adecuado para escenarios prácticos.

Este estudio constituye una aportación novedosa a la investigación data-centric, poco explorada debido al esfuerzo arduo y prolongado que requiere la gestión de un dataset de imágenes. Aunque se necesita más trabajo para resolver por completo la falta de calidad y generalización de los datasets de FER, los prometedores resultados de nuestra tesis pueden ayudar y guiar futuras investigaciones.

4.5.1 Refinamiento iterativo

Uno de los problemas más influyentes en los datasets de emociones categóricas es la clasificación errónea. En este trabajo, presentamos e implementamos un método para reclasificar todas las imágenes faciales de un dataset generando una nueva distribución que aumenta la precisión de los modelos de FER. El método propuesto mantiene fija la red convolucional y mejora iterativamente los datos en sucesivos entrenamientos. Después de cada entrenamiento, el dataset se evalúa con la matriz de confusión, y las imágenes faciales correspondientes a las predicciones correctas (en la diagonal) se seleccionan para formar los datos de entrenamiento posteriores. Este proceso genera gradualmente un modelo más preciso y características más distintivas para cada categoría de emoción. El modelo del último entrenamiento se utiliza para reclasificar todas las imágenes creando una nueva y mejor distribución del dataset.

Experimentamos con datasets de FER populares y CNNs creadas desde cero y aprendizaje por transferencia. La eficacia del método propuesto se valida con el aumento de la precisión de validación en un 20.45%, 14.47% y 39.66%, para FER2013, NHFI y AffectNet, respectivamente. Los resultados sugieren que la calidad y el tamaño del dataset determinan el tipo de modelo más adecuado. Los datasets más grandes generalmente tienden a una menor calidad, por lo que necesitan un modelo desde cero, con un entrenamiento más largo y más parámetros. NHFI es más pequeño y mejor anotado, por lo que es conveniente un modelo preentrenado.

Las versiones reclasificadas de estos datasets mantienen el mismo número de imágenes que el dataset original, pero con menos solapamientos entre categorías y menos variabilidad dentro de una misma categoría de emoción. Esto nos permite alcanzar el rendimiento del estado del arte de los modelos de FER de red simple con un 86.71%, 70.44% y 89.17%, para FER2013, NHFI y AffectNet, respectivamente. Las tasas de reconocimiento mejoraron de forma más significativa para los datasets más grandes y peor clasificados,

es decir, el método propuesto funciona mejor para datasets con un alto nivel de imágenes mal clasificadas.

El proceso de refinamiento también sirve como herramienta de depuración en la recopilación automática de datasets de imágenes. Hemos mantenido el tamaño del dataset original, teniendo en cuenta que la cantidad es importante. Sin embargo, hay imágenes irrelevantes que deberían eliminarse y el desequilibrio puede abordarse con imágenes sintéticas a partir de los recientes modelos generativos.

4.5.2 Combinación de datasets de FER

La idea central para mejorar el reconocimiento de emociones basada en imágenes faciales es que el entrenamiento y la evaluación de modelos con datasets combinados se asemeja más a los escenarios del mundo real. Nuestro método aborda tres aspectos fundamentales:

1. Generar un dataset categórico de emociones más amplio y diverso, creado mediante la la combinación de varios datasets de FER in-the-wild. De este modo, evitamos la recopilación y el etiquetado manual desde cero, que exige tiempo, esfuerzo y son tareas propensas a errores. Aportamos un extenso dataset de 71336 imágenes faciales, mezcladas en cuanto a resolución, color, fondo, iluminación y tipo de archivo. Creemos que se convierte en el primer y más extenso dataset combinado que funciona como una solo en el ámbito de FER. Este producto es útil para entrenar un modelo de reconocimiento de emociones capaz de alcanzar una mejor generalización en entornos reales.

2. Crear un novedoso dataset combinado, imparcial y equilibrado, que está diseñado en respuesta a la necesidad de disponer de una referencia para una evaluación más sólida y veraz del rendimiento de los modelos de reconocimiento de emociones y de la calidad de los datasets en los que se han entrenado. La condición de imparcialidad se aborda mediante una organización del dataset basada en características representativas de la población. Con esta base, realizamos una selección equitativa de imágenes faciales según las diferentes categorías de género, edad y etnia de los individuos. La ausencia de imágenes faciales para determinadas categorías motiva el uso de imágenes sintéticas, que es una estrategia conveniente según la literatura. Utilizamos el reciente modelo Stable Diffusion, que resulta eficaz para generar artificialmente las imágenes que faltan para equilibrar el dataset. Obtuvimos un dataset evaluador de 1680 imágenes faciales destinado a probar y comparar los distintos modelos de FER, y que proporciona una métrica que consideramos una

primera aproximación a un estado del arte más general y no específico como en la actualidad.

3. Realizar el entrenamiento y la evaluación de modelos de DL, basados en CNNs personalizadas y preentrenadas (aprendizaje por transferencia), en los diferentes datasets de entrenamiento y prueba presentados y generados en este trabajo. Principalmente, la comparación se ha realizado en los tres datasets de FER in-the-wild que funcionan como uno solo, a diferencia del enfoque tradicional que utiliza el mismo dataset o uno diferente, pero único.

Con respecto a los resultados de los experimentos single- y cross-dataset, la precisión single-dataset es mayor porque el modelo se evalúa con imágenes del mismo dataset, pero no es necesariamente una medida del rendimiento del modelo en otros datasets. Cuando el modelo se evalúa con imágenes del mismo dataset, la precisión es mayor, pero no es un fiel reflejo de la generalización. Por ende, es una métrica específica de escasa o nula utilidad en entornos reales. La precisión cross-dataset es menor debido a la discrepancia entre los datasets de entrenamiento y de prueba. En la evaluación cross-dataset típica sólo hay un dataset de prueba, por lo que no es una medida fiable de la generalización. En cambio, nosotros evaluamos el modelo en el dataset combinado, equilibrado e imparcial, que proporciona una métrica más representativa que puede utilizarse como referencia para FER en aplicaciones prácticas. Nuestro enfoque muestra que el entrenamiento en el dataset combinado logra una mayor generalización, superando a FER2013 en un 13.93%, un 24.17% sobre NHFI y un 7.45% sobre AffectNet. De este modo, identificamos el modelo más adecuado para la tarea de reconocimiento de emociones que se utiliza para nuestro sistema de FER.

4.5.3 Imágenes faciales artificiales

Utilizamos un modelo texto-imagen para la generación de imágenes sintéticas y diseñamos una estructura adecuada del prompt, que es la clave para obtener un buen resultado. Esta técnica es adecuada para tratar desequilibrios y sesgos en datasets de imágenes faciales. Produce imágenes de gran calidad y realismo, y en menos tiempo en comparación con técnicas tradicionales como GAN. No es necesario entrenar el modelo y las identidades no son reales, lo que evita problemas de privacidad. Estas ventajas motivan la creación de un dataset completamente artificial de emociones categóricas, cuyo subconjunto de prueba está equilibrado.

Aunque el rendimiento del dataset artificial es comparable al de los datasets reales, el rendimiento cross-dataset es muy bajo. Esto sugiere la imposibilidad de sustituir comple-

tamente las imágenes faciales reales por las sintéticas, sin embargo, su complementariedad ha sido de gran ayuda en este trabajo, dado que se necesita una pequeña cantidad de datos sintéticos. El dataset artificial es un producto novedoso para el ámbito de FER, creado con IA para ayudar a la IA, que promete reducir costes, tiempo y esfuerzo, y mejorar la calidad de los datos, en contraste con el paradigma clásico de recopilación manual de datos.

Capítulo 5

Aplicaciones

Este capítulo cubre uno de nuestros objetivos que es llevar a la práctica los datasets de mejor calidad y los modelos de DL entrenados en estos datasets en el desarrollo de aplicaciones del mundo real. Probamos nuestra propuesta en condiciones y entornos reales. La Sección 5.1 corresponde al sistema de categorización de páginas Web, que puede ayudar a diseñadores, webmasters y usuarios en general, mientras que la Sección 5.2 presenta el sistema de reconocimiento de emociones dirigido al ámbito educativo.

5.1 Sistema de categorización Web

El propósito es desarrollar un sistema que clasifique las páginas Web en varias categorías, pero basándose únicamente en sus capturas de pantalla[1]. Un sistema capaz de clasificar una página Web según su aspecto visual puede servir de apoyo a los diseñadores en la creación o el mantenimiento de una página Web, comprobar si la página encaja en la categoría adecuada, identificar los patrones de categoría que deben utilizarse como directrices de diseño, optimizar los resultados de los motores de búsqueda para que coincidan con lo que desean los usuarios, organizar el contenido de Internet, tarea poco práctica para los humanos y esencial de los directorios Web.

La implementación se realiza con Python sobre la plataforma de ejecución de *Google Colab*[2] para facilitar el acceso en línea y sea disponible para los usuarios en general. Básicamente, el funcionamiento consiste de tres etapas.

