Universidad Autónoma de Madrid

Escuela Politécnica Superior

22|23

Trabajo fin de grado

Hacia una visión más justa: abordando el desbalanceo de clases en el reconocimiento de género en imágenes

Daniel Barahona Martín

Escuela Politécnica Superior
Universidad Autónoma de Madrid
C\Francisco Tomás y Valiente nº 11

Campus Internacional excelencia UAM CSIC

www.uam.es

UNIVERSIDAD AUTÓNOMA DE MADRID
ESCUELA POLITÉCNICA SUPERIOR

Grado en Ingeniería Informática

TRABAJO FIN DE GRADO

Hacia una visión más justa: abordando el desbalanceo de clases en el reconocimiento de género en imágenes

Autor: Daniel Barahona Martín
Tutor: Lara Quijano

julio 2023

Daniel Barahona Martín
Hacia una visión más justa: abordando el desbalanceo de clases en el reconocimiento de género en imágenes

Daniel Barahona Martín
C\ Francisco Tomás y Valiente Nº 11

IMPRESO EN ESPAÑA – PRINTED IN SPAIN

A mis padres

I was an ordinary kid who studied hard. There are no miracle people. It happens they get interested in this thing and they learn all this stuff, but they're just people.

Richard P. Feynman

Agradecimientos

Este Trabajo Fin de Grado culmina mi paso de cuatro años por la Escuela Politécnica Superior. Quiero expresar mi más sincero agradecimiento a todas las personas que me han ayudado a llegar a este momento en el que finaliza mi etapa en esta Universidad.

En primer lugar, me gustaría agradecer a mis padres y abuelos por su apoyo incondicional. Gracias por ser una fuente constante de inspiración y motivación en cada paso de este proceso. Sin su guía y apoyo, nunca habría llegado hasta aquí.

Asimismo, no puedo dejar de mencionar a mi tutora, Lara Quijano, por su orientación, su dedicación y su experiencia en el campo de la investigación. Gracias por su paciencia y por compartir conmigo sus conocimientos y consejos, los cuales me han ayudado a crecer tanto académica como personalmente. También quiero agradecer a mi profesora de física de secundaria, Gloria Moro, por infundirme confianza a la hora de afrontar retos difíciles e introducirme al campo de la ciencia y la tecnología.

Quiero mencionar también a aquellas personas que, sin formar parte directa de este proyecto, han sabido transmitirme su confianza y apoyo a lo largo de estos meses. Gracias a Héctor, Sergio, Pablo, Tracy, Carlos, David y Dani por estar siempre ahí.

Por último, quiero recordar al Dr. Richard Feynman (1918-1988), el célebre docente y astrofísico cuya filosofía de aprendizaje y humildad trato de aplicar cada día.

Resumen

En la sociedad actual, cada vez más procesos y dominios están incorporando técnicas de aprendizaje automático y visión artificial, en especial para la clasificación de características faciales (por ejemplo, el reconocimiento de género usado en sistemas de videovigilancia, o demoscópica en redes sociales). Pese a todas sus aplicaciones beneficiosas, existen ciertos temores asociados a los sesgos subyacentes a estas técnicas, ya sea por la propagación de sesgos existentes en la sociedad, o las dudas concernientes al diseño de los algoritmos y datos. Es por tanto crucial abordar las distintas fuentes de sesgo para conseguir resultados verdaderamente objetivos.

Este trabajo se enfoca en el problema del sesgo causado por desbalanceo de clases, cuya mitigación puede resolverse con técnicas de perfil algorítmico, o a nivel de datos. Se han detectado carencias en la bibliografía revisada, en cuanto a que faltan estudios que comparen ambos tipos de técnicas, y determinen cuál es más adecuado. Aparte, se quiere saber si las condiciones concretas de desbalanceo y complejidad de los datos influyen en la elección de un método mitigante. Todas estas brechas bibliográficas se han condensado en forma de tres preguntas de investigación. Para resolverlas, se han diseñado varios experimentos en los que se contrastan diferentes métodos de la bibliografía aplicados a dos datasets: UTKFace y PlantVillage (de características y complejidad diferentes), sobre los que se aplican diferentes escenarios de desbalanceo. Además, se propone un método novedoso de mitigación del desbalanceo a nivel de algoritmo, llamado Loss Focal Difusa. Los resultados obtenidos muestran la elevada eficacia de los métodos de sobremuestreo con generación sintética por medio de redes generativas antagónicas (GAN), en su variante *Wasserstein* GAN, frente a los métodos de mitigación a nivel algorítmico. Entre éstos, el método nuevo propuesto obtiene las mejores métricas.

Con todo, los métodos algorítmicos pueden resultar más prácticos y rápidos de aplicar cuando la complejidad del problema no es muy elevada, lo cual se ha cuantificado mediante métricas de distancias y entropía de imágenes. Finalmente, los resultados experimentales sugieren que, según las prioridades del investigador, la aplicación de estos métodos puede ser la única vía para mitigar el desbalanceo, pues su tiempo de entrenamiento es menor que los de las GAN, y no corren el riesgo de generar características falsas en muestras artificiales.

Palabras clave

Aprendizaje automático, clasificación en imágenes, mitigación de sesgo, desbalanceo de clases, redes neuronales, lógica difusa

Abstract

In today's society, an increasing number of processes and domains are incorporating machine learning and computer vision techniques, particularly for facial feature classification (e.g., gender recognition used in surveillance systems or demographic analysis in social networks). Despite their beneficial applications, there are concerns associated with the biases inherent in these techniques, whether due to the propagation of existing biases in society or doubts regarding algorithm and data design. Therefore, it is crucial to address the various sources of bias to achieve truly objective results.

This work focuses on the problem of bias caused by class imbalance, which can be mitigated using algorithmic or data-level techniques. Some gaps have been identified in the reviewed literature, as there is a lack of studies comparing both types of techniques and determining which is more suitable. Additionally, it is of great interest to investigate whether specific conditions of imbalance and data complexity influence the choice of a mitigation method. These gaps in the literature have been condensed into three research questions. To address them, several experiments have been designed to compare different methods from the literature applied to two datasets: UTKFace and PlantVillage, which have different characteristics and complexity. Different imbalance scenarios are applied to these datasets. Furthermore, a novel algorithm-level mitigation method called Diffuse Focal Loss is proposed. The results obtained demonstrate the high efficacy of synthetic oversampling methods using Generative Adversarial Networks (GANs), specifically the *Wasserstein* GAN variant, compared to algorithm-level mitigation methods. Among these methods, the proposed novel approach achieves the best metrics.

However, algorithmic methods may be more practical and faster to apply when the complexity of the problem is not very high, as quantified by distance metrics and image entropy. Finally, the experimental results suggest that, depending on the researcher's priorities, the application of these methods may be the only way to mitigate class imbalance, as their training time is way shorter than that of GANs and they do not risk generating false features in synthetic samples.

Keywords

Machine learning, image classification, bias mitigation, class imbalance, neural networks, fuzzy logic

Índice

Listas

Lista de algoritmos

Lista de códigos

Lista de cuadros

Lista de ecuaciones

Lista de figuras

Lista de tablas

Lista de cuadros

1

Introducción

En la sociedad, cada vez son más los procesos que pueden automatizarse gracias a la Inteligencia Artificial (IA), empleando algoritmos basados en aprendizaje automático [1]. Sin embargo, al aplicarse a los datos del mundo real, estos algoritmos pueden verse afectados por diversas fuentes de sesgo, entre las que pueden encontrarse la insuficiente representatividad de una clase con respecto a otra [2], el etiquetado poco preciso [3], el solapamiento ocasional entre clases distintas [4], o la desproporción en la cantidad de información que aportan ciertas muestras con respecto a otras [5]. A causa de esto, en ocasiones las ventajas de la IA pueden verse eclipsadas por el temor a que los sesgos presentes se reproduzcan y propaguen en la sociedad [6]. Es por ello que es imperativo ajustar los algoritmos y datasets utilizados en el aprendizaje para minimizar al máximo todas las posibles fuentes de sesgo. Este proyecto está orientado a ese objetivo, donde, dada la imposibilidad de abarcar todas las fuentes de sesgo en un único trabajo, se opta por estudiar en profundidad el sesgo causado por el *desbalanceo de clases* [7].

El desbalanceo de clases en un dataset se define como una situación en la que una o varias clases se encuentran infrarrepresentadas numéricamente en relación a otra clase (o conjunto de clases) [8]. Los motivos detrás del desbalanceo pueden ser muchos: el grupo minoritario está infrarrepresentado en la vida real (por ejemplo, las desigualdades de género presentes en muchos ámbitos) [9]; es imposible recolectar el mismo número de muestras para cada clase (pensemos en las radiografías de enfermedades raras) [10]; o incluso el desconocimiento o equivocación de aquellos encargados de recolectar y anotar los datos [3]. Sea cual sea la causa, ante la aparición de este sesgo los diseñadores de algoritmos de IA deben tratar de reconocerlo, y actuar de acuerdo a ello.

De entre los dominios que se ven afectados por esta problemática, el reconocimiento facial (dentro del campo de la visión artificial) ocupa un lugar importante, dado el crecimiento de estos modelos en campos como la psicología, la interacción con vehículos autónomos, o los sistemas de videovigilancia policiales [11, 12]. Generar modelos fiables de clasificación de imágenes faciales no es tarea sencilla, y el rendimiento de dichos modelos se ve condicionado por la existencia de conjuntos de datos suficientemente grandes y diversos [11] que representen todas las etiquetas que se pretende cubrir (género, raza, edad...). Más en detalle, la clasificación de género, que pretende discernir el género

biológico [1] de una persona basándose en distintas características faciales, ha cobrado popularidad en la última década por sus múltiples aplicaciones, sobre todo en sistemas de videovigilancia o estudios demoscópicos en redes sociales [6, 13]. Por ejemplo, los anuncios que provocan la identificación con el género propio son más probables de tener un impacto favorable en la imagen de la marca desde la perspectiva de los consumidores [14]. Mitigar los sesgos presentes en estos algoritmos es esencial no sólo para lograr una mejora técnica en el rendimiento de los modelos, sino también para trasladar esa idea de igualdad que se persigue en la sociedad.

Tras una extensa revisión bibliográfica centrada en el sesgo causado por el desbalanceo de clases, se ha observado que no existe un consenso entre los autores respecto a cómo tratarlo. Las diversas metodologías [15, 16, 17, 18, 19] pueden agruparse en dos tipos de estrategias a grandes rasgos: i) aquellas que atacan el problema del desbalanceo alterando los conjuntos de datos [8, 20, 21], y ii) aquellas que lo hacen modificando los algoritmos de aprendizaje [10, 22, 23, 24]. Pero, a día de hoy, faltan comparativas concluyentes entre ambos tipos de estrategias (de datos y algorítmicas), limitándose la mayoría de revisiones [2, 8, 20] a poner en contraste métodos del mismo tipo [21]. Esta falta de comparativas entre ambos tipos de estrategias puede deberse a la heterogeneidad de los entornos en los que se trabaja cada técnica (clasificaciones de imágenes faciales, de plantas, de radiografías...), pero también a la falta de puntos en común como serían el aplicar cada técnica a un mismo conjunto de datos. Aparte, las distintas publicaciones recogidas en las revisiones a menudo registran métricas distintas [8, 10, 20, 21, 22, 23, 24], con lo que es complicado determinar a priori qué método es objetivamente mejor. En general, hay cierta confusión con respecto a si existen metodologías aplicables a múltiples escenarios, o son necesarias soluciones *ad hoc*.

Por todo ello, en este Trabajo Fin de Grado se pretende hacer esa comparación entre técnicas de los dos tipos sobre un "punto en común", que permita establecer pautas para atacar el sesgo por desbalanceo de clases en función de cada escenario de forma más completa. Así, al comienzo de esta investigación se formulan las siguientes hipótesis:

Hipótesis 1: Es posible mejorar *las bases de datos* de imágenes usadas en problemas de clasificación para mitigar el sesgo por desbalanceo de clases.

Hipótesis 2: De la misma forma, es posible mejorar *los algoritmos* de clasificación de imágenes para el mismo fin.

Para responder a estas hipótesis se han formulado las siguientes preguntas de investigación (*Research Questions*, en adelante RQ):

RQ1: Para mitigar el sesgo por desbalanceo, ¿es siempre mejor atacar los datos, los algoritmos, o depende de cada situación?

[1] No se debe confundir esta clasificación con la distinción del *género percibido*, ya que la naturaleza subjetiva de éste lo convertiría en un problema totalmente distinto. De hecho, sería más apropiado referirse al problema como "clasificación de sexo" y no de "género", pero en este trabajo se opta por lo segundo, ya que es la terminología usada en la literatura especializada.

RQ2: ¿Es necesario conocer de antemano si el conjunto de datos presenta desbalanceo y, en su caso, el grado de desbalanceo, para determinar el método a aplicar?

RQ3: ¿Cómo afecta la complejidad del problema a la toma de decisiones sobre el mejor método de mitigación del desbalanceo a aplicar?

Para responder a estas preguntas, se ha realizado una comparación entre las diferentes estrategias identificadas en la bibliografía a nivel de algoritmos y de datos. Además, de forma novedosa, se propone un método de mitigación a nivel de algoritmo, denominado *Loss Focal Difusa*, que se ha comparado junto con el resto de métodos. Por último, para dotar de algo más de generalidad a las conclusiones, se ha tratado de extender esta comparativa a otros dominios de la clasificación de imágenes aparte del de reconocimiento de género facial, pero manteniéndonos en el dominio de clasificación binaria. Por ello, se han empleado dos datasets: uno de género en imágenes y otro de infecciones en plantas de tomate.

El desarrollo de este Trabajo Fin de Grado ha conducido finalmente a las siguientes contribuciones:

1: Se aporta una comparativa en común de estrategias de mitigación del desbalanceo de clases, tanto a nivel de algoritmo como de datos.