[1] Los términos *imagen* y *captura de pantalla* se utilizarán indistintamente

[2] https://drive.google.com/file/d/1M4ZYwJWH1zO7D_86u4YdO8cbQ_71tBM-/view?usp=sharing

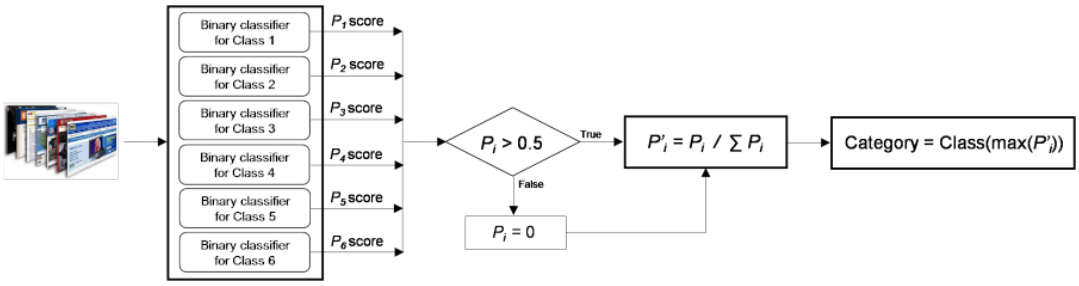


Figura 5.1: Mecanismo de predicción de la categoría Web utilizando los clasificadores binarios.

- Primero, es necesario instalar EfficientNetB0, la cual es la arquitectura de los clasificadores binarios One. vs Rest que permitieron mejorar en conjunto el rendimiento de la categorización Web multiclase, tal como se explicó en el apartado 3.6. También se tienen almacenados en el directorio de Google Drive, los modelos en formato h5 que contienen los pesos resultantes del entrenamiento de cada clasificador binario. Aprovechamos una de las librerías de Colab para realizar la selección de archivos dentro del cuaderno de programación. A través de un botón de selección, el usuario puede cargar el archivo de imagen de la captura de pantalla, misma que se guarda temporalmente en esta plataforma.

- El siguiente paso es el análisis de la imagen cargada, que es recibida por una función de predicción que se encarga de redimensionar la imagen a 224x224 píxeles, convertirla en un array y normalizarla. Esto con el fin de aplicar el método *predict* de Keras para cada uno de los 6 clasificadores binarios. Se genera un valor de predicción (puntuación) por clasificador, que es la probabilidad de que la imagen pertenezca a la categoría. Hemos creado un mecanismo de cálculo que toma en cuenta todos estos valores para la predicción final, tal como se muestra en la Figura 5.1. Si el resultado es menor al umbral establecido por la curva ROC para cada clasificador (0.5 por defecto en los clasificadores convencionales), indicaría una baja probabilidad de pertenecer a la categoría, por lo que se le asigna el valor 0. En caso contrario, se conserva el mismo valor. Todos los valores son considerados para realizar un cálculo de suma relativa a 1 junto con la ponderación respectiva.

- Finalmente, una función de ploteo basada en la librería Matplotlib ayuda a la representación gráfica que permite comparar todos los valores mediante barras, cuya máxima puntuación corresponde a la categoría de la página Web.

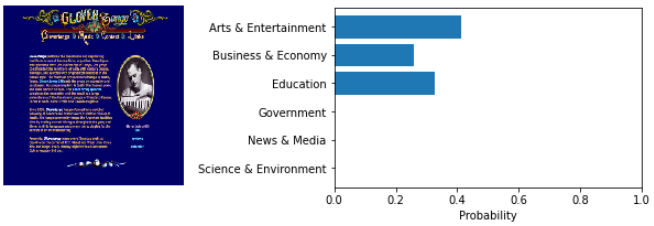
La Figura 5.2 presenta algunos ejemplos de predicción, uno por categoría, con capturas de pantalla tomadas de la sección de páginas Web de Imagenet [17].

5.2 Sistema de reconocimiento de emociones

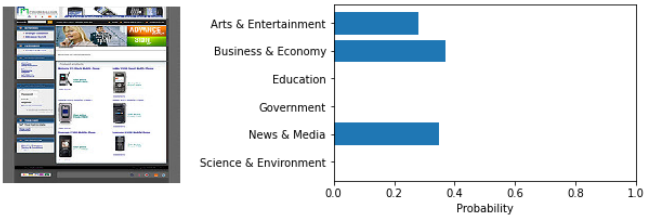
Una de las aplicaciones más importantes de las FER es en la educación. Las expresiones faciales son clave en el aula, tanto física como virtualmente. Los gestos de los alumnos sirven de retroalimentación al profesor para detectar el compromiso o la falta de interés [97]. Mientras que los gestos del profesor pueden ser un aviso de cumplimiento o incumplimiento de los objetivos educativos. Hoy en día, un sistema de reconocimiento de la expresión facial se hace imprescindible en el cada vez más presente aprendizaje online o virtual, que ha provocado el abandono de muchos alumnos, la falta de compromiso y la baja calidad de la enseñanza.

La expresión facial es la mejor prueba del estado emocional de una persona. Un sistema de FER puede ser embebido en cámaras de seguridad para monitorizar en tiempo real o post-procesado, el nivel de atención y concentración de los alumnos, lo que se convierte en un apoyo para la toma de decisiones por parte del profesor. Hemos aprovechado el modelo obtenido a partir del entrenamiento personalizado de CNN sobre el conjunto de datos combinados para la implementación de un sistema online de acceso público para el reconocimiento de expresiones faciales en reuniones de vídeo que pueden tener lugar en plataformas como Google Meet, Microsoft Teams o Zoom.

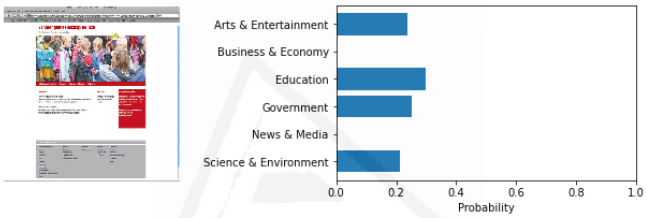
La Figura 5.3 presenta una captura de pantalla del sistema de reconocimiento de expresiones faciales en acción utilizando el modelo obtenido del entrenamiento personalizado de la CNN sobre el conjunto de datos combinados. Podemos apreciar cómo se encuadra la región facial de cada uno de los participantes y se le asigna la etiqueta de categoría de expresión correspondiente. Incluso es posible distinguir por color el tipo de expresión facial, donde las positivas o neutras tienen un color tenue, mientras que las negativas, como el caso de “enfado”, destacan por el color rojo, lo que sería perceptible para quien supervise la atención y el compromiso de las personas durante la reunión, además de ser un aviso para tomar una decisión.



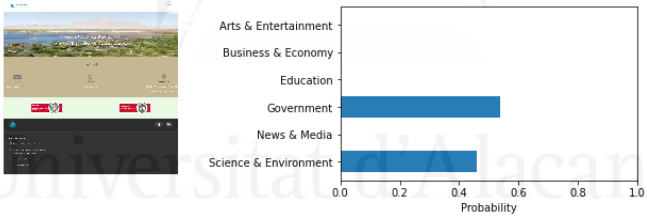
(a) Arte y entretenimiento



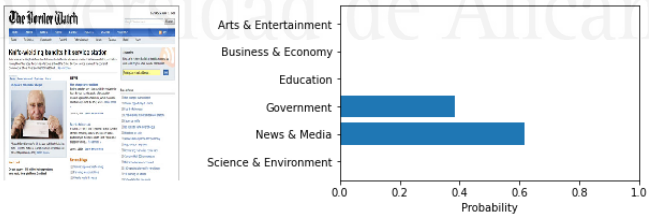
(b) Negocios y economía



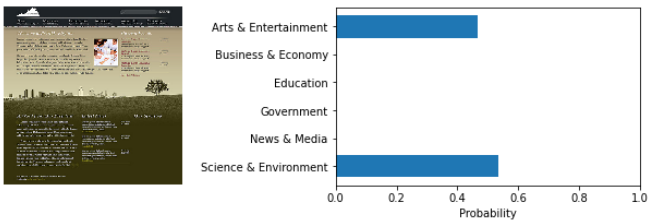
(c) Educación



(d) Gobierno



(e) Noticias y medios



(f) Ciencia y ambiente

Figura 5.2: Resultados del sistema de categorización Web con un ejemplo de cada categoría.

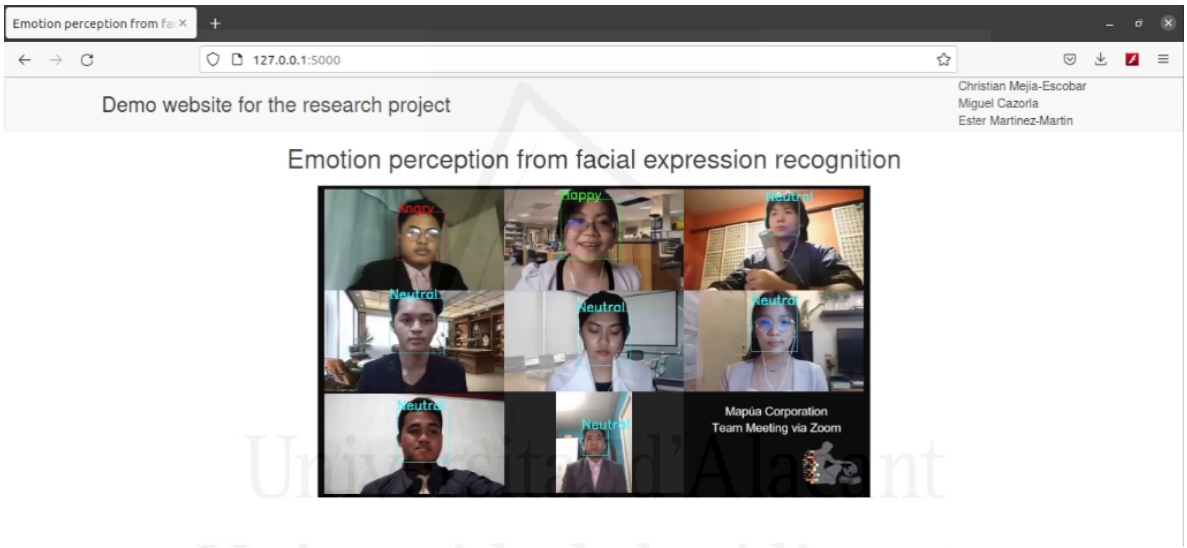


Figura 5.3: Captura de pantalla del sistema de reconocimiento de emociones en pleno funcionamiento.

Capítulo 6

Conclusiones

Este capítulo final comprende el siguiente contenido. La Sección 6.1 establece las conclusiones generales asociadas con los objetivos de nuestro trabajo de investigación. La Sección 6.2 enuncia las contribuciones que se desprenden de esta tesis. La Sección 6.3 se refiere al acceso y la disponibilidad de los recursos utilizados y productos generados en este trabajo. La Sección 6.4 cita las publicaciones científicas derivadas del desarrollo de la presente tesis. Por último, la Sección 6.5 menciona las futuras líneas de investigación.

6.1 Conclusiones generales

Esta tesis se centró en el problema de mejorar el reconocimiento y la clasificación de imágenes utilizando DL en aplicaciones del mundo real. Aunque los modelos de DL y los datos comparten la responsabilidad, comprobamos que mejorar la calidad de los datasets de imágenes ayuda a incrementar la precisión y la mayor generalización de los modelos en esta tarea.

Hemos diseñado e implementado una novedosa propuesta basada en los datos como una contribución al poco explorado enfoque de investigación data-centric. Creamos estrategias de ingeniería de datos apoyadas en técnicas de DL que abordan los inconvenientes más comunes de los datasets de imágenes como la clasificación errónea, las imágenes irrelevantes, el desequilibrio de clases, el sesgo o falta de representatividad y las propiedades únicas de los datasets.

Demostramos la eficacia de nuestro método en dos aplicaciones del mundo real de gran relevancia y utilidad, pero de alta complejidad. Tanto la categorización de páginas Web como el reconocimiento de emociones humanas, ambas utilizando solamente imágenes

como entrada, son problemas de alta dificultad para las computadoras, incluso para las personas.

En el caso de las páginas Web, fue necesario crear desde cero un dataset y mejorarlo, mientras que en las emociones, aprovechamos lo existente mediante la combinación y mejora de varios datasets in-the-wild disponibles. Los datasets mejorados permitieron aumentar el rendimiento y la capacidad de generalización de los modelos de DL para estas aplicaciones. Por lo tanto, una metodología de DL debería tener en cuenta la calidad del dataset como requisito previo a la búsqueda de mejores arquitecturas de red y configuraciones de modelos.

6.2 Contribuciones de la tesis

Proporcionamos las siguientes aportaciones en la categorización de páginas Web basada en capturas de pantalla:

- Un amplio dataset mixto en tipo de datos de libre acceso sobre la estructura, el contenido y la apariencia visual de las páginas Web. Puede ser un producto útil para la investigación y desarrollo Web.
- Un flujo de trabajo apoyado en herramientas computacionales para automatizar el proceso de recopilación, organización y depuración de capturas de pantalla y atributos de páginas Web. La metodología diseñada puede adaptarse a problemas que requieran la recopilación, organización, análisis y publicación de grandes cantidades de datos.
- Un sistema de categorización de páginas Web utilizando únicamente capturas de pantalla, evitando el análisis del contenido textual y código. Puede ser útil para optimizar recursos de Internet y ahorrar tiempo y costes en sistemas de recuperación de información como motores de búsqueda, clasificadores, sistemas de recomendación, directorios Web y rastreadores.

En el reconocimiento y la clasificación de emociones humanas basada en expresiones faciales, contribuimos con:

- Un novedoso método de refinamiento iterativo para reclasificar las imágenes de un dataset que permite una mayor precisión de un modelo de FER. Aunque hemos mantenido el tamaño del dataset, considerando que la cantidad es importante, este método también demostró ser útil como una herramienta de depuración, ya que

puede detectar imágenes irrelevantes que deberían eliminarse. Para completar o equilibrar un dataset, puede aprovecharse el modelo generativo de difusión.

- Una metodología aplicable a otros datasets de diferentes dominios y soportada por herramientas informáticas, especialmente librerías de Python y DL.

- Una versión reclasificada de cada dataset, que puede ser útil para futuras investigaciones, disponible públicamente para FER2013 y NHFI, mientras que para AffectNet esto no es posible debido a restricciones de licencia.

- Un dataset de imágenes faciales formado a partir de la combinación de FER2013, NHFI y AffectNet. Hasta donde sabemos, se trata del mayor dataset combinado de emociones categóricas que abarca distintas propiedades de color, tamaño, resolución y formato de imagen. Es útil para entrenar un modelo más general en el reconocimiento de emociones, en contraste con los modelos actuales que se basan en un dataset específico.

- Un dataset más pequeño que el anterior, pero equilibrado e imparcial en términos de género, edad y origen étnico para evaluar la generalización de un modelo de FER y el dataset en el que se ha entrenado. Esta métrica es una aproximación a un estado del arte más general en el reconocimiento de emociones.

- Un novedoso dataset de imágenes faciales sintéticas que utiliza la herramienta generativa de vanguardia Stable Diffusion. Hasta donde sabemos, se trata del primer dataset artificial de FER, que puede dar lugar a una nueva categoría además de los datasets in-the-lab e in-the-wild. Describimos la ingeniería de prompts diseñada para controlar el contenido de las imágenes y el etiquetado automático mediante unidades de acción de expresión facial y las categorías de género, edad y etnia. Se trata de un producto útil para la investigación y el desarrollo en FER, que proporciona imágenes sintéticas de alta calidad en menos tiempo en comparación con las técnicas tradicionales, sin necesidad de entrenamiento y sin los riesgos potenciales de invasión de la privacidad de las personas como ocurre con los datasets reales.

- Un sistema basado en la Web para la categorización de emociones utilizando el modelo entrenado en el dataset combinado de FER.

6.3 Disponibilidad de datos

El dataset de páginas Web y el código utilizado en este trabajo son públicos y están disponibles a través de OSF[1] (*Open Science Foundation*), una plataforma libre y abierta para apoyar, difundir y permitir la colaboración de la investigación científica. El sitio web de la OSF ofrece una interfaz fácil de usar para gestionar todo lo relacionado con el proyecto, cuya organización jerárquica consta de: a) el dataset dividido en la parte visual (webshots en categorías), y la parte textual y numérica (hojas de datos); b) el código desarrollado; y c) la documentación de apoyo. El proyecto es público[2] y sus recursos pueden seleccionarse y visualizarse dentro de la misma página y descargarse a través del botón de la barra superior.

Otras contribuciones a disposición del público[3] son: el conjunto de datos refinados, el código implementado, el modelo y sus mejores ponderaciones, y un sistema en línea para la categorización Web basada en temas.

Respecto al trabajo relacionado con las emociones, hemos aprovechado la plataforma GitHub para poner a disposición del público el código utilizado y los modelos obtenidos. El código desarrollado a partir de este estudio está disponible en el repositorio de GitHub[4], mientras que los datasets generados (excepto la versión reclasificada de AffectNet) están disponibles bajo petición al autor.

6.4 Publicaciones

Artículos derivados del trabajo de tesis:

- Mejia-Escobar, C., Cazorla, M., Martinez-Martin, E. (2023). A Large Visual, Qualitative and Quantitative Dataset for Web Intelligence Applications. Computational Intelligence and Neuroscience. Nature-inspired Computing for Web Intelligence. Hindawi. Aceptación: 05/04/2023.

- Mejia-Escobar, C., Cazorla, M., Martinez-Martin, E. (2022). Webpage Categorization Using Deep Learning. In 16th International Conference on Soft Computing Models in Industrial and Environmental Applications (SOCO 2021) (pp. 358-368). Springer International Publishing.

[1] https://osf.io/
[2] urlhttps://osf.io/7ghd2/
[3] https://osf.io/vbm4w/
[4] https://github.com/cimejia/novel-FER-datasets.git

- Mejia-Escobar, C., Cazorla, M., Martinez-Martin, E. (2022). Towards a better performance in facial expression recognition: a data-centric approach. Computational Intelligence and Neuroscience. Advances in the Application of Human Activity Recognition. Hindawi. Aceptación: 19/09/2022.

- Mejia-Escobar, C., Cazorla, M., Martinez-Martin, E. (2023). Improving Facial Expression Recognition through Data Preparation & Merging. IEEE Access. En revisión. Fecha de envío: 30/04/2023.