2: Se estudia el impacto de la complejidad del dominio del problema a la hora de dirimir la adecuación de una familia de métodos u otros.

3: Se propone un método algorítmico propio, que mezcla conceptos de lógica difusa con funciones de pérdida adaptadas al desbalanceo, llamado Loss Focal Difusa.

El resto de este trabajo se ordena de la siguiente forma: el Capítulo 2 realiza una revisión del estado del arte más relevante de acuerdo a los objetivos de este proyecto, y justifica la formulación de las preguntas de investigación. El Capítulo 3 expone cada una de las técnicas comparadas en los experimentos (tanto las estrategias de datos como las algorítmicas), así como el funcionamiento del método algorítmico propio, la Loss Focal Difusa. En el Capítulo 4 se presenta el diseño de cada uno de los experimentos y pruebas llevadas a cabo, así como los datasets sobre los que se realizan. Los resultados de dichos experimentos se discuten y comparan en el Capítulo 5. Para terminar, el Capítulo 6 expone las conclusiones extraídas de este proyecto de investigación, y las posibles líneas de investigación futura.

2

Trabajo relacionado

En este capítulo se realiza una revisión del estado del arte, comenzando por describir el dominio de interés: el problema de la clasificación de género en imágenes. Hecho esto, se pasa a analizar el problema del sesgo en dichos algoritmos de clasificación, y dentro de este, el sesgo causado por el desbalanceo de clases, en el cual se centra el resto de este trabajo. Continuamos analizando los métodos de la bibliografía empleados para mitigarlo, tanto a nivel de datos como de algoritmo. Finalmente, se analizan los *research gaps* detectados en el estado del arte, justificando la formulación de las preguntas de investigación que han guiado este trabajo. El protocolo de búsqueda bibliográfica se describe en el Apéndice A.

2.1. El problema de la clasificación de género

La clasificación de género facial (o cualquier otro tipo de clasificación basada en imágenes) es un proceso que, a rasgos generales, suele dividirse en dos fases. La primera consiste en la extracción de características en forma de vectores de las imágenes que componen el dataset, y en la segunda fase se envían dichos vectores a un modelo que predice la etiqueta final, como puede ser una Máquinas de Vector Soporte (SVM) ó un Perceptrón Multicapa (MLP). En el estado del arte estudiado, los algoritmos de clasificación se dividen entre los que usan clasificadores tradicionales para aprender de la información de los vectores de características [9, 25, 26, 27], y los que utilizan Redes Neuronales Convolucionales (CNNs) para unificar las fases de extracción y predicción en una única arquitectura [14, 28, 29, 30, 31, 32]. Dichas familias de métodos se resumen a continuación:

2.1.1. Métodos tradicionales

En estas técnicas, el primer paso es realizar una extracción de características manual, pre-procesando las imágenes para extraer sus Histogramas de Gradientes Orientados (HOG) [33], o Patrones Locales Binarios (LBP) [9]. Luego, estos conjuntos de características sirven para alimentar el aprendizaje de modelos de mayor o menor complejidad. Una técnica común a muchos trabajos es el uso de un clasificador sencillo como las SVM [9, 25, 26, 27, 34, 35], especializado en problemas de clasificación

binaria; lo cual, por otro lado, lo hace ideal para un problema como la distinción de género [27]. De forma similar, en Poornima et al. [13] se compara el rendimiento del SVM con respecto al *Random Forest* (RF), o su variante *Adaptive Boosting* (AdaBoost), concluyendo que RF puede obtener un rendimiento notablemente superior. Estudios más complejos y recientes han empleado técnicas de aprendizaje evolutivo [36], pero también los hay que emplean métodos tan clásicos como los K vecinos próximos [26].

2.1.2. Redes convolucionales

No obstante, en comparación con los métodos tradicionales, en el campo de la clasificación de género las CNNs han probado tener el mejor rendimiento [14, 37, 38]. Esto es debido a que los filtros de extracción de características se optimizan automáticamente en el entrenamiento de la propia red, mientras que en los métodos anteriores esta extracción se realiza manualmente en una fase previa. Además, cuando se usan algoritmos tradicionales, basados en vectores de píxeles, se puede perder información muy valiosa sobre la interacción espacial entre ellos. Las CNNs pueden extraer esa información de los "vecindarios de píxeles" de forma eficiente, por medio de la convolución, y luego usar esa información en las capas densamente conexas para la predicción final.

Los modelos convolucionales más relevantes son los basados en VGG [39] y ResNet [40], aunque hay otros como GoogLeNet o Inception [41]. Por ejemplo, en Gurnani et al. [42], una combinación de VGG-16 para la extracción de características faciales con un clasificador basado en AlexNet [43] da lugar a un clasificador combinado de edad, género y expresión facial. O en el trabajo de Zhang et al. [32] proponen una VGG-19 con imágenes RGB-D (combinación de canales de color y un "mapa de profundidad") para detectar el género en imágenes de cuerpo entero. Por otro lado, en Lin and Xie [26] un extractor de características basado en las capas convolucionales de ResNet-50 se une a un clasificador SVM para distinguir género en imágenes faciales. Por último, algunos estudios proponen arquitecturas prometedoras construidas prácticamente desde cero: la red propuesta en Sumi et al. [31] (aunque lejanamente inspirada en VGG) logra resultados comparativamente superiores a métodos de estado del arte que combinaban LBP con redes neuronales no convolucionales.

2.2. Sesgos en la clasificación de imágenes

Expuesto el dominio de la clasificación de imágenes, se presentan los sesgos que pueden surgir en estos algoritmos. Ello permite introducir el sesgo por desbalanceo de clases, en el que se centra el resto de esta investigación.

El sesgo en el aprendizaje automático se produce cuando un algoritmo presenta resultados con un prejuicio sistemático debido a asunciones erróneas en el proceso de aprendizaje [2]. La literatura es-

tudiada [4, 5, 8, 44, 45] identifica múltiples fuentes de sesgo posibles en los algoritmos de clasificación en imágenes, como las que se describen a continuación:

- En primer lugar, el sesgo puede derivar de que los etiquetados de las muestras sean incorrectos. Los errores en el etiquetado pueden deberse a errores en etiquetadores automáticos (no supervisados), o al fallo de anotadores humanos. Por ejemplo, en Kafkalias et al. [3] se explora el sesgo introducido en las etiquetas en función del género y la raza del propio anotador, y se trata de cuantificarlo. Así, algunos estudios del estado del arte dan lugar a técnicas de "etiquetado suave" [12] que defina automáticamente las fronteras entre clases; o diseñan modelos capaces de evaluar si unas u otras etiquetas se adecúan a cada muestra utilizando métodos de *clustering*, como Aka et al. [4].
- En segundo lugar, está el sesgo producido por conjuntos de datos que no representan cuantitativamente la distribución de la realidad o que tienen una falta de representación para ciertos valores. Para nuestro dominio, si se trata de crear un algoritmo de clasificación que tenga en cuenta el género, deberá asegurarse una distribución paritaria entre las clases "hombre" y "mujer", o de lo contrario existirá una descompensación en el rendimiento del modelo al predecir la clase más numerosa frente a la otra. En consecuencia, el algoritmo aprende mejor las clases más representadas, pero se encuentra con dificultades a la hora de distinguir el resto. Se dice entonces que hay un "desbalanceo" entre las clases [21], y es en este problema en el que se centra nuestra investigación. Las soluciones propuestas al desbalanceo van desde el aumento de datos de las clases minoritarias (bien con copias modificadas [8], bien con generación sintética de muestras [46, 47, 48, 49]) hasta el refinado de los algoritmos para que presten especial atención a dichas clases infrarrepresentadas. Nos referiremos a estas familias de técnicas en las Secciones 2.3.1 y 2.3.2, respectivamente.
- En tercer lugar, es fundamental tener presente que para lograr predicciones óptimas, no solo es necesario abordar el desbalanceo de clases, sino también considerar la representación adecuada de todos los atributos relevantes [1, 14, 36]. Por ejemplo, aunque un conjunto de datos pueda contener un número significativo de imágenes que representen tanto hombres como mujeres, será insuficiente si no se ha considerado adecuadamente la diversidad étnica de las personas. Según Loo et al. [25], diferentes etnias pueden exhibir características distintivas de género, lo cual implica que un conjunto de datos compuesto predominantemente por individuos de origen caucásico no será eficaz al clasificar personas de otras razas.
- En cuarto lugar, hay que considerar que algunas técnicas de mitigación, como la generación sintética de muestras, pueden tener efectos adversos e incluso introducir sus propios sesgos [17]. Crear muestras que no existen en la realidad puede inducir fallos no intencionados en los sistemas de visión artificial. Por ejemplo, en Rahimzadeh and Attar [10] se desaconseja la generación de imágenes de radiografías para distinguir pulmones enfermos de los sanos, ya que a fin de cuentas se estaría generando "pacientes" que no existen realmente. Esto se traduce en un sesgo entre las imágenes reales y las sintéticas [50]. Estudios como Yan et al. [17] usan funciones de pérdida propias para influir el proceso de retropropagación, logrando mayor separación entre instancias de diversas clases; y más proximidad entre instancias dentro de la misma clase.
- Por último, otras fuentes secundarias de sesgo pueden ser el ruido presente en las propias imágenes [5, 19]; o el llamado "sesgo contextual" [51], en el que por ejemplo, un algoritmo

entrenado exclusivamente con imágenes de hombres haciendo deporte tendría dificultades a la hora de detectar mujeres realizando las mismas actividades.

2.3. Mitigación del sesgo por desbalanceo de clases

Vistas las posibles fuentes de sesgo, este trabajo se centra en la mitigación del sesgo producido por el desbalanceo de clases, debido a su impacto en la precisión de los modelos, la necesidad de equidad en la IA, y su naturaleza inherente a los datos de entrenamiento. Abordar las demás fuentes de sesgo es material para trabajo futuro. Las múltiples estrategias existentes para este fin pueden agruparse en dos familias: aquellas que atacan el problema del desbalanceo alterando la distribución de los datasets [21], y aquellas que lo hacen modificando los algoritmos de aprendizaje [10, 22, 23, 24] para que se preste más atención a las muestras de clases minoritarias. Las siguientes sub-secciones resumen los estudios bibliográficos de ambas familias de estrategias.

2.3.1. Técnicas de mitigación a nivel de datos

El objetivo de estas técnicas es modificar las distribuciones de los datos para disminuir el grado de desbalanceo y/o reducir el ruido. Matemáticamente, se define el "ratio de desbalanceo" (IR) de un dataset [4, 8, 21] como:

$$IR = \frac{tamaño_{clase\ grande}}{tamaño_{clase\ pequeña}} \geq 1 \tag{2.1}$$

El objetivo al aplicar estas técnicas es lograr un valor del IR cercano o igual a 1. Hecho esto, se entrena un modelo sobre la distribución modificada, comparando si su rendimiento mejora frente a un entrenamiento con la distribución original. Existen múltiples técnicas para lograr este fin, desde las más simples a otras más complejas.

Entre los métodos más sencillos, el submuestreo aleatorio (*random undersampling*) descarta muestras al azar del grupo mayoritario; mientras que el sobremuestreo aleatorio (*random oversampling*) duplica al azar muestras del grupo minoritario [8], pudiendo rotar o transponer imágenes, ampliar ciertas zonas o cambiar los colores a través de filtros de contraste. En Lin et al. [52] se usa un método basado en *clusters* de K-means para submuestrear la clase mayoritaria a los K centroides calculados, con K igual al número de muestras de la clase minoritaria. Por su simpleza, estos métodos son poco efectivos: el submuestreo descarta datos, reduciendo la cantidad de información de la que los modelos disponen para aprender [8]; y el sobremuestreo tiende a producir sobreajuste (*overfitting*), aún si se acompaña de cambios aleatorios en las copias (rotación, reflejo...), pues no aporta suficiente variedad [47, 49].

Por ello, en la bibliografía se han desarrollado métodos de sobremuestreo "informados" que refuer-

cen las fronteras inter-clase, reduzcan el sobreajuste y mejoren la clasificación. Destaca el *Synthetic Minority Oversampling Technique* (SMOTE) [49], una técnica que produce nuevas muestras minoritarias artificiales por medio de interpolar muestras existentes con sus vecinos cercanos. Algunas variantes, como *borderline*-SMOTE y ADASYN [21], tratan de mejorar la calidad del concepto original en las fronteras inter-clase, así como prestar más atención a las muestras más difíciles de aprender. Los estudios al respecto coinciden en que SMOTE es sustancialmente mejor que los métodos clásicos de sub y sobremuestreo [8, 21, 49]. Por ejemplo, en Nafi and Hsu [21], se consigue una F1-score del 62 % en SMOTE frente al 58 % del sobre/submuestreo.

No obstante, las técnicas como SMOTE y similares tienen un límite en la cantidad de variedad e "información útil" que pueden introducir en las muestras sintetizadas [21]. Es por ello que recientemente se han desarrollado modelos más complejos para aumentar la calidad de las muestras minoritarias, mejorando las métricas existentes hasta el momento [8, 53]. Destacan aquí la familia de las redes antagónicas generativas (GAN) [8, 17, 46, 47, 53]. Una GAN es un modelo compuesto de dos redes neuronales: generador y discriminador. El objetivo del generador es crear artificialmente resultados que pretenden confundirse con datos reales. El discriminador trata de identificar cuándo una muestra es real o sintética. A medida que continúa el ciclo de entrenamiento entre las redes antagónicas, el generador comenzará a producir resultados "más creíbles", y el discriminador aprenderá a distinguirlos de los verídicos. Existen numerosas variantes de esta técnica, entre las que destacan la DCGAN [21], que usa redes profundas para estabilizar el aprendizaje; y la WGAN [47], que imita a la anterior pero introduce la función de pérdida *Wasserstein-1* para aproximar mejor la distribución de los originales. En el mencionado artículo de Nafi and Hsu [21], la WGAN supera a todos los métodos anteriormente enunciados, alcanzando una F1-score del 65 %.