Colaboraciones:

- Guerrero-Rodriguez, B., Garcia-Rodriguez, J., Salvador, J., Mejia-Escobar, C., Bonifaz, M., Gallardo, O. (2022). Defining High Risk Landslide Areas Using Machine Learning. In Bio-inspired Systems and Applications: from Robotics to Ambient Intelligence: 9th International Work-Conference on the Interplay Between Natural and Artificial Computation, IWINAC 2022, Puerto de la Cruz, Tenerife, Spain, May 31 – June 3, 2022, Proceedings, Part II. Springer-Verlag, Berlin, Heidelberg, 183–192. https://doi.org/10.1007/978-3-031-06527-9_18

- Guerrero-Rodriguez, B., Garcia-Rodriguez, J., Salvador, J., Mejia-Escobar, C., Bonifaz, M., Gallardo, O. (2022). Landslide Prediction with Machine Learning and Time Windows. In: Ferrández Vicente, J.M., Álvarez-Sánchez, J.R., de la Paz López, F., Adeli, H. (eds) Bio-inspired Systems and Applications: from Robotics to Ambient Intelligence. IWINAC 2022. Lecture Notes in Computer Science, vol 13259. Springer, Cham. https://doi.org/10.1007/978-3-031-06527-9_19

- Aldás-Núñez, R. J., Tuz-Chamorro, K. V., Vega-Ocaña, J. A., Velasco-Haro, M. S., & Mejía-Escobar, C. I. (2022). Delimitación automática de ceniza volcánica en imágenes satelitales mediante Deep Learning. FIGEMPA: Investigación Y Desarrollo, 13(1), 48–58. https://doi.org/10.29166/revfig.v13i1.3121

- Menéndez, B., Ormasa, R., Peñafiel, J., Yauli, V, Mejia-Escobar, C. Automatic recognition and description of sedimentary rocks through Artificial Intelligence. VIII Congreso Internacional de Investigación REDU. Proceedings of the VIII REDU International Research Congress, Ambato, Ecuador. Medwave. 2022 Apr 25;22(S1):CI01-CI148. Spanish. doi: 10.5867/Medwave.2022.S1. PMID: 35467094. CI38, pág. 30. https://www.medwave.cl/medios/medwave/Marzo2022/PDF/CongresoCI2022/medwave-2022-S1-CI.pdf

6.5 Trabajo futuro

Nuestro estudio pretende ser un paso más en la solución de la compleja problemática de la calidad de los datos en el campo de DL. Las estrategias de ingeniería de datos y las técnicas de DL que hemos expuesto, pueden ser extendidas más allá de la aplicación al dominio de la Web y de FER, también son útiles para una variedad de problemas de visión por computadora, donde el recurso principal son las imágenes.

La combinación de datasets in-the-wild demostró ser eficaz para aumentar la capacidad de generalización de un modelo DL. El dataset combinado puede ampliarse con más datasets in-the-wild teniendo en cuenta que se han refinado previamente y presentan diferentes propiedades de tamaño, resolución, color, fondo, iluminación y formato de imagen para lograr una mayor variabilidad y permitir que un modelo aprenda patrones generalizables a nuevos datos y escenarios reales. También es posible organizar este gran dataset según categorías de género, edad y etnia, así como equilibrarlo artificialmente con el fin de entrenar un modelo de EERR mucho más general.

Es necesario estudiar a fondo las imágenes sintéticas para comprender sus particularidades y reducir la brecha entre los dominios real y artificial. Esto permitiría integrar datos sintéticos y reales en la fase de entrenamiento de los modelos de DL y no sólo en la de evaluación. Adicionalmente, considerar imágenes faciales con posturas de perfil, lo que es frecuente observar en la vida real.

Se sugiere probar arquitecturas basadas en redes Transformer, ya sea de forma independiente o combinadas con CNNs. Combinar este trabajo con sistemas de categorización basados en código en el caso de las páginas Web y con sistemas de reconocimiento multimodal en el caso de emociones humanas, lo que proporcionaría soluciones más robustas. Creemos que estas contribuciones mejoran la calidad de los datasets de imágenes y la precisión de los modelos que se entrenan en ellos.

Referencias

[1] Abbasnejad, I., Sridharan, S., Nguyen, D., Denman, S., Fookes, C., and Lucey, S. (2017). Using synthetic data to improve facial expression analysis with 3d convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 1609–1618.

[2] Abou Zafra, R., Abdullah, L. A., Alaraj, R., Albezreh, R., Barhoum, T., and Al Jallad, K. (2022). An experimental study in real-time facial emotion recognition on 3rl dataset.

[3] Ahmed, N., Al Aghbari, Z., and Girija, S. (2023). A systematic survey on multimodal emotion recognition using learning algorithms. *Intelligent Systems with Applications*, 17:200171.

[4] Baek, K., Bang, D., and Shim, H. (2021). Gridmix: Strong regularization through local context mapping. *Pattern Recognition*, 109:107594.

[5] Banerjee, S., Bernhard, J. S., Scheirer, W. J., Bowyer, K. W., and Flynn, P. J. (2017). Srefi: Synthesis of realistic example face images. In *2017 IEEE International Joint Conference on Biometrics (IJCB)*, pages 37–45.

[6] Barsoum, E., Zhang, C., Ferrer, C. C., and Zhang, Z. (2016). Training deep networks for facial expression recognition with crowd-sourced label distribution.

[7] Bhalla, A. (2020). Facial expression recognition.

[8] Bhatti, Y., Jamil, A., Nida, N., Yousaf, M. H., Viriri, S., and Velastin, S. (2021). Facial expression recognition of instructor using deep features and extreme learning machine. *Computational Intelligence and Neuroscience*, 2021:1–17.

[9] Boer, V., Someren, M., and Lupascu, T. (2010). Classifying web pages with visual features. *Journal of Clinical Virology - J CLIN VIROL*, 1:245–252.

[10] Boutros, F., Huber, M., Siebke, P., Rieber, T., and Damer, N. (2022). Sface: Privacy-friendly and accurate face recognition using synthetic data. In *2022 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–11.

[11] Bozorgtabar, B., Rad, M. S., Kemal Ekenel, H., and Thiran, J.-P. (2019). Using photorealistic face synthesis and domain adaptation to improve facial expression analysis. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–8.

[12] Brownlee, J. (2019). *Generative adversarial networks with python: deep learning generative models for image synthesis and image translation*. Machine Learning Mastery.

[13] Canal, F. Z., Müller, T. R., Matias, J. C., Scotton, G. G., de Sa Junior, A. R., Pozzebon, E., and Sobieranski, A. C. (2022). A survey on facial emotion recognition techniques: A state-of-the-art literature review. *Information Sciences*, 582:593–617.

[14] Chaudhari, A., Bhatt, C., Krishna, A., and Mazzeo, P. L. (2022). Vitfer: Facial emotion recognition with vision transformers. *Applied System Innovation*, 5(4):80.

[15] Chen, L.-F., Wu, M., Pedrycz, W., and Hirota, K. (2021). *Emotion Recognition and Understanding for Emotional Human-Robot Interaction Systems*. Springer.

[16] Colbois, L., de Freitas Pereira, T., and Marcel, S. (2021). On the use of automatically generated synthetic image datasets for benchmarking face recognition. In *2021 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–8. IEEE.

[17] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. IEEE.

[18] DeVries, T. and Taylor, G. W. (2017). Improved regularization of convolutional neural networks with cutout.

[19] Dias, W., Andaló, F., Padilha, R., Bertocco, G., Almeida, W., Costa, P., and Rocha, A. (2022). Cross-dataset emotion recognition from facial expressions through convolutional neural networks. *Journal of Visual Communication and Image Representation*, 82:103395.

[20] Ekman, P. (1992). An argument for basic emotions. *Cognition & Emotion*, 6:169–200.

[21] Ekman, P. and Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17 2:124–9.

[22] Ekman, P. and Friesen, W. V. (1978). Facial action coding system. *Environmental Psychology & Nonverbal Behavior*.

[23] Falconieri, V. (2019a). Circl images ail dataset.

[24] Falconieri, V. (2019b). Circl images phishing dataset.

[25] Faria, A. R., Almeida, A., Martins, C., Gonçalves, R., Martins, J., and Branco, F. (2017). A global perspective on an emotional learning model proposal. *Telematics and Informatics*, 34(6):824–837. SI: IT Education and Training.

[26] Fiorini, L., Loizzo, F. G., D'Onofrio, G., Sorrentino, A., Ciccone, F., Russo, S., Giuliani, F., Sancarlo, D., and Cavallo, F. (2023). Can i feel you? recognizing human's emotions during human-robot interaction. In *Social Robotics: 14th International Conference, ICSR 2022, Florence, Italy, December 13–16, 2022, Proceedings, Part I*, pages 511–521. Springer.

[27] Fisogni, P. (2023). Machine learning and emotions. In *Encyclopedia of Data Science and Machine Learning*, pages 961–970. IGI Global.

[28] Fu, T., Abbasi, A., and Chen, H.-c. (2010). A focused crawler for dark web forums. *JASIST*, 61:1213–1231.

[29] Gao, H. and Ogawara, K. (2022). Face alignment by learning from small real datasets and large synthetic datasets. In *2022 Asia Conference on Algorithms, Computing and Machine Learning (CACML)*, pages 397–402.

[30] Ghosh, T., Banna, M. H. A., Nahian, M. J. A., Kaiser, M. S., Mahmud, M., Li, S., and Pillay, N. (2023). A privacy-preserving federated-mobilenet for facial expression detection from images. In *Applied Intelligence and Informatics: Second International Conference, AII 2022, Reggio Calabria, Italy, September 1–3, 2022, Proceedings*, pages 277–292. Springer.

[31] Gigerenzer, G. (2022). *How to stay smart in a smart world: Why human intelligence still beats algorithms*. MIT Press.

[32] Gomez-Donoso, F. (2020). Contributions to 3d object recognition and 3d hand pose estimation using deep learning techniques.

[33] Goodfellow, I. J., Erhan, D., Carrier, P. L., Courville, A., Mirza, M., Hamner, B., Cukierski, W., Tang, Y., Thaler, D., Lee, D.-H., et al. (2013). Challenges in representation learning: A report on three machine learning contests. In *International conference on neural information processing*, pages 117–124. Springer.

[34] Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial networks.

[35] Gori, M., Schiatti, L., and Amadeo, M. (2021). Masking emotions: Face masks impair how we read emotions. *Frontiers in Psychology*, 12.

[36] Gozalo-Brizuela, R. and Garrido-Merchan, E. C. (2023). Chatgpt is not all you need. a state of the art review of large generative ai models.

[37] Greco, A., Strisciuglio, N., Vento, M., and Vigilante, V. (2023). Benchmarking deep networks for facial emotion recognition in the wild. *Multimedia Tools and Applications*, 82:11189—-11220.

[38] Groenendijk, L. (2003). Planning and management tools.

[39] Haghighi, S., Jasemi, M., Hessabi, S., and Zolanvari, A. (2018). Pycm: Multiclass confusion matrix library in python. *J. Open Source Softw.*, 3:729.

[40] Harl, M., Herchenbach, M., Kruschel, S., Hambauer, N., Zschech, P., and Kraus, M. (2022). A light in the dark: Deep learning practices for industrial computer vision.

[41] He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition.

[42] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

[43] Henderson, P., Islam, R., Bachman, P., Pineau, J., Precup, D., and Meger, D. (2018). Deep reinforcement learning that matters. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press.

[44] Hendler, J. and Hall, W. (2016). Science of the world wide web. *Science*, 354(6313):703–704.

[45] Hendrycks, D., Mu, N., Cubuk, E. D., Zoph, B., Gilmer, J., and Lakshminarayanan, B. (2020). Augmix: A simple data processing method to improve robustness and uncertainty.

[46] Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., and Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications.

[47] Hwang, H. and Matsumoto, D. (2022). Functions of emotions.

[48] Jiao, Y., Niu, Y., Tran, T. D., and Shi, G. (2020). 2d+3d facial expression recognition via discriminative dynamic range enhancement and multi-scale learning.

[49] Jin, X., Sun, W., and Jin, Z. (2020). A discriminative deep association learning for facial expression recognition. *International Journal of Machine Learning and Cybernetics*, 11.

[50] Kamruzzaman, S. M. (2010). Web page categorization using artificial neural networks.

[51] Kandeel, A., Rahmanian, M., Zulkernine, F., Abbas, H. M., and Hassanein, H. (2021). Facial expression recognition using a simplified convolutional neural network model. In *ICCSPA 2020 - 4th International Conference on Communications, Signal Processing, and their Applications*, volume 2021-January. Cited By :3.

[52] Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., and Aila, T. (2019). Analyzing and improving the image quality of stylegan.

[53] Khaireddin, Y. and Chen, Z. (2021). Facial emotion recognition: State of the art performance on fer2013.

[54] Khan, A. R. (2022). Facial emotion recognition using conventional machine learning and deep learning methods: Current achievements, analysis and remaining challenges. *Information*, 13(6):268.

[55] Khani, M. G., Mazinani, M. R., Fayyaz, M., and Hoseini, M. (2016). A novel approach for website aesthetic evaluation based on convolutional neural networks. In *2016 Second International Conference on Web Research (ICWR)*, pages 48–53.

[56] Khder, M. A., Bahar, A. Y., and Fujo, S. W. (2022). Effect of artificial intelligence in the field of games on humanity. In *2022 ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems (ICETSIS)*, pages 199–204. IEEE.

[57] Kim, D. Y. and Wallraven, C. (2022). Label quality in affectnet: results of crowd-based re-annotation. In *Asian Conference on Pattern Recognition*, pages 518–531. Springer.

[58] Kim, J. H. and Han, D. S. (2020). Data augmentation & merging dataset for facial emotion recognition. In *Proceedings of the Symposium of the 1st Korea Artificial Intelligence Conference, Jeju, Korea*, pages 12–16.

[59] Kim, J. H., Poulose, A., and Han, D. S. (2021). The extensive usage of the facial image threshing machine for facial emotion recognition performance. *Sensors*, 21(6).

[60] King, A. (2008). *Website Optimization: Speed, Search Engine & Conversion Rate Secrets*. O'Reilly Media.

[61] Ko, B. (2018). A brief review of facial emotion recognition based on visual information. *Sensors*, 18:401.

[62] Kollias, D., Cheng, S., Ververas, E., Kotsia, I., and Zafeiriou, S. (2020). Deep neural network augmentation: Generating faces for affect analysis. *Int. J. Comput. Vision*, 128(5):1455–1484.

[63] Kumar, A., Bi, L., Kim, J., and Feng, D. D. (2020). Chapter five - machine learning in medical imaging. In Feng, D. D., editor, *Biomedical Information Technology (Second Edition)*, Biomedical Engineering, pages 167–196. Academic Press, second edition edition.

[64] Lassri, S., Benlahmar, E. H., and Tragha, A. (2019). Machine learning for web page classification: A survey. *International Journal of Information Science and Technology*, 3.

[65] Li, Z., Zhang, T., Jing, X., and Wang, Y. (2021). Facial expression-based analysis on emotion correlations, hotspots, and potential occurrence of urban crimes. *Alexandria Engineering Journal*, 60(1):1411–1420.

[66] Liao, Y.-H., Kar, A., and Fidler, S. (2021). Towards good practices for efficiently annotating large-scale image classification datasets.

[67] Lipton, Z. C. and Steinhardt, J. (2019). Research for practice: Troubling trends in machine-learning scholarship. *Commun. ACM*, 62(6):45–53.

[68] Liu, D., Lee, J.-H., Wang, W., and Wang, Y. (2019). Malicious websites detection via cnn based screenshot recognition. In *2019 International Conference on Intelligent Computing and its Emerging Applications (ICEA)*, pages 115–119.

[69] Liu, J., Wang, H., and Feng, Y. (2021). An end-to-end deep model with discriminative facial features for facial expression recognition. *IEEE Access*, 9:12158–12166.

[70] López-Sánchez, D., Corchado Rodríguez, J., and González, A. (2017). A cbr system for image-based webpage classification: Case representation with convolutional neural networks.

[71] López-Sánchez, D., González, A., and Corchado Rodríguez, J. (2019). Visual content-based web page categorization with deep transfer learning and metric learning. *Neurocomputing*, 338.

[72] Martínez-Miranda, J. and Aldea, A. (2005). Emotions in human and artificial intelligence. *Computers in Human Behavior*, 21(2):323–341.

[73] Mazen, F. M. A., Nashat, A. A., and Seoud, R. A. A. A. A. (2021). Real time face expression recognition along with balanced fer2013 dataset using cyclegan. *International Journal of Advanced Computer Science and Applications*, 12(6).

[74] Mehrabian, A. (1968). Communication without words.

[75] Mejia-Escobar, C., Cazorla, M., and Martinez-Martin, E. (2022a). Towards a better performance in facial expression recognition: a data-centric approach. *Computational Intelligence and Neuroscience. Advances in the Application of Human Activity Recognition.* (in press).

[76] Mejia-Escobar, C., Cazorla, M., and Martinez-Martin, E. (2022b). Webpage categorization using deep learning. In Sanjurjo González, H., Pastor López, I., García Bringas, P., Quintián, H., and Corchado, E., editors, *16th International Conference on Soft Computing Models in Industrial and Environmental Applications (SOCO 2021)*, pages 358–368, Cham. Springer International Publishing.

[77] Mejia-Escobar, C., Cazorla, M., and Martinez-Martin, E. (2023a). Improving facial expression recognition through data preparation & merging. *IEEE Access.* (in press).

[78] Mejia-Escobar, C., Cazorla, M., and Martinez-Martin, E. (2023b). A large visual, qualitative and quantitative dataset for web intelligence applications. *Computational Intelligence and Neuroscience. Nature-inspired Computing for Web Intelligence.* (in press).

[79] Meng, H., Yuan, F., Tian, Y., and Yan, T. (2022). Cross-datasets facial expression recognition via distance metric learning and teacher-student model. *Multimedia Tools and Applications*, 81(4):5621–5643.

[80] Michailidou, E., Harper, S., and Bechhofer, S. (2008). Visual complexity and aesthetic perception of web pages. In *Proceedings of the 26th Annual ACM International Conference on Design of Communication*, SIGDOC '08, page 215–224, New York, NY, USA. Association for Computing Machinery.

[81] Mollahosseini, A., Hasani, B., and Mahoor, M. H. (2017). Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, PP(99):1–1.

[82] Mondal, B. (2020). Artificial intelligence: state of the art. *Recent Trends and Advances in Artificial Intelligence and Internet of Things*, pages 389–425.

[83] Ng, A. (2021). Data-centric ai competition. https://https-deeplearning-ai.github.io/data-centric-comp/. Accessed: 2023-05-20.

[84] Nikolova, D., Vladimirov, I., and Terneva, Z. (2022). Artificial humans: an overview of photorealistic synthetic datasets and possible applications. In *2022 57th International Scientific Conference on Information, Communication and Energy Systems and Technologies (ICEST)*, pages 1–4.

[85] Nordhoff, M., August, T., Oliveira, N., and Reinecke, K. (2018). A case for design localization: Diversity of website aesthetics in 44 countries.

[86] Nyaupane, B. and Shakya, S. (2022). Age, gender, and ethnicity prediction using deep separable convolutional neural networks. In *International Multidisciplinary Conference*, Dubai.

[87] Pathak, A., Bhalsing, S., Desai, S., Gandhi, M., and Patwardhan, P. (2020). Deep learning model for facial emotion recognition.

[88] Pawara, P., Okafor, E., Groefsema, M., He, S., Schomaker, L. R., and Wiering, M. A. (2020). One-vs-one classification for deep neural networks. *Pattern Recognition*, 108:107528.