2.3.2. Técnicas de mitigación a nivel de algoritmo

En determinadas ocasiones, no siempre se puede alterar artificialmente la distribución de los datos [17]. Véase el ejemplo de un dataset compuesto de radiografías para una determinada enfermedad [10]; Usar aquí métodos generativos, como las GAN, implica crear radiografías de pacientes que realmente no existen, pudiendo llegar a "confundir" al algoritmo [10]. Además, los métodos a nivel de datos tienen el inconveniente de añadir tiempo extra al entrenamiento de los clasificadores [8, 12], pues hay un paso previo de entrenamiento y generación sobre el dominio objetivo (imágenes faciales, radiografías...).

Por todo ello, la otra familia de métodos que mitigan el desbalanceo de clases lo hacen modificando la estructura interna de los algoritmos clasificadores, sin alterar la de los datasets [12, 54, 55]. Así, se persigue que todas las clases tengan una contribución más igualitaria al modelo, pese a no estar balanceadas en número absoluto de muestras. Varios estudios [2, 7, 18, 19, 22, 56, 57, 58] utilizan funciones de pérdida dinámicas para ajustar el nivel de contribución de cada clase al aprendizaje, y

quizás la más representativa es la Loss Focal de Lin et al. [22]. Otro ejemplo es Li et al. [5], que diseña un sistema de reponderación ubicua (del inglés *Ubiquitous Reweighting*) para balancear la contribución de clases y muestras individuales a la pérdida de entropía cruzada.

Otros autores aplican el llamado "suavizado de etiquetas" [55, 59] al problema del desbalanceo, si bien su efectividad sólo llega a ser notable en entornos con gran cantidad de clases (habitualmente más de 50), por lo que pierde interés para este trabajo. En Liu et al. [60], se resuelve la infra-representación de radiografías de pacientes con covid-19 frente a los sanos con un sistema mixto que utiliza dos redes convolucionales y un mecanismo de repetición para que el modelo visite más frecuentemente las muestras minoritarias. Este sistema se conoce como PRM-IM (*Probabilistic Relational Model for IMbalanced learning*), y en Guo and Viktor [61] incluyen la creación de conjuntos de clasificadores (*ensembles*) para mejorar el rendimiento de un único modelo . Por último, en Khan et al. [23] se usan matrices de costes para imponer penalizaciones más altas al algoritmo cuando éste clasifica erróneamente una muestra de una clase minoritaria, que cuando lo hace con una mayoritaria. Otros estudios [7, 56, 57] usan sistemas de control por lógica difusa para adaptar algoritmos de índole diversa a entornos desbalanceados.

2.4. Resumen

Tras esta revisión bibliográfica se detectan las siguientes brechas de investigación (*research gaps*) que dan lugar a nuestra propuesta de investigación.

Del estudio realizado podemos deducir que la problemática del sesgo por desbalanceo de clases en la clasificación de imágenes es un asunto estudiado extensivamente [4, 5, 8, 44, 45]. La frecuencia con la que los datasets de imágenes del mundo real presentan infrarrepresentación para ciertas clases hace que sea imperativo ahondar en este problema y buscar maneras de paliarlo (especialmente en ámbitos en los que se persigue la igualdad plena, como es el reconocimiento de género en caras humanas). A grandes rasgos, la investigación se divide en métodos mitigantes que modifiquen las distribuciones de datos, y métodos que alteren el aprendizaje de los propios algoritmos.

Un *research gap* que se ha detectado es que las revisiones paraguas tratan únicamente la comparativa entre métodos de la misma familia [8, 21]. Por ejemplo, en Nafi and Hsu [21] se comparan únicamente métodos a nivel de datos (sub/sobremuestreo, SMOTE y GAN) sobre un problema de clasificación binaria, concluyendo que las mejores métricas de clasificación provienen de los datos aumentados con GAN. Otro *research gap* aparece en el caso de los métodos de mitigación algorítmicos, para los que ninguna revisión aporta conclusiones sobre qué método impera sobre el resto, ni en qué situaciones. Según Johnson and Khoshgoftaar [8], esto es debido a que las distintas publicaciones cubren conjuntos de datos con grados de desbalanceo diferentes, y aportan métricas de rendimiento inconsistentes. Al limitarse a exponer los resultados de varios artículos, sin poner un conjunto de mé-

tricas ni datasets comunes, sus resultados son difícilmente interpretables, puesto que cada uno viene de un dominio y condiciones de desbalanceo distintas. La raíz del problema reside en que hasta el momento no se ha encontrado una comparativa concluyente que explore si una familia de técnicas (datos *versus* algoritmos) es inherentemente más adecuada que la otra, o si por el contrario distintos escenarios se verían beneficiados por técnicas distintas. Recordemos además las hipótesis formuladas al comienzo de esta investigación:

Hipótesis 1: Es posible mejorar *las bases de datos* de imágenes usadas en problemas de clasificación para mitigar el sesgo por desbalanceo de clases.

Hipótesis 2: De la misma forma, es posible mejorar *los algoritmos* de clasificación de imágenes para el mismo fin.

En consecuencia, para poder probar dichas hipótesis, y con el fin de afrontar los *research gaps* enumerados, se formula la primera pregunta de investigación:

RQ1: Para mitigar el sesgo por desbalanceo, ¿es siempre mejor atacar los datos, los algoritmos, o depende de cada situación?

Al formular esta pregunta, cabe también cuestionarse cuál es el nivel de reproducibilidad de los métodos “mejores” en dominios y escenarios de desbalanceo distintos. Es intuitivo pensar que si el desbalanceo del problema cambia, métodos como las GAN deberían pasar por una nueva fase de generación para adaptar el nuevo dataset. Así, cabe preguntarse si esto sucede en otras técnicas. Para ello, se formula una nueva pregunta de investigación:

RQ2: ¿Es necesario conocer de antemano si el conjunto de datos presenta desbalanceo y, en su caso, el grado de desbalanceo, para determinar el método a aplicar?

Y bien es cierto que, en problemas con una complejidad pequeña o donde las muestras presentan alto grado de similaridad (radiografías, infecciones botánicas...), pueden ser útiles los modelos de generación sintética de muestras. Pero, ¿qué ocurre cuando el espacio de representación aumenta, por ejemplo, al incluir varias etnias y edades en el dominio del género facial? ¿Existen escenarios en los que sería incorrecto usar muestras que no existen en el mundo real? Resolver estas preguntas relativas a la complejidad es esencial a la hora de dotar de transparencia a las futuras investigaciones que aborden estos escenarios, y con ese fin se formula la tercera pregunta de investigación:

RQ3: ¿Cómo afecta la complejidad del problema a la toma de decisiones sobre el mejor método de mitigación del desbalanceo a aplicar?

En el siguiente capítulo se detalla la propuesta de resolución de cada una de las preguntas de investigación formuladas.

3

Métodos empleados

Para resolver la **RQ1**, y como primera contribución de este trabajo, es necesario comparar métodos representativos de ambas estrategias sobre los mismos datasets. Recogiendo además las mismas métricas para todas las pruebas realizadas, se pretende llegar a un "punto en común" sobre el que obtener respuestas concluyentes. La **RQ2** se puede resolver una vez determinada la respuesta de la primera, pues los "mejores" métodos determinarán si es necesario o no conocer de antemano el desbalanceo presente en los datos. Además, la **RQ2** motiva que los experimentos se repliquen sobre un mismo dataset con distintos grados de desbalanceo. Por tanto, el primer paso para resolver ambas preguntas es escoger qué métodos mitigantes de cada familia de estrategias se propone comparar.

En primer lugar, elegir los métodos de mitigación a nivel de datos es más sencillo, al haber comparativas concluyentes en la bibliografía [20, 21]. Como se ha mencionado, en Nafi and Hsu [21], se compara el rendimiento del submuestreo, el sobremuestreo, el SMOTE y la GAN para mejorar el rendimiento en un clasificador de imágenes de plantas de tomate con y sin infecciones. Los autores llegan a la conclusión de que las GAN son, de lejos, el método más efectivo de mitigación del desbalanceo a nivel de datos. Esta conclusión ha sido ratificada por otras revisiones, como Johnson and Khoshgoftaar [8]. Para empezar, el sub/sobremuestreo tienen una marcada tendencia al sobreajuste y reportan las métricas más bajas [8, 20, 21]. Por su parte, SMOTE no logra que las muestras artificiales sean lo suficientemente parecidas a las reales como para aportar "variedad útil" al conjunto de datos [21]. En cambio, la arquitectura convolucional de las GAN las convierte en algoritmos mucho más potentes, capaces de generar imágenes artificiales con un sorprendente nivel de realismo [8]. En concreto, se ha escogido la variante WGAN [47], ya que como se muestra en Nafi and Hsu [21], la función de pérdida *Wasserstein-1* permite un mejor ajuste del sistema y produce mejores resultados que el esquema clásico o la DCGAN. En cuanto a las demás técnicas bibliográficas de esta familia, la Tabla 3.1 resume el motivo por el que no se consideran para las comparativas.

En segundo lugar están los métodos de mitigación a nivel de algoritmo. Durante el análisis bibliográfico se ha notado una ausencia de comparativas concluyentes que afirmen la superioridad de un método sobre el resto, como sí ocurre con las estrategias a nivel de datos. Con todo, resalta la popularidad de métodos que modifican las funciones de pérdida a fin de modificar la "atención" dada por el modelo a las clases mayoritaria y minoritaria: en especial la Loss Focal (con un 91 % de Accuracy en

Método	Descripción	Motivo de descarte
Submuestreo aleatorio (RUS)	Descartar muestras de la clase minoritaria hasta alcanzar el equilibrio.	Métricas notablemente más bajas, aprendizaje pobre [8, 20, 21].
Sobremuestreo aleatorio (ROS)	Duplicar y/o alterar muestras minoritarias existentes hasta alcanzar el equilibrio.	Mismas deficiencias que submuestreo, más el riesgo elevado de sobreajuste [8, 20, 21].
RUS basado en clústers [52]	Sustituir muestras mayoritarias por tantos centroides de K-means como muestras haya en la clase minoritaria.	Aprendizaje pobre y difícil en entornos complejos como el aprendizaje de imágenes (diseñado para datos tabulares).
SMOTE [49]	Interpolar nuevas muestras minoritarias sintéticas a partir de las existentes.	Menos capacidad de generar información nueva y "variedad útil" que las GAN [21].
ADASYN [48]	Misma idea que SMOTE, dando más importancia a muestras difíciles al interpolar.	Menos capacidad de generar información nueva y "variedad útil" que las GAN [21].
DCGAN [21]	Familia de redes convolucionales generativas antagónicas (GANs) profundas.	Funciona bien, pero la mejora introducida en la WGAN ayuda aún más a estabilizar el aprendizaje [21].
WGAN [47]	DCGAN adaptada con la función de pérdida *Wasserstein-1* para mayor estabilidad y calidad de las muestras sintéticas.	

Tabla 3.1: Motivo de descarte de los métodos a nivel de datos no escogidos para los experimentos de este trabajo. El método escogido aparece subrayado.

Lin et al. [22]) y la red sensible a costes (CoSen CNN) de Khan et al. [23], que en los datasets probados obtiene una F1-score promedio superior al 80 % [8]. También son prometedores los resultados de la técnica PRM-IM [60], con más de un 93 % de F1-score en clasificación de radiografías. Otras técnicas tienen el inconveniente de estar diseñadas para problemas de aprendizaje sencillos (con datos tabulares y clasificadores simples), y por ello no son adecuadas para mitigar el sesgo en problemas más complejos, como lo es la clasificación de imágenes con CNNs. La Tabla 3.2 resume el motivo de descarte de las otras técnicas identificadas en la bibliografía.

Por último, para resolver la **RQ3** (relativa al impacto de la complejidad del problema en la decisión del método de mitigación), se propone realizar los experimentos sobre dos datasets de distinta complejidad, además de los distintos grados de desbalanceo ya propuestos. Para esto también es necesario contextualizar los resultados de los experimentos de acuerdo a algún tipo de métrica que cuantifique la "complejidad" de los datasets empleados. Es en ello esencial la revisión de Rahane and Subramanian [62], que propone las siguientes métricas estadísticas:

- **Distancia de Minkowski** [63]: es un método usado por defecto en las librerías de aprendizaje automático para calcular la distancia entre cada par de imágenes de un dataset. Es útil para dirimir el grado de similitud entre dos muestras (la distancia es próxima a cero cuanto más similares sean), y es preferible a la distancia euclídea [62].
- **Entropía**: es una medida de la incertidumbre o la aleatoriedad de un conjunto de datos. Si las imágenes en un dataset son muy parecidas, se espera que tengan una entropía baja, mientras que si son muy diferentes, se espera que tengan una entropía alta. En Rahane and Subramanian [62] se recomiendan:

Método	Descripción	Motivo de descarte
Reponderación ubicua [5]	Balancear la contribución entre clases completas y muestras individuales a la pérdida por entropía cruzada.	Arquitectura no disponible en dominio público.
Suavizado de etiquetas [55]	Suavizado de la distribución "*long-tailed*" dentro de la función de pérdida de entropía cruzada.	Solamente válido para un gran número de clases (en la bibliografía es habitual que sean 100 o más).
Vecinos difusos [57]	K-vecinos próximos con pesos difusos para mejorar la detección de clase minoritaria.	El algoritmo K-NN es menos apropiado que las CNN para el caso concreto de la clasificación de imágenes.
IEFSVM [56]	SVM modificada con un Sistema de Control Difuso (FCS) para mejorar la detección de clase minoritaria.	El algoritmo SVM es menos apropiado que las CNN para el caso concreto de la clasificación de imágenes.
Loss Focal [22]	Nueva función de pérdida que repondera muestras de la clase mayoritaria, reduciendo su impacto en la pérdida total.	
CoSen CNN [23]	Red con matriz de costes dinámica para aprender distintas penalizaciones en los errores de clasificación.	
PRM-IM [60]	Sistema de repetición de la clase minoritaria junto a subconjuntos de la mayoritaria, unido a un extractor de características formado por dos CNNs.	