[89] Pinney, J., Carroll, F., and Newbury, P. (2021). Robots and uncertainty: An investigation into the impact of the aesthetic visualisation on peoples trust of robots. In Chew, E., P. P. Abdul Majeed, A., Liu, P., Platts, J., Myung, H., Kim, J., and Kim, J.-H., editors, *RiTA 2020*, pages 1–10, Singapore. Springer Singapore.

[90] Pramerdorfer, C. and Kampel, M. (2016). Facial expression recognition using convolutional neural networks: State of the art.

[91] Qi, X. and Davison, B. (2009). Web page classification: Features and algorithms. *ACM Comput. Surv.*, 41.

[92] Ramis, S., Buades, J. M., Perales, F. J., and Manresa-Yee, C. (2022). A novel approach to cross dataset studies in facial expression recognition. *Multimedia Tools and Applications*, pages 1–38.

[93] Reinecke, K. and Gajos, K. (2014). Quantifying visual preferences around the world. *Conference on Human Factors in Computing Systems - Proceedings*.

[94] Rifkin, R. and Klautau, A. (2004). In defense of one-vs-all classification. *J. Mach. Learn. Res.*, 5:101–141.

[95] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2021). High-resolution image synthesis with latent diffusion models.

[96] Rosenberg, E. L. and Ekman, P. (2020). *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press.

[97] Savchenko, A. V., Savchenko, L. V., and Makarov, I. (2022). Classifying emotions and engagement in online learning based on a single facial expression recognition neural network. *IEEE Transactions on Affective Computing*, 13(4):2132–2143.

[98] Schraml, D. (2019). Physically based synthetic image generation for machine learning: a review of pertinent literature. In *Other Conferences*.

[99] Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C. W., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S. R., Crowson, K., Schmidt, L., Kaczmarczyk, R., and Jitsev, J. (2022). LAION-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

[100] Spezialetti, M., Placidi, G., and Rossi, S. (2020). Emotion recognition for human-robot interaction: Recent advances and future perspectives. *Frontiers in Robotics and AI*, 7.

[101] Sunitha, G., Geetha, K., Neelakandan, S., Pundir, A. K. S., Hemalatha, S., and Kumar, V. (2022). Intelligent deep learning based ethnicity recognition and classification using facial images. *Image and Vision Computing*, 121:104404.

[102] Tan, M. and Le, Q. V. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks.

[103] Tan, M. and Le, Q. V. (2020). Efficientnet: Rethinking model scaling for convolutional neural networks.

[104] Vaidya, S. (2020). Detecting human emotions - facial expression recognition.

[105] Vaidya, S. (2021). Natural human face images for emotion recognition. https://www.kaggle.com/sudarshanvaidya/random-images-for-face-emotion-recognition. Accessed: 2023-05-20.

[106] Velasquez, J. and Jain, L. (2010). Advanced techniques in web intelligence - i.

[107] Vonikakis, V., Dexter, N. Y. R., and Winkler, S. (2021). Morphset: Augmenting categorical emotion datasets with dimensional affect labels using face morphing. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 2713–2717.

[108] Wang, Y., Li, Y., Song, Y., and Rong, X. (2019). Facial expression recognition based on random forest and convolutional neural network. *Information*, 10(12).

[109] Wang, Y., Song, W., Tao, W., Liotta, A., Yang, D., Li, X., Gao, S., Sun, Y., Ge, W., Zhang, W., and Zhang, W. (2022). A systematic review on affective computing: emotion models, databases, and recent advances. *Information Fusion*, 83-84:19–52.

[110] Webpages, D. (2019). Web image dataset, institute for computer research, university of alicante.

[111] Wen, Z., Lin, W., Wang, T., and Xu, G. (2021). Distract your attention: Multi-head cross attention network for facial expression recognition.

[112] Yang, J., Shen, J., Lin, Y., Hristov, Y., and Pantic, M. (2023). Fan-trans: Online knowledge distillation for facial action unit detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6019–6027.

[113] Yun, S., Han, D., Chun, S., Oh, S., Yoo, Y., and Choe, J. (2019). Cutmix: Regularization strategy to train strong classifiers with localizable features. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6022–6031, Los Alamitos, CA, USA. IEEE Computer Society.

[114] Zeiler, M. and Fergus, R. (2013). Visualizing and understanding convolutional neural networks. *ECCV 2014, Part I, LNCS 8689*, 8689.

[115] Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. (2018). mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*.

[116] Zhao, H., Wang, Q., Jia, Z., Chen, Y., and Zhang, J. (2021). Bayesian based facial expression recognition transformer model in uncertainty. In *2021 International Conference on Digital Society and Intelligent Systems (DSInS)*, pages 157–161.

[117] Zhong, Y., Wu, L., Liu, X., and Jiang, J. (2022). Exploiting the potential of datasets: A data-centric approach for model robustness.

[118] Zhou, C., Zhao, J., Ma, T., and Zhou, X. (2021a). Haif: A hierarchical attention-based model of filtering invalid webpage. *IEICE TRANSACTIONS on Information and Systems*, 104(5):659–668.

[119] Zhou, L., Fan, X., Tjahjadi, T., and Choudhury, S. D. (2021b). Discriminative attention-augmented feature learning for facial expression recognition in the wild. *Neural Computing and Applications*, pages 1–12.

Apéndice A

Análisis estadístico de datos Web

La gran cantidad de datos recopilados acerca de las páginas Web recolectadas puede proporcionar información útil, la cual puede convertirse en conocimiento. Mediante el análisis estadístico de los atributos cualitativos y cuantitativos de las páginas Web, podemos identificar patrones sobre su estructura, lo cual es fundamental en el diseño Web. Distinguimos dos fuentes de datos: 1) la navegación, que es más estricta, y 2) la búsqueda, que prácticamente no tiene restricciones. Utilizamos el lenguaje de programación R para procesar los datos obtenidos y comparar ambas fuentes. Debido a la gran heterogeneidad que caracteriza a las variables estudiadas, se excluyen los *outliers* en el cálculo de los indicadores estadísticos y las gráficas resultantes. Estos valores difieren mucho de los considerados comunes y pueden causar distorsiones en el análisis matemático y visual. Utilizando la conocida regla "1.5 veces el *Rango Intercuartílico*", se pueden identificar y omitir los valores atípicos. Por este motivo, el número de valores de cada una de las variables puede diferir.

A.1 Atributos cualitativos

Aunque intentamos obtener un conjunto de URLs uniformemente distribuido con respecto a las categorías, los errores citados en la Sección 3.4 sobre el reconocimiento automático de páginas Web erróneas, generaron la distribución irregular mostrada en la Figura A.1a. En la técnica de búsqueda se observa un menor desequilibrio, al contrario que en la navegación, donde predominan las páginas Web relacionadas con negocios y economía, y gobierno, posiblemente debido a una mayor necesidad de difusión y capacidad económica para registrar estas páginas Web en un servicio de pago. Por otra parte, la mayoría de las páginas Web se localizan geográficamente en Europa y Asia, tanto para la navegación como para la búsqueda, siendo estos continentes los que concentran un mayor número de

(a) **(b)**

Figura A.1: Distribución de los atributos cualitativos sobre las páginas Web: a) categoría y b) continente.

países. Además, su potencial económico podría explicar que ocupen los primeros puestos (Figura A.1b).

A.2 Atributos cuantitativos

La variabilidad que se presnta en la técnica de búsqueda es más evidente si se tiene en cuenta el número de caracteres que conforman una URL (Figura A.2a), lo que da lugar a un rango más amplio (la diferencia entre el mínimo y el máximo) y a una media y una desviación estándar mayores. En ambos gráficos, los valores se acumulan más en la parte inferior de la variable y son menos frecuentes en la parte superior, por lo que hay una cola hacia la derecha. Este comportamiento es deseable al leer o teclear una URL en un navegador e indica que no hay demasiados niveles hasta llegar a una página Web concreta. El tiempo de descarga del código fuente de las páginas Web en la Figura A.2b se comporta de forma similar para la navegación y la búsqueda. En ambos casos, aunque existe una variabilidad considerable debido a la amplitud del intervalo y la desviación estándar, los valores se sitúan en torno a los 20 milisegundos y son en su mayoría bajos, lo que supone una ventaja para el usuario que desea ver la página Web en el menor tiempo posible. Según los errores citados en la Sección 3.4, 254 de 3182 URLs (7.98%) pertenecientes a navegación no estaban disponibles, mientras que en búsqueda, 4079 de 46256 (8.82%), no eran accesibles para descargar el código fuente y, por tanto, no era posible la extracción de los parámetros cuantitativos. El comportamiento del tamaño en bytes (Figura A.2c) es casi idéntico en navegación y búsqueda, donde la media de las páginas Web es de aproximadamente 50 KB. Tanto la variabilidad como las colas de las distribuciones son prácticamente iguales, favoreciendo una visualización rápida de la página Web, lo que está en relación directa con un tamaño pequeño. No obstante, todavía existen algunas páginas Web con un tamaño considerable, que puede deberse a elementos gráficos u objetos externos vinculados a la página. En la Figura A.2d, casi el 60% de las

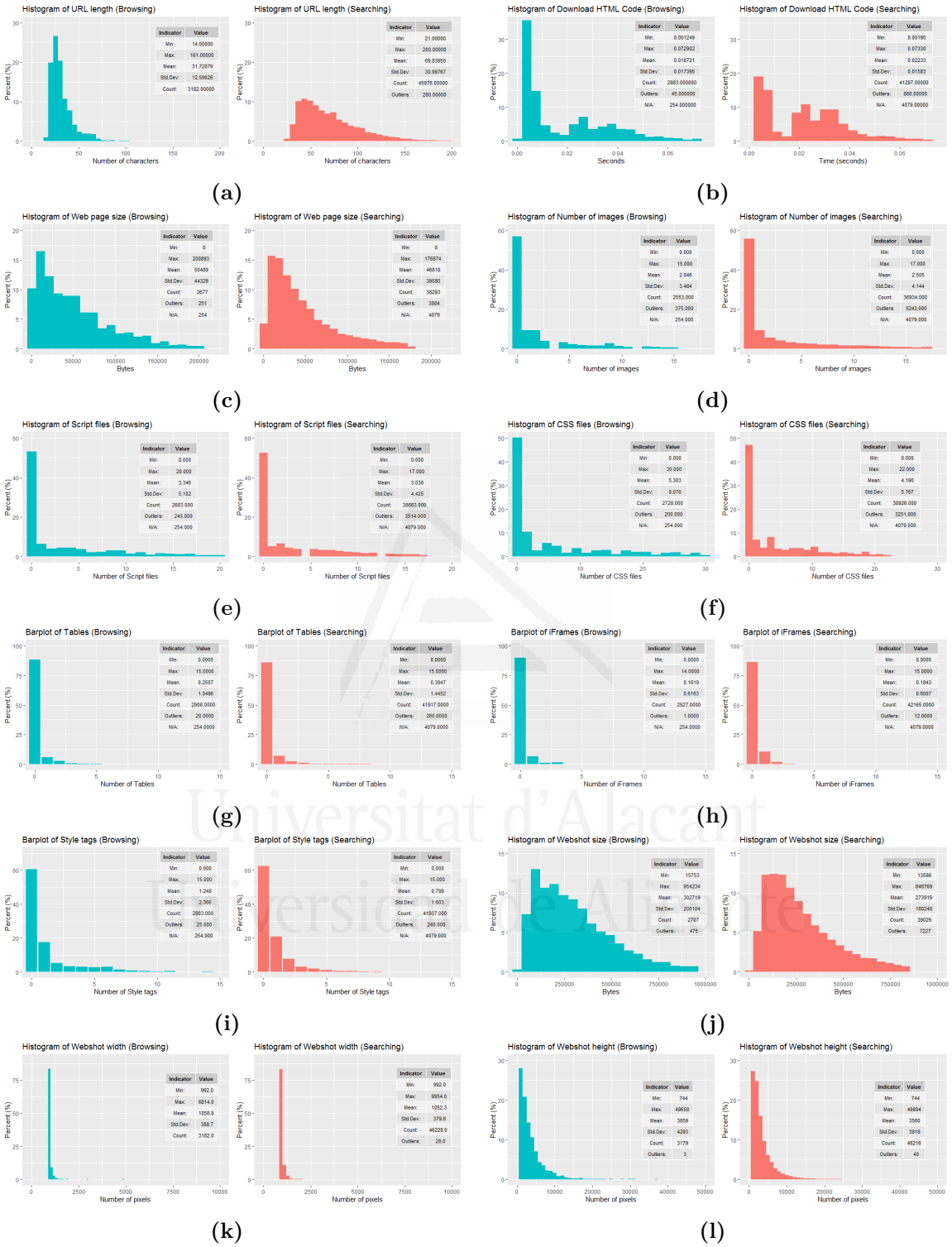


Figura A.2: Distribución de los atributos cuantitativos sobre las páginas Web de navegación y búsqueda.

páginas Web no incluyen imágenes dentro de su código fuente, sólo información textual. Sin embargo, pueden incluir imágenes a través de archivos de estilo CSS, lo cual es una buena práctica [60]. Aunque el rango del número de imágenes es amplio, la media es baja, 2 o 3 imágenes por página Web, por lo que se tiende a utilizar pocas imágenes dentro de una página Web para aligerarla. Las páginas Web sin scripts superan el 50% en navegación y búsqueda. En este sentido, los scripts son programas complementarios que proporcionan funciones adicionales a las páginas Web. Sin embargo, su uso puede provocar incompatibilidades con los navegadores y hacer la página más compleja y pesada. En la Figura A.2e, la tendencia es minimizar la presencia de scripts, 3 scripts por página Web de media. En la Figura A.2f, aproximadamente el 50% de las páginas Web no utilizan archivos de hojas de estilo en cascada (CSS), cuyo uso se recomienda como buena práctica en el diseño Web. La cifra ideal sería un archivo de estilo por página Web, pero la media para navegación y búsqueda es de 5 y 4, respectivamente. La tendencia es hacia valores bajos, aunque hay algunos casos con muchos archivos de estilo, lo que dificulta una visualización ágil de la página. Los gráficos de la Figura A.2g presentan un aspecto muy similar. La mayoría de las páginas Web (alrededor del 85%) ya no utilizan tablas dentro del código fuente. En el pasado, las tablas se utilizaban generalmente para estructurar el contenido, esta práctica ha sido sustituida por la etiqueta “div”, con lo que se consigue un diseño más elegante y profesional. Más del 85% de las páginas Web no utilizan iFrames. Para la navegación y la búsqueda, las barras del gráfico se agrupan a la izquierda (Figura A.2h). Podemos deducir que incrustar otro documento en el documento HTML mediante la etiqueta “iFrame” ya no es una práctica habitual, pues ahora existen mejores opciones. Más de la mitad de las páginas Web (cerca del 60%) ya no utilizan etiquetas “style” en su código fuente. Ambos gráficos de la Figura A.2i tienen barras que disminuyen hacia la derecha. La tendencia es reducir al mínimo el número de etiquetas de este tipo, ya que es más apropiado utilizar archivos CSS. Así, el código fuente de una página Web no se extiende en exceso. El tamaño de la captura de pantalla completa (webshot) es decisivo para determinar cuánto espacio consumirá nuestro dataset en un dispositivo de almacenamiento. En la Figura A.2j, el peso fluctúa en un amplio rango de valores, con una media de 300 KB. La mayoría de los valores se concentran en tamaños bajos, pero con una presencia considerable de imágenes de tamaño medio y alto. Este comportamiento requeriría no sólo una gran cantidad de espacio, sino un preprocesamiento de las imágenes para aplicaciones de ML y DL. Los gráficos de búsqueda y navegación son bastante similares para el ancho de la captura de pantalla (Figura A.2k). En ambos casos, predomina significativamente la primera barra, que muestra el mínimo, ya que la captura de pantalla por defecto establece 992

píxeles. El valor medio está muy próximo al mínimo, ya que casi todas las imágenes se capturaron con este valor por defecto (alrededor del 85%), aunque también hay imágenes con un ancho mayor, especialmente en la técnica de búsqueda con un máximo de casi 10000 píxeles. En el caso de la variable altura (Figura A.2l), también predomina el valor mínimo, aunque en menor medida (alrededor del 28%), que coincide con el valor por defecto establecido en la captura de pantalla de 744 píxeles. A diferencia del caso anterior, se observa una distribución menos desequilibrada de los valores, con una variabilidad más amplia en la que la mayor acumulación se produce hasta los 5000 píxeles, una acumulación considerable entre 5000 y 10000 píxeles y, por último, se obtienen imágenes con una altura de hasta casi 50000 píxeles. En cuanto al ancho y la altura, la mayoría de las páginas Web tienen una disposición vertical. Estos parámetros están estrechamente relacionados con la resolución o calidad de la imagen. A mayor número de píxeles, mayor es la resolución y calidad de las imágenes, aunque exige más espacio de almacenamiento.

Tabla A.1: Resumen de los indicadores estadísticos de los atributos cuantitativos de las páginas Web de navegación y búsqueda.

Atributo	**Navegación**				**Búsqueda**			
	Mín.	**Máx.**	**Media**	**Desv. Std.**	**Mín.**	**Máx.**	**Media**	**Desv. Std.**
Longitud URL	14	161	31.73	12.59	21	200	69.84	30.99
Tiempo (ms)	1.25	72.9	18.72	17.39	1.9	73.3	22.33	15.83
Tamaño (KB)	0	200.89	50.49	44.33	0	176.87	46.82	39.58
Imágenes	0	15	2.05	3.48	0	17	2.51	4.14
Scripts	0	20	3.35	5.1	0	17	3.04	4.43
Archivos CSS	0	30	5.38	8.07	0	22	4.19	5.77
Tablas	0	15	0.25	1.05	0	15	0.39	1.45
iFrames	0	14	0.16	0.62	0	15	0.18	0.6
Etiquetas style	0	15	1.25	2.37	0	15	0.8	1.6
Peso (KB)	15.75	954.23	302.72	200.18	13.59	846.77	273.92	180.25
Ancho (px)	992	6814	1058.8	388.7	992	9954	1052.3	379.6
Altura (px)	744	49658	3859	4293	744	49894	3560	3918

Por último, la tabla A.1 resume los principales indicadores estadísticos de los parámetros cuantitativos de las páginas Web, tanto para el conjunto de navegación como para el de búsqueda.

A.3 Conclusión

Los atributos cualitativos y cuantitativos del dataset de páginas Web presentado en este trabajo nos permitieron obtener información útil sobre la estructura de las páginas Web. El análisis estadístico de estos atributos mostró una distribución muy heterogénea, una

gran variabilidad y una tendencia a los valores bajos. Esto sugiere que el diseño Web sigue una regla implícita de optimización, ya que cuanto más altos son los valores, mayor es el tiempo de descarga y visualización de la página, lo que provoca incomodidad en el usuario.