Tabla 3.2: Motivo de descarte de los métodos a nivel de algoritmo no escogidos para los experimentos de este trabajo. Los métodos escogidos aparecen subrayados.

- **Entropía de Shannon**: también llamada entropía de la información, se puede calcular a partir de la distribución de probabilidad de los píxeles de las imágenes:

$$H = -\sum_{i=0}^{n-1} p_i log\ p_i \tag{3.1}$$

- **GLCM** (*Gray-Level Co-occurrence Matrix*): un histograma de co-ocurrencia de píxeles que ayuda a caracterizar la textura de las imágenes [62], al relacionar los valores de instensidad (0-255) con el vecindario para hallar relaciones espaciales. Su fórmula es:

$$H_g = -\sum_{i=0}^{n-1}\sum_{j=1}^{n-1} p(i,j) log\ p(i,j) \tag{3.2}$$

En el resto del capítulo se detallan a nivel teórico los métodos y algoritmos de mitigación de sesgo por desbalanceo de clases que han sido comparados en los experimentos, a fin de dar respuesta a las preguntas de investigación **RQ1** y **RQ2**. Las técnicas probadas pueden agruparse como estrategias a nivel de datos y algorítmicas. Dentro de esta última categoría, y como aportación novedosa de este trabajo, se ha diseñado un método de mitigación a nivel de algoritmo propio, llamado Loss Focal Difusa.

3.1. Método a nivel de datos

En la Sección anterior se ha justificado la elección exclusiva de las GAN por su rendimiento dominante dentro de esta familia de métodos mitigantes del desbalanceo.

Las redes generativas antagónicas (GAN) son modelos compuestos de dos redes neuronales, que en el dominio de las imágenes habitúan a ser CNNs. Estas redes reciben el nombre de “discriminador” y “generador” (puede verse en la Figura 3.1). El generador toma una entrada de “ruido” (*espacio latente aleatorio*), y genera una imagen sintética. El discriminador, dado un dataset, toma una imagen como entrada y produce una salida binaria que indica si esa imagen es real o falsa. Durante el entrenamiento, el discriminador se entrena para clasificar correctamente las imágenes como reales o falsas, mientras que el generador se entrena para producir imágenes que confundan al discriminador y se clasifiquen como reales. A medida que el generador mejora, el discriminador empeora, y en teoría el final del entrenamiento se produce cuando este último tiene una exactitud del 50 %. No obstante, este “estado de convergencia” de una GAN suele ser fugaz, en vez de estable [46]: si el discriminador elige al azar, el generador comienza a entrenar para recibir predicciones no deseadas, y su propia calidad disminuye.

En concreto, se ha escogido implementar la variante WGAN [47], ya que produce resultados sustancialmente mejores que el esquema clásico, e incluso que la DCGAN [21]. Esto es debido a varias novedades introducidas en Arjovsky et al. [47]. Primero, mientras que una GAN entrena al discriminador con la función de entropía binaria cruzada (BCE), la WGAN utiliza su propia función de pérdida

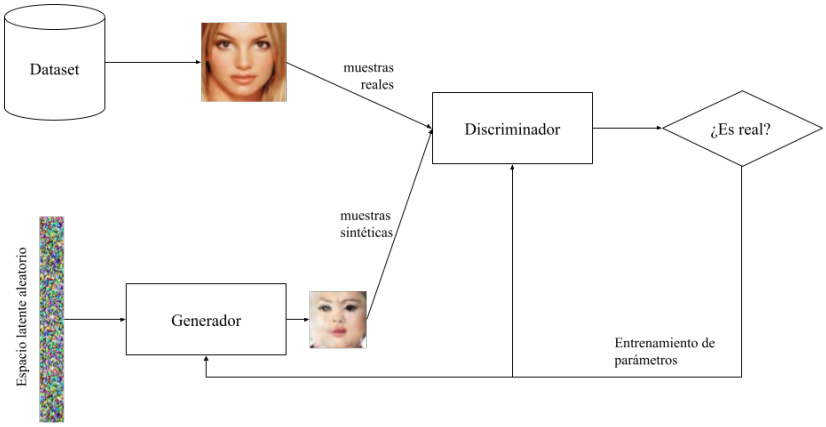

Figura 3.1: Esquema de funcionamiento de una red antagónica generativa (GAN).

(*Wasserstein-1*) para que el discrimador prediga una "probabilidad de realismo", en lugar de una etiqueta binaria. Esto transforma el rol del discriminador en el de un *crítico* que dirime el "nivel de realismo" de una entrada dada. Además, el crítico se actualiza más veces que el generador por cada iteración del entrenamiento (en el paper original, son 5 iteraciones de crítico por cada iteración global). La última diferencia importante es el uso del optimizador RMSProp frente al Adam usado por la DCGAN.

Las funciones de pérdida son realmente simples [46]: sea $D(x)$ el resultado del crítico para una instancia real x, $G(z)$ la salida del generador cuando recibe un vector z de ruido; y $D(G(z))$ el resultado del crítico para una muestra falsa. Se llama $LossD$ a la pérdida del crítico, y $LossG$ a la del generador:

$$\begin{aligned} LossD &= D(x) - D(G(z)) \\ LossG &= D(G(z)) \end{aligned} \tag{3.3}$$

Al maximizar su función, el crítico intenta maximizar la diferencia entre su resultado en instancias reales y su resultado en instancias falsas; y el generador intenta maximizar el resultado del crítico para sus instancias falsas. Gracias a todas estas novedades, las WGAN son capaces de sintetizar muestras con un mayor nivel de similitud a las reales, mejorando las métricas de clasificación en los dominios probados [47].

3.2. Métodos a nivel de algoritmo

A continuación, se describen los métodos de mitigación a nivel de algoritmo estudiados: la Loss Focal, el PRM-IM, la red sensible a costes (CoSenCNN), y la novedosa Loss Focal Difusa.

3.2.1. Loss Focal

La Entropía Cruzada (en adelante, *Cross Entropy* ó CE), es una función de pérdida que funciona bien en muchos escenarios, pero ve su efectividad reducida cuando las clases están desbalanceadas [18]. En Lin et al. [22] proponen la llamada *Loss Focal*, que centra su aprendizaje en las muestras minoritarias, como sustituto de la pérdida por entropía cruzada. Aplicándola al desbalanceo en problemas de detección de objetos, los autores encontraron una mejora sustancial en todas las métricas recogidas cuando se comparaban a otras del estado del arte [22]. Matemáticamente:

$$Focal\ Loss = -\sum_{n=1}^{m} T_n log\ t_n (1 - t_n)^{\gamma} \tag{3.4}$$

Donde m el número de clases, T_n es la clase real de la muestra y t_n es probabilidad o "confianza" para la clase predicha. El factor $(1 - t_n)^{\gamma}$ es el encargado de modular el valor de la función de pérdida: cuando t_n es pequeña, el factor es muy cercano a 1 y por tanto la pérdida no cambia. A medida que $t_n \to 1$, el factor se acerca a 0 y la pérdida para muestras bien clasificadas se degrada. Intuitivamente, este "factor modulador" reduce la contribución de las muestras "fáciles" (alta t_n), y extiende el rango en que una muestra recibe una pérdida baja. Cabe resaltar cómo, sin este factor (o si $\gamma = 0$), la Loss Focal no es más que la pérdida por Entropía Cruzada. La Figura 3.2 resume las características de esta función de pérdida.

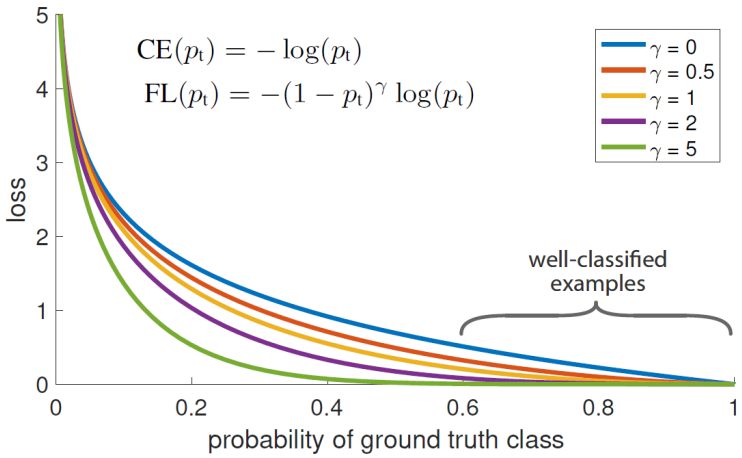

Figura 3.2: Curvas de pérdida para la Loss Focal bajo diversos valores de γ.

3.2.2. PRM-IM

En Rahimzadeh and Attar [10], se ataca el problema de desbalanceo en un dataset de radiografías con covid-19 por medio de una estructura que combina dos extractores de características basados en CNNs, y un algoritmo de repetición que permite al modelo revisar la clase minoritaria más veces que la mayoritaria. La técnica original surge de Ghanem et al. [64], donde se aplica a procesamiento

del lenguaje y recibe el nombre de PRM-IM (del inglés *Probabilistic Relational Model for IMbalanced class problem*). El modelo propuesto en Rahimzadeh and Attar [10] adapta la idea anterior al problema del desbalanceo en clasificación de radiografías de pulmones infectados con covid-19 y neumonía, mejorando las métricas para la clase más infrarrepresentada (covid-19) con una Accuracy del 90 %. La Figura 3.3 muestra el flujo de dicho sistema, ejemplificado para nuestro dominio de la clasificación de género.

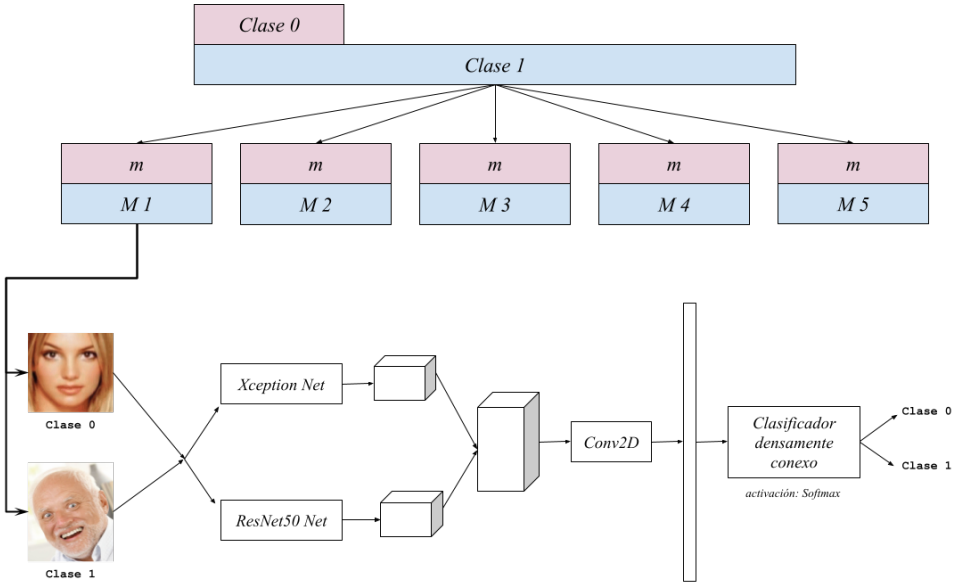

Figura 3.3: Esquema de flujo en el sistema PRM-IM de [10].

Los datos se dividen entre entrenamiento y validación, en un ratio 70-30 %. Para el subconjunto de entrenamiento, sean M y m los tamaños de la clase mayoritaria y minoritaria, respectivamente. A partir de ahí:

1.– Se dividen los datos de entrenamiento en $k = \frac{M}{m}$ subconjuntos. Cada uno estará formado por las m muestras de la clase minoritaria, y un subconjunto único de m muestras de la clase mayoritaria $M_i, i \in \{1, ..., k\}$.

2.– Durante E épocas de entrenamiento, iterar sobre los k subconjuntos:

2.1.– Para cada muestra de los k subconjuntos:

2.1.1.– Se introduce la muestra en dos extractores convolucionales en paralelo: uno se basa en *Xception* [65], y el otro en *ResNet50* [40], ambas sin la parte densamente conexa (sólo las capas convolucionales).

2.1.2.– Los vectores de características resultantes se concatenan y se hacen pasar por una última capa convolucional, aplanando su salida.

2.1.3.– Este vector aplanado se hace pasar por un clasificador formado por capas densamente conexas, para producir finalmente una

predicción para cada una de las clases del problema (previa activación por Softmax).

Así, el propósito de la arquitectura es, en esencia, que el modelo sea expuesto a la clase minoritaria más veces que a la mayoritaria, para compensar por el desbalanceo presente, y dependiendo el "número de exposiciones" en el grado de desbalanceo ($k = \frac{M}{m}$). Todo ello, sin alterar la distribución de clases del conjunto de datos (es decir, sin aplicar sobremuestreo en ningún momento).

Otras investigaciones han sugerido métodos de *ensemble* basados en esta idea, como Guo and Viktor [61]. No obstante, en este proyecto se implementa la arquitectura de la Figura 3.3 para las comparaciones por sus resultados prometedores (más de 93 % de F1-score en el campo de las radiografías de covid-19 y neumonía).

3.2.3. Red sensible a costes (CoSenCNN)

En Khan et al. [23], se reduce el sesgo hacia la clase mayoritaria utilizando el concepto de aprendizaje con costes. En este escenario, la función de costes penaliza al clasificador por las clasificaciones erróneas de distinta forma dependiendo de qué clase se está clasificando mal. En concreto, una matriz de costes define que la penalización por clasificar una muestra minoritaria como mayoritaria es mayor que si se clasifica una muestra mayoritaria como minoritaria.

La matriz de costes es una estructura $\mathcal{E}_{p,q}$ donde $\mathcal{E}_{p,p} = 0 \, \forall p$ y $\mathcal{E}_{p,q} > 0 \, \forall p, q$. Su función es tomar las salidas de la red convolucional que conforma el modelo y aplicar los costes antes de pasar por la función de pérdida (es decir, se aplica sobre los llamados *logits*). La función de pérdida en este caso es la de entropía cruzada. Los autores definen, pues, el siguiente modelo de funcionamiento del clasificador (Figura 3.4):

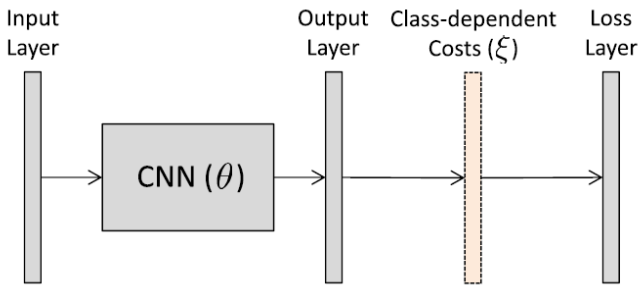

Figura 3.4: Esquema de funcionamiento en la red sensible a costes (CoSenCNN). Imagen de: Khan et al. [23].

Los valores de la matriz de coste, $\mathcal{E}_{p,q}, p \neq q$, se aprenden junto a los pesos de la red convolucional durante el proceso de entrenamiento.

3.2.4. Loss Focal Difusa

La motivación de querer diseñar un método algorítmico frente a uno a nivel de datos es que se ha demostrado que los primeros tienen menor probabilidad de afectar a los tiempos de entrenamiento. En esencia, suelen ser más fáciles de "portar" a diferentes problemas con un tuneado de parámetros mínimo. Es por ello que se considera que debe hacerse un esfuerzo por desarrollar métodos más eficientes dentro de esta familia.

La Loss Focal (Sección 3.2.1) tiene el inconveniente de que el parámetro γ mantiene un mismo valor para todas las clases, y durante todo el proceso de entrenamiento. En Hong et al. [18], se propone que a mayor tamaño de clase (mayor *grado de balanceo* respecto a dicha clase), mayor sea el valor de γ, pero los autores de este artículo siguen manteniendo el mismo valor durante todo el aprendizaje. Como el objetivo es minimizar la función de pérdida, la idea intuitiva es que a mayor γ, menor sea la Loss (Ecuación 3.4); y a menor γ, mayor Loss. Con esto se pretende que las muestras pertenecientes a la clase mayoritaria tengan una menor pérdida, facilitando su aprendizaje; mientras que las muestras de la clase minoritaria tendrán una mayor pérdida, y se les deberá prestar más atención para aprenderlas.

Por otra parte, durante el estudio bibliográfico se ha dado con varios estudios [7, 56, 57] en los que se hace uso de lógica difusa en algoritmos de aprendizaje automático. Vemos en Burduk [66] cómo se diseña una función de pérdida basada en reglas linguísticas difusas; en Cho et al. [56] cómo se modifica un SVM difuso (FSVM) para compensar el desbalanceo; o en Patel and Thakur [57] como hacen lo propio con un algoritmo de vecinos próximos. Es por ello que se plantea en este trabajo la idea de aplicar lógica difusa para mejorar las deficiencias detectadas en la Loss Focal tradicional, y dar lugar a un método llamado Loss Focal Difusa. Para dar un poco de contexto, un Sistema de Control Difuso (FCS) de Mamdani [67] tiene tres componentes, que ilustra la Figura 3.5:

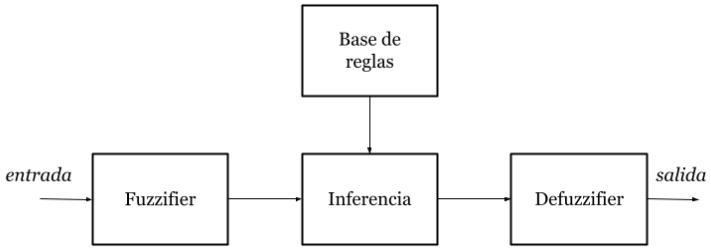

Figura 3.5: Esquema de flujo de información de un sistema de control difuso (FCS).

1.– El *fuzzifier* convierte la entrada a variables linguísticas de acuerdo a funciones de pertenencia ("*membership functions*") predefinidas.

2.– Una base de reglas linguísticas es usada para inferir conclusiones en base a las variables de entrada.

3.– El *defuzzifier* convierte las conclusiones a variables del mismo formato que la entrada.

Para definir la base de reglas y funciones de pertenencia del FCS de la Loss Focal Difusa, se ha pensado en la forma en que deben variar las pérdidas de ambas clases, mayoritaria y minoritaria, y en el papel del parámetro γ en todo ello. Se sabe que cuanto mayor es una clase, menor debe ser su pérdida (y viceversa), y también se quiere que ésta se modifique de forma activa durante el aprendizaje por medio de modificar el valor del parámetro γ para cada clase, para compensar las muestras difíciles de aprender. Entonces surge la duda de cuán grande debe ser esa variación de γ, y si esta debe ser positiva (aumentar) o negativa (disminuir). Así, de forma similiar a como Patel and Thakur [57] modifican el algoritmo de K-vecinos ponderado para modificar de forma difusa los pesos, aquí se utiliza un FCS para modificar el valor del parámetro γ para cada clase, al acabar cada iteración del aprendizaje. Dicho FCS tiene dos entradas. La primera es el grado de balanceo de cada clase (la inversa del IR de la Ecuación 2.1):

$$GB^i = \frac{tamaño\ clase_i}{Max(tamaño\ clase_i)} \leq 1, \quad i \in 0,1 \text{ para clasificación binaria.} \tag{3.5}$$

La segunda entrada es el valor de la Loss para cada clase en la iteración anterior, L^i_{t-1}, y su propósito es acoplar el aprendizaje pasado a la actualización de γ en la iteración presente (idea inspirada por Burduk [66]). Tanto GB^i como L^i_{t-1} pueden tomar los valores “alto”, “medio” y “bajo”. Por otra parte, la salida del FCS es la variación del parámetro γ, que puede ser “positivo” (aumenta), “negativo” (disminuye) o “cero”. Las funciones de pertenencia para las entradas y la salida se han diseñado como triangulares por ser rápidas de computar, y se ilustran en la Figura 3.6. El dominio de la función de pertenencia para la variable L^i_{t-1} se ha determinado empíricamente por la variación máxima de la pérdida en la Loss Focal tradicional. Por su parte, el dominio de GB^i solo puede variar, por definición, entre 0 y 1. Los límites de la salida se establecen entre -0,2 y 0,2 para evitar que la variación de γ sea muy abrupta de una iteración a otra.

Por último, y de acuerdo a las relaciones entre la actualización de γ y el tamaño de clases que ya se han razonado, se propone la siguiente base de reglas:

1.– **Regla 1**: SI GB^i ES “bajo” Y L^i_{t-1} ES “bajo” ENTONCES delta_gamma ES “*negativo*”
2.– **Regla 2**: SI GB^i ES “bajo” Y L^i_{t-1} ES “medio” ENTONCES delta_gamma ES “*negativo*”
3.– **Regla 3**: SI GB^i ES “bajo” Y L^i_{t-1} ES “alto” ENTONCES delta_gamma ES “*cero*”
4.– **Regla 4**: SI GB^i ES “alto” Y L^i_{t-1} ES “bajo” ENTONCES delta_gamma ES “*cero*”
5.– **Regla 5**: SI GB^i ES “alto” Y L^i_{t-1} ES “medio” ENTONCES delta_gamma ES “*positivo*”
6.– **Regla 6**: SI GB^i ES “alto” Y L^i_{t-1} ES “bajo” ENTONCES delta_gamma ES “*positivo*”

Con estas reglas, se logra una Loss elevada para la clase minoritaria (bajo GB^i), y una Loss reducida para la mayoritaria (alto GB^i), pero la actualización dinámica de γ pretende además “sortear” las muestras difíciles que aparezcan durante el entrenamiento. Finalmente, para convertir las salidas de las reglas a un valor de actualización de γ, se usa el método de centroides para “*deffuzificar*”, pues es el estándar en la bibliografía especializada [7, 56, 57, 68].

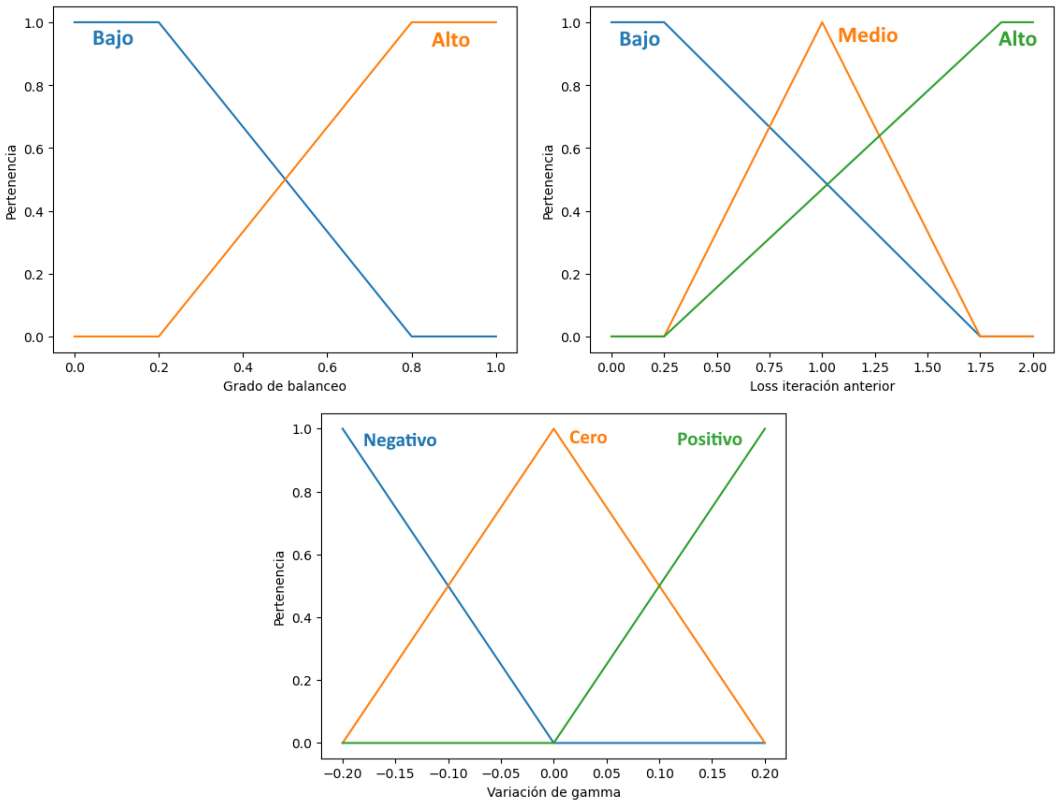

Figura 3.6: Funciones de pertenencia para las entradas (arriba) y la salida (abajo) del sistema de control de la Loss Focal Difusa.

4

Implementación y experimentos

En este capítulo, primero se justifica la elección de los datasets empleados durante los experimentos propuestos para responder a las preguntas de investigación. Después, se describe el diseño de dichos experimentos. Finalmente, se ofrecen algunos detalles técnicos de la implementación, así como las métricas recogidas durante la fase de ejecución.

4.1. Justificación de los datasets

Como ya se ha mencionado en la Sección 2.4, la pregunta **RQ2** (sobre la necesidad de conocer de antemano el desbalanceo del dataset) requiere un estudio sobre un dataset bajo diferentes grados de desbalanceo, de forma que se pretende observar las variaciones entre escenarios. Por otra parte, la resolución de la **RQ3** (sobre cómo afecta la complejidad del dataset al método a elegir) pasa por realizar los experimentos sobre dos conjuntos de datos con distinta complejidad. El primero, y más acorde al dominio de este trabajo, se compone de imágenes faciales de hombres y mujeres, sobre las que se han probado dos escenarios de desbalanceo distintos. Se trata del dataset "complejo", y sobre él se van a crear varios escenarios de desbalanceo. El segundo consiste en un conjunto de imágenes con una gran similitud entre muestras, por lo que constituye el dataset "sencillo".

En primer lugar, para el dominio del reconocimiento de género facial, de entre los tres datasets más mencionados en la bibliografía (FERET, UTKFace e IMDB-WIKI [1, 9, 11, 25, 27, 28, 29, 30, 36, 69]), dos presentan ciertos inconvenientes que motivan descartarlos:

1.– **FERET** [1] [1, 9, 25, 27, 36, 69]: no está disponible al dominio público.

2.– **IMDB-WIKI** [2] [11, 28, 30]: pese a contener información de género y raza para muchos miles de instancias, exige que se trate con MATLAB, lo cual dificulta su integración en el ecosistema de PyTorch utilizado.

Es por ello que se elige el dataset UTKFace [70, 71], que dispone de tres clases: género, edad y raza para 23.708 imágenes faciales a color, y en diversos escenarios de postura, resolución e ilumi-

[1] Fuente: https://www.nist.gov/itl/products-and-services/color-feret-database

[2] Fuente: https://data.vision.ee.ethz.ch/cvl/rrothe/imdb-wiki/

nación. Esta variedad asegura que el modelo pueda aprender las características diferenciadoras de hombres y mujeres en cuantos más escenarios de etnia, edad y condiciones posible. Su distribución de género, ilustrada en la Figura 4.1, es bastante igualitaria, con lo que permite generar artificialmente diferentes escenarios de desbalanceo.