Apéndice B

Aplicaciones del reconocimiento de emociones

B.1 Robótica social

- Aplicación más citada en la literatura [61][87]
- Los robots están cada vez más presentes en una amplia gama de entornos y tareas
- La interacción human-robot (HRI) será una práctica común en un futuro cercano
- Personificación de los robots
  - Forma humana
  - Mostrar una expresión facial
  - Capacidad de percibir emociones
- Comunicación más natural, emocional e inteligente
- Similar a la interacción humano-humano

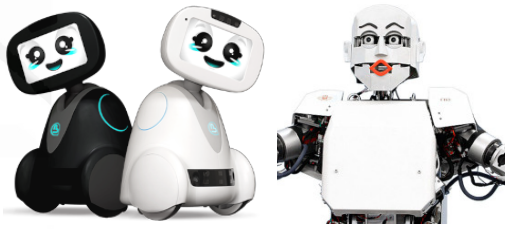


Figura B.1: Robots con expresión facial para ganar confianza y afecto de las personas.

Figura B.2: Sofía y Saya expresan emociones como los humanos sintéticamente.

B.2 Robótica médica

- Hospitales: sitios frecuentes para robots
- Ayudan en una multitud de tareas
- Coexistencia con humanos en un entorno de trabajo
- La expresión facial permite obtener la aceptación del personal y pacientes [87][89]
  - Para comunicar o tomar datos
  - Minimizar el riesgo del contacto humano-humano
  - Combatir futuras pandemias
- Casas de salud y de cuidados
  - Robots de servicio
  - Rehabilitación
  - Reducir la soledad

Figura B.3: Cirugía robótica, limpieza, registro de signos vitales y organización de estanterías.

Figura B.4: Robots con rostro. Feliz o sonriente: aceptado. Enfadado o molesto: incómodo y poco confiable.

B.3 Salud y medicina

- Expresión: indicador del estado de salud
- Diagnóstico y prevención
  - Reconocer síntomas a partir del rostro
  - Emoción negativa: stress, ansiedad, depresión
  - Prevención del suicidio
  - Predecir desórdenes psicóticos
  - Detectar enfermedades
- Monitorizar pacientes
  - En tiempo real
  - Necesidad de asistencia
  - Observación durante el tratamiento
- Tratamiento
  - Oportuno
  - Terapia (e.g., musical)
  - Reconocer emociones (autismo)

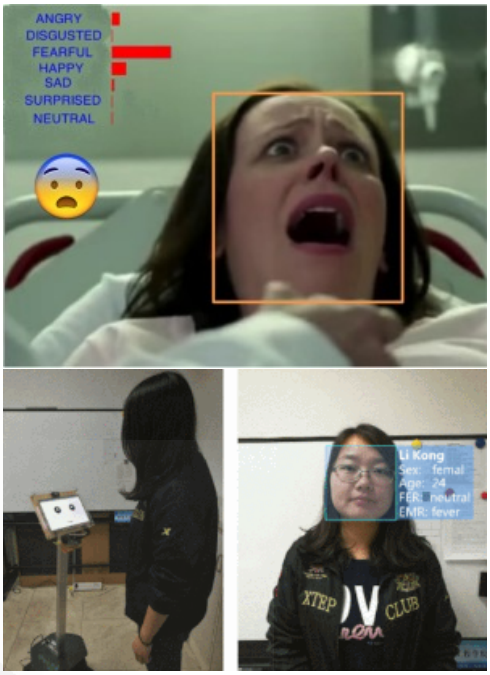


Figura B.5: Sentimientos de los pacientes sobre el tratamiento.

B.4 Seguridad vial

- Las emociones influyen en nuestro estilo de conducción
  - Negativas: más agresivos (ira, irritado, disgusto)
  - Positivas: menos atentos (distraído o entretenido)
  - En ambos casos: más accidentes
  - Lo recomendable: un estado neutral
- Sistemas FER en vehículos
  - Monitorizar en tiempo real al conductor
  - Detectar y advertir el estrés
  - Apagado automático del vehículo
- Extender a cualquier medio de transporte
- Conducción más segura para reducir accidentes fatales

Figura B.6: La ira y descuido del conductor: entre los mayores peligros en las carreteras.

B.5 Estudios de mercado

- Soporte para cualquier empresa o industria
- Previa al lanzamiento de contenidos, productos y servicios
- Analizar la reacción del consumidor
- Monitorizar la expresión facial
  - Observando programas de TV, películas o anuncios
  - Evaluación de productos
  - Videojuegos
- Ayuda a determinar el interés e intento de compra

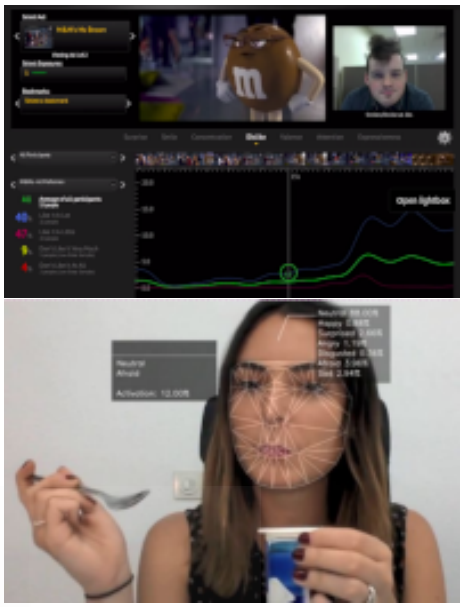


Figura B.7: Cámaras y software para detectar la expresión facial al evaluar productos.

B.6 Satisfacción del cliente/marketing

- Los productos inspiran emociones a los consumidores
  - Rechazo: falta de satisfacción
    negativas: indiferencia, disgusto, tristeza, miedo, ira
  - Aceptación: satisfacción
    positivas: felicidad, sorpresa
- Un sistema FER puede medir la satisfacción del cliente
  - Reemplazar encuestas de calidad del servicio
  - Retroalimentación para detectar deficiencias en los servicios
  - Diseñar anuncios y promociones de impacto
  - Mejorar los escaparates de los minoristas
- Herramienta de neuromarketing del consumidor

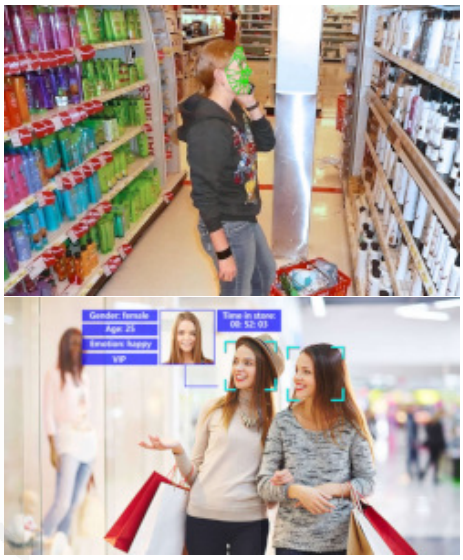


Figura B.8: La emoción es vital en las decisiones de compra.

B.7 Educación

- Expresión facial: clave en el aula [8]
  - Física o virtual
  - Rostro del estudiante: retroalimentación para el profesor (compromiso o falta de interés)
  - Rostro del profesor: cumplimiento o no de los objetivos educacionales
- Aprendizaje en línea: abandono, menos compromiso, baja calidad de la educación
- Sistema FER: detección de emociones
  - Integrado en cámaras de seguridad
  - Monitorizar atención y concentración
  - En tiempo real o procesamiento posterior
  - Detector de compromiso
  - Soporte para toma de decisiones

Figura B.9: Detectar el estado del alumno, cómo de comprometidos (o no) están los estudiantes.

B.8 Empleo, profesiones y ocupaciones

- Entrevistas de trabajo en vídeo
- Durante la entrevista
  - Evaluación del estado emocional
  - Conocer la personalidad de la persona
  - Selección de preguntas basada en la expresión
- Luego de la entrevista
  - Soporte para decisión de los reclutadores de personal
  - Identificar candidadtos adecuados
  - Atención de empleados y estilo de trabajo
  - Controlar los estados de ánimo
- Audiciones para películas, teatro y televisión

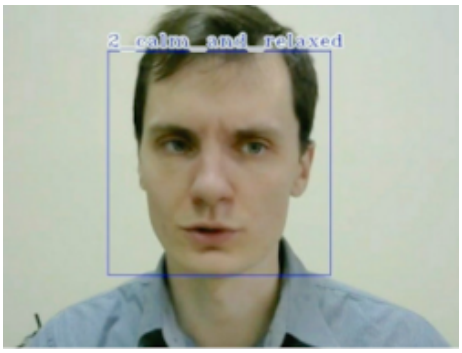


Figura B.10: Entrevistar a las personas es valioso, pero se obtienen más detalles observando sus expresiones.

Figura B.11: Evaluación de las competencias comunicativas del comentarista de noticias.

B.9 Seguridad pública

- Los delitos pueden ser guiados por las emociones [65]
- Emociones negativas
  - Sentimientos de irritación, indignación, frustración, hostilidad, enfado, furia, rabia
  - Puede provocar lesiones, homicidios, venganzas, violaciones
  - No se evalúan las consecuencias de los actos
- Cámaras con sistemas FER integrados
  - Para inferir malas intenciones
  - Comportamiento sospechoso
  - Predecir acciones futuras
  - Descubrir a los delincuentes: detección de mentiras, control fronterizo inteligente, amenaza terrorista, uso de cajeros automáticos/ATM
- Herramienta de investigación, detección, prevención y predicción de delitos

Figura B.12: La policía escanea los rostros mediante cámaras de vigilancia. Un cajero automático no dispensa dinero si el usuario está asustado.