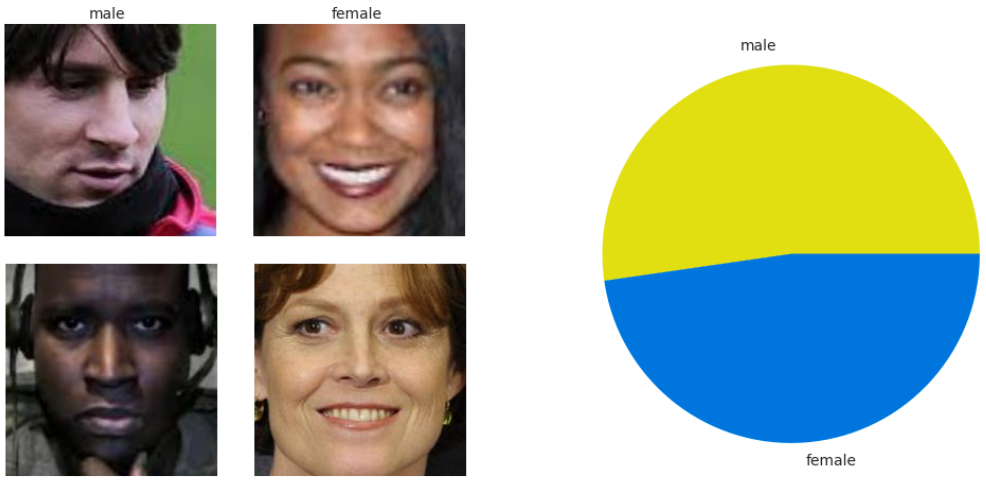

Figura 4.1: Estudio estadístico del dataset UTKFace.

En segundo lugar, se realizan los experimentos sobre un conjunto de datos con una gran similitud entre muestras, menos "complejo". En concreto, se ha escogido el dataset PlantVillage [72], por varios motivos. Primero, sus muestras recogen imágenes de plantas de tomate "sanas" e "infectadas", tomadas sobre el mismo fondo e iluminación. Esta alta similaridad entre muestras hace que PlantVillage sea más simple que UTKFace. El otro motivo de escoger este dataset es que se emplea en la revisión de Nafi and Hsu [21], en la que se demuestra empíricamente la idoneidad de las WGAN sobre otras técnicas a nivel de datos, pudiendo encontrar así un punto común sobre el que experimentar. Como este trabajo se centra en los problemas de clasificación binaria, se utiliza un subconjunto de este dataset. En concreto, solo se preservan las clases para tomates sanos, y para tomates infectados con el "*mosaic virus*", pues es el escenario de referencia descrito en Nafi and Hsu [21] (Figura 4.2). Así, el subconjunto de PlantVillage usado consta de 1964 imágenes: 1591 para tomates sanos, y 373 para los infectados (es decir, existe un desbalanceo por defecto).

En resumen, a partir de los dos datasets expuestos, se han creado los siguientes escenarios en los experimentos. Junto al alias de cada dataset se indica su IR (Ecuación 2.1):

1.– **UTKFaceBias** [IR = 4,23 (12391 hombres, 3056 mujeres)]: imágenes faciales de hombres y mujeres con un escenario de balanceo 4 a 1, en desventaja para los hombres (proveniente del dataset UTKFace [70, 71]).

2.– **UTKFaceFull** [IR = 1,095 (12391 hombres, 11317 mujeres)]: imágenes faciales de hombres y mujeres con clases balanceadas (mismo dataset que **UTKFaceBias**). Al incluir dos escenarios de balanceo para el mismo dataset, se pretende responder a la **RQ2**.

3.– **PlantVillage** [IR = 4,23 (1591 tomates sanos, 373 infectados)]: imágenes de plantas de

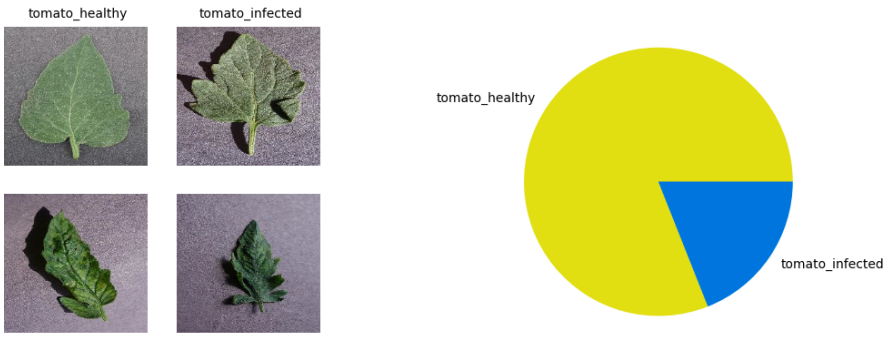

Figura 4.2: Estudio estadístico del dataset PlantVillage.

tomates sanos e infectados con un escenario de balanceo 4 a 1, en desventaja para los infectados (proveniente del dataset PlantVillage [72]). Esto nos permite experimentar sobre distintos escenarios de complejidad y dar respuesta a la **RQ3**.

Como puede observarse, el desbalanceo de **UTKFaceBias** y **PlantVillage** es el mismo: 4,23 (que es el IR "nativo" de PlantVillage). **UTKFaceBias** se ha adaptado a este IR para que las comparativas entre datasets con desbalanceo estén en la mayor igualdad de condiciones.

4.2. Diseño de los experimentos

En relación con la **RQ1** (sobre cuál de las dos familias de técnicas mitigantes del desbalanceo es más adecuada), para cada uno de los tres datasets enumerados, se entrenan los modelos del Capítulo 3. Poniendo en común los conjuntos de datos y el objetivo de cada experimento, se pretende que las métricas de rendimiento recogidas sean coherentes y significativas para obtener resultados concluyentes. Los modelos lanzados sobre cada uno de los tres datasets son los siguientes:

1.– **Baseline**: se entrena un modelo de clasificación estándar sobre los datos. Se utiliza como función de pérdida la de entropía cruzada (*Cross Entropy Loss*). Esta es la referencia contra la que se mide la mejora (o ausencia de mejora) de los métodos de mitigación del desbalanceo seleccionados, y de la Loss Focal Difusa.

2.– **WGAN**: consiste en el mismo modelo y configuración que **Baseline**, pero en este caso el desbalanceo de cada dataset se compensa por medio de la generación sintética de muestras con una WGAN entrenada previamente sobre la clase minoritaria. El código es el presentado en Arjovsky et al. [47].

3.– **Loss Focal**: al modelo de referencia del **Baseline** se le cambia la función de pérdida por una implementación oficial de la Loss Focal (código provisto en Lin et al. [22]), probando distintos valores de γ.

4.– **Loss Focal Difusa**: se trata de casi el mismo experimento que **Loss Focal**, pero incluyendo la actualización dinámica del hiperparámetro γ de la función de pérdida por medio del Sistema de Control Difuso diseñado en este proyecto.

5.– **PRM-IM**: aplica el sistema de Liu et al. [60] de repetición de la clase minoritaria junto a subconjuntos de la mayoritaria, unido a un extractor de características formado por dos CNNs (*Xception* [65], y *ResNet50* [40]) y una capa densamente conexa común. Se usa la entropía cruzada como función de pérdida.

6.– **CoSenCNN**: replica el escenario descrito en Khan et al. [23] para la red sensible a costes, utilizando como clasificador la misma arqutiectura de CNN que **Baseline**. La implementación es la propuesta en Khan et al. [23].

4.3. Detalles de implementación

Los experimentos y modelos han sido escritos en el lenguaje de programación Python, haciendo uso de las librerías de aprendizaje automático PyTorch [73], TensorFlow [63], Scikit-Learn [74] y Fuzzy-Logic [68]. Tanto para la **WGAN** [47], como para la **CoSenCNN** [23], se utilizan las implementaciones originales, adaptándolas al entorno de PyTorch. A nivel hardware, se dispone de una tarjeta gráfica NVIDIA GeForce GTX 1070, sistema operativo Windows 11 y CPU Intel Core i7-8700K de 3.70GHz .

En cuanto al clasificador de los experimentos **Baseline**, **WGAN**, **Loss Focal** y **Loss Focal Difusa**, la bibliografía estudiada motiva que sea una CNN [14, 28, 29, 30, 31, 32]. En este punto, hay dos opciones: bien diseñar la red desde cero, o bien aprovechar alguna arquitectura existente en el estado del arte. Se opta por la segunda opción, debido a que en los últimos años los modelos disponibles son bastante difíciles de superar, y el margen de aportación es limitado para la naturaleza de este trabajo. El modelo elegido es la arquitectura ResNet [40], concretamente ResNet-50, y al igual que otros estudios, se usa la técnica de *transfer learning* para adaptarlo a nuestras necesidades particulares. Se trata de un modelo entrenado en el dataset ImageNet, compuesto de cientos de miles de imágenes pertenecientes a 1000 categorías. En el *Reto de Reconocimiento Visual a Gran Escala* [75], ResNet es una de las arquitecturas con mejores resultados.

Exceptuando los experimentos de la **Loss Focal**, la **Loss Focal Difusa** y la **CoSen CNN**, el resto de clasificadores (**Baseline**, **WGAN** y **PRM-IM**) utilizan la función de pérdida por entropía cruzada, por ser la más usada en la bibliografía [14, 28, 29, 30, 31, 32]. Otros aspectos que cabe subrayar son el uso de planificadores de la tasa de aprendizaje (del inglés *learning rate schedulers*) y el optimizador elegido, cuando estos no vienen especificados por la implementación original. En ambos casos se han seguido las recomendaciones de la bibliografía [14, 28, 29, 30, 31, 32]: las redes utilizan un planificador "*One-Cycle*" con tasa máxima de 0.01; y un optimizador Adam, el cual suele ser preferido frente a otros como el de SGD por la mayoría de autores.

Para encontrar la configuración óptima de cada experimento, cada bucle de entrenamiento-validación se ha lanzado con diferentes combinaciones de hiperparámtros en lo que se conoce como búsqueda en malla (*grid search*) con validación cruzada de 3 iteraciones. Por último, cada uno de los bucles de entrenamiento-validación realiza una validación cruzada de 5 iteraciones (exceptuando **PRM-IM**, cuyo

algoritmo de entrenamiento es el de la Sección 3.2.2). La Tabla 4.1 resume los distintos hiperparámetros optimizados en el conjunto de los experimentos:

Hiperparámetro	Descripción
`epochs`	Número de iteraciones dentro de uno de los 5 bucles de entrenamiento-validación del *5-K-fold*.
`batch size`	Número de muestras dentro de cada "lote" en los que se divide el dataset durante el bucle de entrenamiento, para evitar cargar el conjunto de datos completo en memoria.
`learnign rate`	Define el "ritmo de avance" de la minimización de la función de pérdida.
`gamma`	Parámetro γ incluido en los métodos **Loss Focal** y **Loss Focal Difusa**.
`beta(Adam)`	Parámetro β del optimizador Adam [73] para el método **WGAN**.

Tabla 4.1: Hiperparámetros optimizados a lo largo de los distintos experimentos.

4.4. Métricas recogidas

A la hora de elegir las métricas que recoger en los experimentos, la bibliografía especializada coincide en las siguientes cinco: Accuracy (exactitud), Recall (exhaustividad), Precisión, F1-score y G-mean. Los autores que han estudiado el desbalanceo de clases [8, 20, 69] recomiendan medir la Precisión y la Recall, y muy encarecidamente la F1-score y la G-mean. La F1-score sirve como alternativa para expresar un concepto parecido al del Accuracy, pero es independiente del desbalanceo subyacente a los datos [8]. De igual forma, la G-mean es una buena métrica para problemas de desbalanceo de clases porque está diseñada para ser sensible tanto a la Precisión como al Recall en ambas clases, lo que significa que no solo se enfoca en la clase mayoritaria, sino también en la minoritaria [20]. Por tanto, de cada experimento realizado se mide su Precisión, Recall, F1-score y G-mean, para cada clase y en global (*macro average*). De acuerdo a la evaluación de la F1-score y la G-mean, se pretende obtener respuestas concluyentes con respecto a la adecuación de unos métodos u otros.

5

Análisis de resultados

En este capítulo se presentan los resultados y su análisis en relación con las respuestas a las preguntas de investigación planteadas.

5.1. Resultados experimentales

Las Tablas 5.1 a 5.3 muestran las métricas de los seis experimentos, para los conjuntos **UTKFace-Bias**, **UTKFaceFull** y **PlantVillage**. El Apéndice B ofrece información adicional relativa al consumo de recursos y configuración óptima de hiperparámetros para cada experimento.

Se denomina "Clase 0" a los hombres en el dataset UTKFace, y a los tomates sanos en PlantVillage. De igual forma, la "Clase 1" son las mujeres para UTKFace, y los tomates infectados para PlantVillage. En todos los escenarios de desbalanceo, la "Clase 1" es la minoritaria, y la 0 la mayoritaria. Es importante señalar que los experimentos **WGAN** y **PRM-IM** no han sido llevados a cabo para el conjunto de datos **UTKFaceFull**, que presenta un grado de desbalanceo de 1 (sus clases están equilibradas). Esta decisión se debe a que generar muestras minoritarias mediante WGAN, o dividir la clase mayoritaria en conjuntos del mismo tamaño que la clase minoritaria con PRM-IM, carece de sentido en un escenario donde los tamaños de las clases son idénticos, y sus métricas serían equivalentes al experimento **Baseline**.

	Macro average				Clase 0			Clase 1		
Experimento	**Recall**	**Precisión**	**F1-score**	**g-mean**	**Recall**	**Precisión**	**F1-score**	**Recall**	**Precisión**	**F1-score**
Baseline	0.9913043	0.9913043	0.9913043	0.9940954	0.9968944	0.9968944	0.9968944	0.9857140	0.9857140	0.9857140
WGAN	0.9985380	0.9983766	**0.9984549**	**0.9992687**	1.0000000	0.9967532	0.9983740	0.9970760	1.0000000	0.9985359
Loss Focal	0.9984326	0.9933333	0.9958594	0.9976486	0.9968652	1.0000000	0.9984301	1.0000000	0.9866667	0.9932886
Loss Focal Difusa	1.0000000	1.0000000	**1.0000000**	**1.0000000**	1.0000000	1.0000000	1.0000000	1.0000000	1.0000000	1.0000000
PRM-IM	0.9935642	0.9752451	0.9845126	0.9967769	1.0000000	0.9501254	0.9742123	0.9865145	1.0000000	0.9932542
CoSen CNN	0.9984520	0.9929577	0.9956787	0.9976777	0.9969040	1.0000000	0.9984496	1.0000000	0.9859155	0.9929078

Tabla 5.1: Resultados para el dataset **PlantVillage**. En negrita, se resaltan los métodos con las mejores métricas, tanto a nivel de datos como de algoritmo.

	Macro average				Clase 0			Clase 1		
Experimento	**Recall**	**Precisión**	**F1-score**	**g-mean**	**Recall**	**Precisión**	**F1-score**	**Recall**	**Precisión**	**F1-score**
Baseline	0.6736993	0.7705148	0.7036496	0.8027902	0.9566170	0.8777644	0.9154959	0.3907810	0.6632650	0.4918030
WGAN	0.8830004	0.8957965	**0.8845087**	**0.8434530**	0.8056769	0.9495625	0.8717222	0.9603239	0.8420305	0.8972953
Loss Focal	0.7457720	0.7184976	0.7212667	0.8043225	0.8674699	0.9054495	0.8860529	0.6240741	0.5315457	0.5741056
Loss Focal Difusa	0.7193296	0.7232835	**0.7300762**	**0.8239759**	0.8985828	0.8938326	0.8962014	0.5400763	0.5527344	0.5463320
PRM-IM	0.6534153	0.7432568	0.6253146	0.7888770	0.9524217	0.6132042	0.7465416	0.3540129	0.8725101	0.5032451
CoSen CNN	0.5936011	0.7620101	0.6133879	0.7617356	0.9774934	0.8447749	0.9063012	0.2097088	0.6792453	0.3204748

Tabla 5.2: Resultados para el dataset **UTKFaceBias**. En negrita, se resaltan los métodos con las mejores métricas, tanto a nivel de datos como de algoritmo.

	Macro average				Clase 0			Clase 1		
Experimento	**Recall**	**Precisión**	**F1-score**	**g-mean**	**Recall**	**Precisión**	**F1-score**	**Recall**	**Precisión**	**F1-score**
Baseline	0.7561197	0.7560004	0.7560406	0.7552294	0.7543403	0.74656357	0.7504318	0.7578991	0.7654372	0.7616495
WGAN	–	–	–	–	–	–	–	–	–	–
Loss Focal	0.7710066	0.7702876	0.7693955	0.7845349	0.7983005	0.7354759	0.7656015	0.7437126	0.8050994	0.7731895
Loss Focal Difusa	0.7878700	0.7871188	**0.7785564**	**0.7916486**	0.7954454	0.7545364	0.6877871	0.7548715	0.7045454	0.7221489
PRM-IM	–	–	–	–	–	–	–	–	–	–
CoSen CNN	0.7253209	0.7250534	0.7251617	0.7205537	0.7158177	0.7069727	0.7113677	0.7348243	0.7431341	0.7389559

Tabla 5.3: Resultados para el dataset **UTKFaceFull**. En negrita, se resaltan los métodos con las mejores métricas, tanto a nivel de datos como de algoritmo.

En general, los resultados de las tablas revelan la preponderancia en términos de *macro* F1-score y G-mean para los métodos **WGAN** y **Loss Focal Difusa**, como se resalta en negrita. Si se presta atención a las métricas desglosadas por clase, se puede ver cómo la presencia de desbalanceo afecta negativamente a la clase minoritaria, y positivamente a la minoritaria. Para encontrar métricas "equilibradas" entre las dos clases, hay que dirigirse bien al conjunto **UTKFaceFull** en la Tabla 5.3 (que tiene la misma proporción de hombres y mujeres); o bien a las métricas de **WGAN** de las Tablas 5.1 y 5.2, que representan los escenarios en los que el desbalanceo ha sido mitigado mediante sobremuestreo con imágenes sintéticas.

Las Figuras 5.1 y 5.2 permiten estudiar de manera detallada las métricas *macro* de F1-score y G-mean para los tres datasets. Estos diagramas de barras ayudan a discernir el impacto de las características de cada dataset por separado. En el caso de PlantVillage, compuesto por imágenes muy similares, se observa que las métricas de clasificación superan consistentemente el umbral del 90 %. En contraste, las métricas asociadas a UTKFace, que abarca una amplia variedad de caras humanas en diversos contextos de iluminación, raza y edad, muestran valores más bajos en general. Este contraste destaca la influencia que la similitud y la diversidad entre muestras en los conjuntos de datos pueden tener en el rendimiento de las métricas de clasificación.

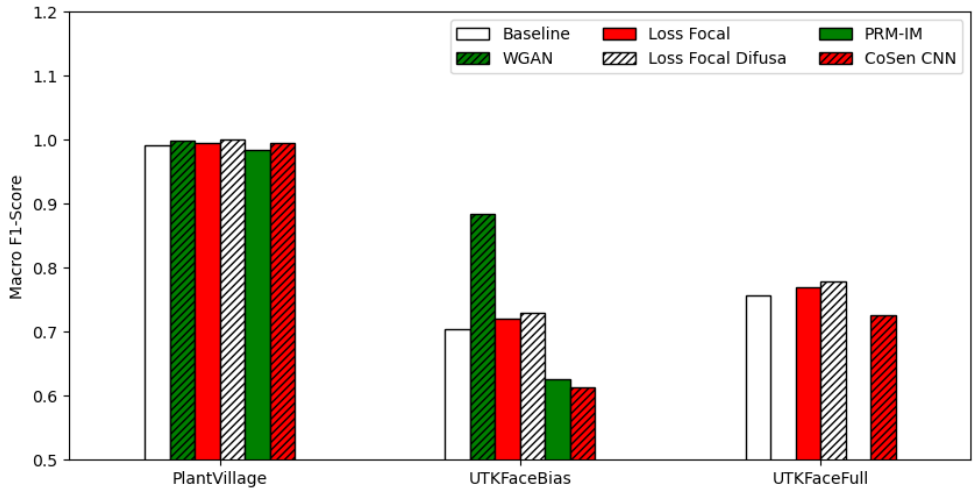

Figura 5.1: F1-score global para los métodos y datasets.

5.2. Respuesta a las preguntas de investigación

Tanto por medio de la F1-score global (Figura 5.1) como por la G-mean global (Figura 5.2), podemos llegar a las mismas conclusiones respecto al rendimiento comparativo de unos métodos y otros en función del escenario de datos y desbalanceo presentes. En este punto, se pueden recordar y responder las dos primeras preguntas de investigación:

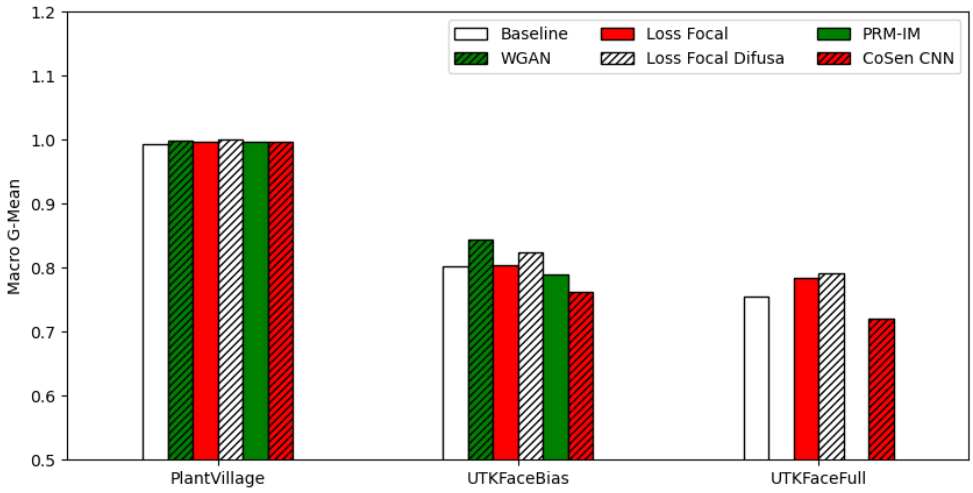

Figura 5.2: G-mean global para los métodos y datasets.

RQ1: Para mitigar el sesgo por desbalanceo, ¿es siempre mejor atacar los datos, los algoritmos, o depende de cada situación?

RQ2: ¿Es necesario conocer de antemano si el conjunto de datos presenta desbalanceo y, en su caso, el grado de desbalanceo, para determinar el método a aplicar?

La respuesta a la **RQ1** es que la elección del modelo depende claramente de escenario. En lo que respecta al dataset sencillo, PlantVillage, las métricas de F1-score y G-mean revelan que no hay tanta diferencia entre emplear WGAN frente a la Loss Focal Difusa o la CoSen CNN. Mientras tanto, si nos centramos en UTKFace, comparando los dos escenarios de desbalanceo, podemos ver que el claro ganador es el sobremuestreo por WGAN, seguido de la Loss Focal Difusa, que supera al resto de métodos de mitigación algorítmicos. Las Figuras 5.1 y 5.2 ilustran estas conclusiones. Por tanto, se puede decir que las características de cada escenario sí determinan el método mitigante a escoger. Si el tiempo de ejecución de los modelos quiere reducirse al mínimo, entonces se recomienda usar un método algorítmico como la Loss Focal Difusa, ya que es el que funciona mejor inmediatamente después de la WGAN para ambos datasets. Como ya se ha mencionado, los métodos algorítmicos no añaden tiempo de ejecución adicional, a diferencia de la mitigación a nivel de datos con WGAN (que requieren un entrenamiento y generación de muestras previo). En el Apéndice B se puede observar cómo la WGAN añade hasta 36 horas en tiempo de entrenamiento del generador al del clasificador, que está en el orden de la semana y media. En cambio, si el tiempo no supone un problema, el sobremuestreo por WGAN es el método más recomendable, ya que no sólo mejora las métricas globales, sino que también equilibra y eleva las métricas de las clases individuales.

En relación a la necesidad de conocer previamente el grado de desbalanceo en relación al método seleccionado (**RQ2**), se puede observar que en la mayoría de las estrategias algorítmicas (Loss Focal, Loss Focal Difusa y CoSenCNN), no es necesario, ya que los pesos y funciones de pérdida se ajustan de manera adecuada, independientemente del grado de desbalanceo. La única excepción es el método

PRM-IM, pues su estrategia de dividir la clase mayoritaria en conjuntos del mismo tamaño que la clase minoritaria no tiene sentido en un escenario donde los tamaños de las clases son idénticos. En cuanto al método de estrategia a nivel de datos (WGAN), sí resulta imprescindible tener conocimiento sobre el desbalanceo del conjunto de datos, dado que ello determina qué hacer (si existe desbalanceo, hay que saber cuántas muestras generar para alcanzar el equilibro de la distribución). Y aún en escenarios de equilibrio de clases, si involucran conjuntos de datos complejos, se puede considerar el realizar un aumento sintético de ambas clases. Aunque el propósito principal no sería mitigar el desbalanceo, se podrían mejorar las métricas de rendimiento al incrementar tanto la densidad de muestras como los atributos presentes.

Para terminar con el análsis de resultados, es necesario hacer comparativas entre los resultados obtenidos para el dataset de complejidad menor (PlantVillage) frente al más complejo (UTKFace), y determinar el método mejor en cada caso, dando respuesta a la tercera pregunta de investigación:

RQ3: ¿Cómo afecta la complejidad del problema a la toma de decisiones sobre el mejor método de mitigación del desbalanceo a aplicar?

Para ello, y de acuerdo a la propuesta de resolución descrita en el Capítulo 3, se han tomado mediciones estadísticas de distancias de Minkowski, entropía de Shannon y entropía de GLCM para ambos datasets. Las Figuras 5.3 a 5.5 muestran histogramas con la distribución de cada una de las tres métricas para cada dataset. Aparte, la Tabla 5.4 resume estadísticamente las medidas de entropía.

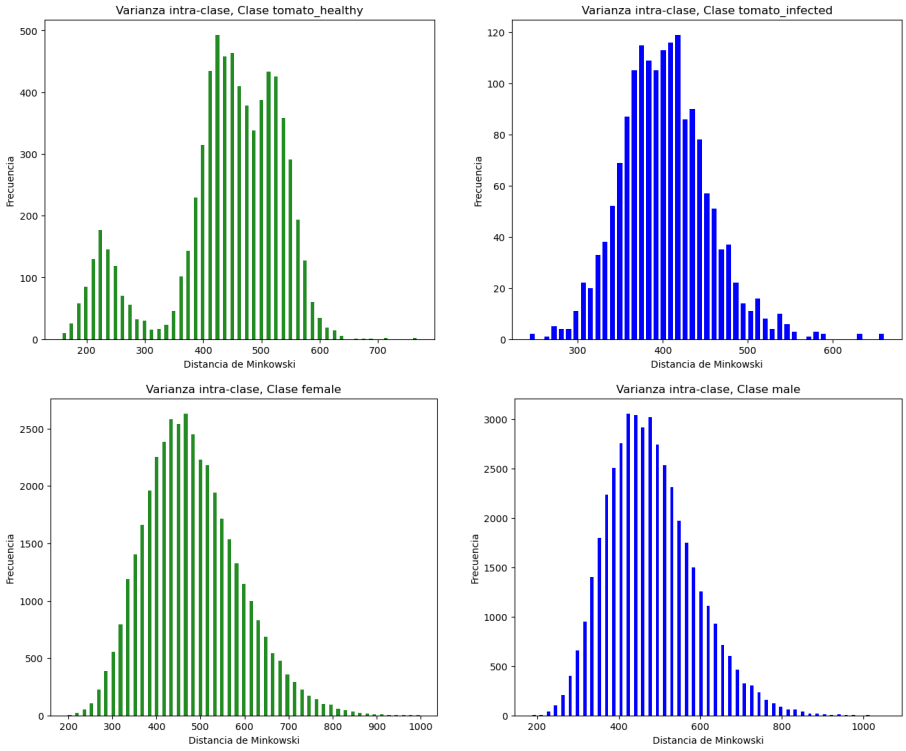

Figura 5.3: Distancias de Minkowski para el dataset PlantVillage (arriba) y UTKFace (abajo).

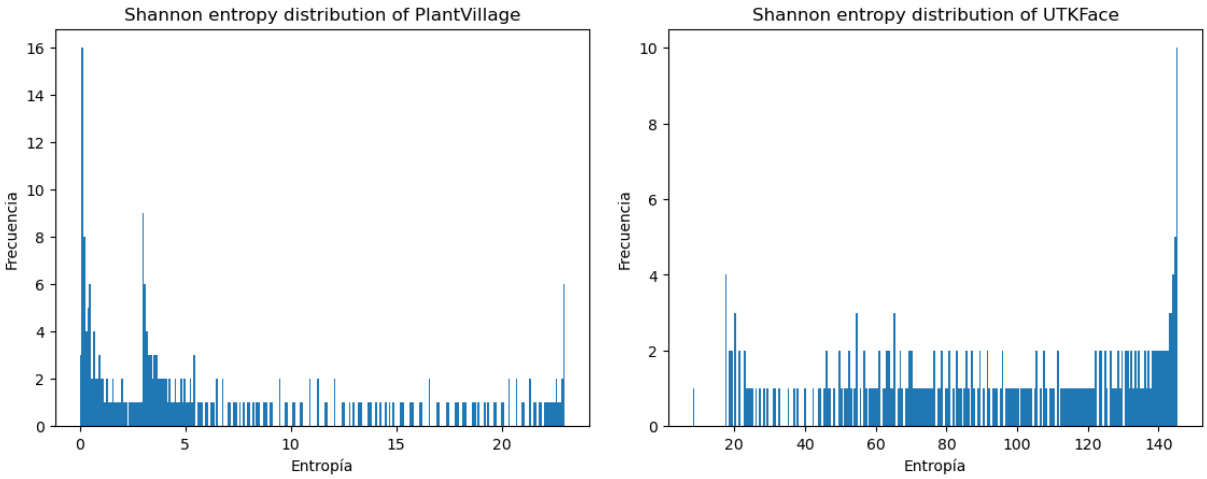

Figura 5.4: Entropías de Shannon para el dataset PlantVillage (izda.) y UTKFace (dcha.).

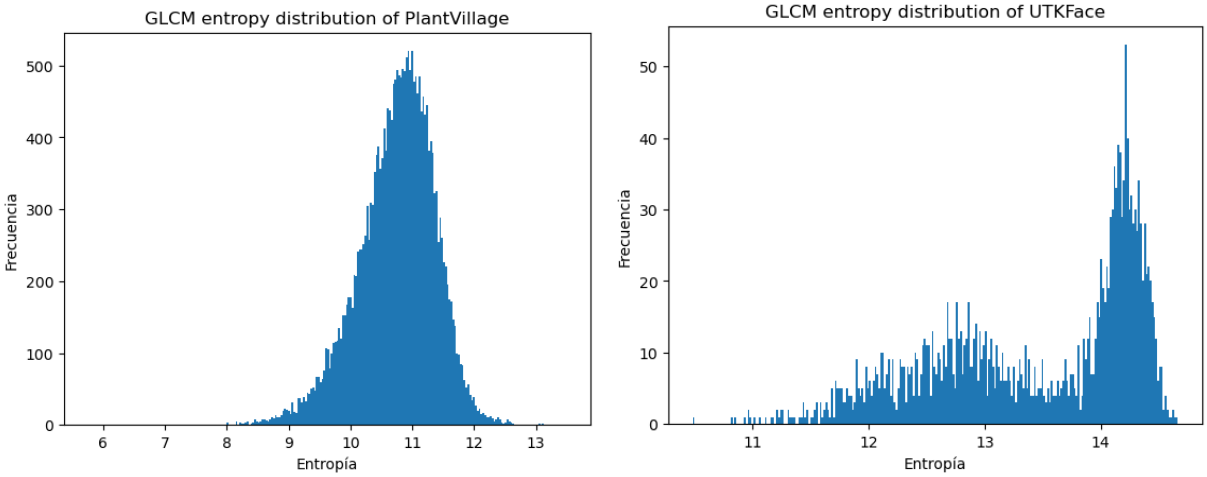

Figura 5.5: Entropías de GLCM para el dataset PlantVillage (izda.) y UTKFace (dcha.).

Dataset	Shannon (media)	GLCM (media/std)
PlantVillage	7.0906	13.4438 / 0.8817
UTKFace	7.2627	10.7452 / 0.6235

Tabla 5.4: Resumen estadístico de las métricas de entropía para ambos datasets.

Las gráficas de la Figura 5.3 son especialmente reveladoras a la hora de hablar de la distribución intra-clase de los conjunto de datos. Aunque para ambos la mayoría de muestras presentan una distancia relativa de entre 400 y 500, es notable cómo en UTKFace hay más *outliers*, para los que la distancia con respecto al resto aumenta a 900 e incluso a 1000. Por su parte, las distancias de Minkowski para PlantVillage siempre están acotadas en un rango reducido que no rebasa los 700.

En cuanto a las métricas de entropía, el histograma de Shannon de PlantVillage (Figura 5.4) muestra una concentración de bajas entropías, lo cual aventura una alta similaridad entre muestras. Esto es comprensible, ya que se trata de imágenes de plantas de tomate, muy similares entre sí. Por otro lado, el histograma de Shannon de UTKFace muestra una mayor concentración en las entropías altas, lo que indica una alta diversidad entre las instancias de imágenes (imágenes faciales en distintas configuraciones de género, raza, edad y condiciones ambientales). Estas conclusiones son extrapolables a la entropía por GLCM (Figura 5.5), que caracteriza mejor la variabilidad de texturas. El motivo de que en este caso ambos histogramas estén en una escala similar (con entropías entre 10 y 15) puede deberse a que, si bien las imágenes de PlantVillage son muy parecidas entre sí, presentan alta cantidad de grano fotográfico, lo que puede hacer que la caracterización de su textura sea más “rugosa”. En definitiva, estas métricas permiten cuantificar la noción de “complejidad” presente en un dataset de imágenes, en base a dos aspectos fundamentales:

1.– Distancia de Minkowski entre imágenes dentro de un rango acotado (más o menos por debajo de 700), lo que implica que las imágenes exhiben patrones de similitud más marcados en cuanto a las relaciones de proximidad.

2.– Una baja entropía, principalmente en términos de la información de Shannon, pero también en la textura caracterizada mediante GLCM. La baja entropía indica una mayor uniformidad y homogeneidad en las características visuales capturadas por las imágenes, y en forma de histograma se caracterizará por un “pico” en los valores próximos a cero.

Volviendo a la pregunta **RQ3**, si la complejidad del dataset es elevada, es mejor idea realizar un aumento de datos con WGAN, siempre y cuando exista una base lo suficientemente grande como para que el generador aprenda a producir muestras aceptables. Ello maximizará las métricas de clasificación. En cambio, para datasets de poca complejidad (como PlantVillage) basta con aplicar un método a algorítmico para paliar el desbalanceo, y seguir logrando métricas elevadas. En este caso, las tablas de resultados y las gráficas parecen indicar que el método propio, la Loss Focal Difusa, es un buen candidato. Por último, cabe recordar que pueden existir escenarios en los que la aplicación de métodos algorítmicos sea la única vía para mitigar el desbalanceo. Como se señala en Liu et al. [60], a veces la generación sintética de muestras puede ser un error ético, o incluso negligente. Por ejemplo, si la clase minoritaria la componen imágenes de radiografías de pulmones enfermos [10, 24], puede que no sea adecuado aplicar la generación sintética para mitigar el desbalanceo (aun cuando puedan crearse muestras altamente realistas), ya que en ese caso los modelos aprenderían las características de la enfermedad sobre muestras de pacientes que realmente no existen.

6

Conclusiones y trabajo futuro

En este capítulo se exponen las conclusiones finales, y se comentan posibles líneas de trabajo futuro que podrían seguirse para expandir el alcance de la presente investigación.

6.1. Conclusiones

El reconocimiento de género en imágenes es un campo del aprendizaje automático que tiene aplicaciones beneficiosas para la sociedad, como la mejora de la seguridad, la personalización de experiencias *online* y la identificación de desigualdades de género en ciertos ámbitos. Por ello, es crucial abordar las distintas fuentes de sesgo en los algoritmos utilizados en este campo. En esta investigación en concreto se ha enfocado en el sesgo por desbalanceo de clases.

La revisión bibliográfica [4, 5, 8, 44, 45] ha revelado que el desbalanceo de clases puede abordarse de dos formas: mediante el equilibrado de las distribuciones de los datasets, o mediante la adaptación de los algoritmos clasificadores para prestar más atención a las clases minoritarias. Se han identificado una serie de vacíos en la literatura (*research gaps*), ya que no se han encontrado comparativas concluyentes sobre cuál es el método algorítmico más efectivo, y tampoco hay comparativas que pongan en contraste ambas estrategias para determinar cuál es más apropiada, y en qué situaciones. Para abordar estos asuntos, se han formulado una serie de preguntas de investigación, y a fin de responderlas, se han elaborado experimentos que pusieran en comparación métodos de ambas categorías sobre dos datasets, en diferentes escenarios de desbalanceo.

Los resultados experimentales parecen indicar que la utilización de WGAN (un método de mitigación a nivel de datos) da la mayor mejora en el reconocimiento de la clase minoritaria, con una mejora de hasta un 21 % en F1-score (y un 2 % en G-mean) frente al mejor método algorítmico, la Loss Focal Difusa, cuando existe desbalanceo. No obstante, entrenar la red generativa implica añadir hasta 36 horas al tiempo de entrenamiento del clasificador. Además, existe el riesgo de generar "información falsa" que, según el dominio de aplicación, puede ser inaceptable. En estos escenarios ganan peso los métodos algorítmicos, que además no alteran la distribución de los datos, por lo que los tiempos de ejecución no aumentan.

Para terminar, se ha estudiado el papel que desempeña la complejidad del problema en la elección del método de mitigación del desbalanceo, por medio de comparar los experimentos sobre un dataset “sencillo” (PlantVillage) y otro más “complejo” (UTKFace). Dicha noción de complejidad se ha cuantificado mediante métricas como la distancia de Minkowski y las entropías de Shannon y GLCM. En resumen, bajo escenarios de alta complejidad, es recomendable emplear el aumento de datos mediante WGAN, siempre y cuando se disponga de una base lo suficientemente amplia para permitir que el generador aprenda a producir muestras aceptables. Por otro lado, en datasets de baja complejidad o donde la generación sintética no sea una alternativa viable, puede ser mejor idea aplicar métodos algorítmicos para abordar el desbalanceo (y si la complejidad es reducida, se seguirán obteniendo métricas elevadas). Aquí, de nuevo, se puede emplear la Loss Focal Difusa.

6.2. Trabajo futuro

Como primera línea de posible investigación futura, y dado que su rendimiento ha resultado exitoso dentro de los métodos mitigantes a nivel de datos, se puede explorar la optimización del FCS de la Loss Focal Difusa. En el campo de la optimización de sistemas difusos, Alcalá et al. [76] revisa diversos tipos de técnicas basadas en algoritmos genéticos [77] o redes neuronales [78]. También destaca el método PSO de Esmin et al. [79]. Una investigación futura podría explorar la capacidad de alguno de estos mecanismos para optimizar la base de reglas y/o las funciones de pertenencia del FCS propuesto.

Se ha hablado además de la complejidad de los datasets, y de cómo las métricas de género facial podrían mejorarse si se aumentara la densidad de muestras, por lo que podrían repetirse los experimentos para todos los métodos estudiados, con otro dataset de imágenes faciales más grande que UTKFace.

En cuanto a la WGAN, se podría evaluar de manera más detallada para determinar si existe un límite en el grado de desbalanceo, a partir del cual no puede aprender a generar muestras con un nivel de realismo suficiente. Comprender este límite es esencial para establecer expectativas realistas sobre las capacidades del sobremuestreo sintético generativo. Otro aspecto importante de esta técnica es que requiere un dataset mínimo para entrenar. Así, una posible idea de investigación futura sería examinar la capacidad de los modelos como *Stable Diffusion* [80] o *Midjourney* [81] para generar datasets completamente nuevos, y comparar su desempeño con las GAN. Este enfoque podría proporcionar una alternativa interesante en situaciones en las que apenas haya muestras de la clase a aumentar.

Por último, sería relevante explorar la aplicabilidad de los métodos más destacados en este trabajo en escenarios de desbalanceo múltiple o multiclase. Por ejemplo, considerando la etnia o la edad además del género en imágenes faciales. Estudiar la efectividad y las consideraciones específicas de estas técnicas en situaciones de desbalanceo más complejas ampliaría su aplicabilidad a una gama más amplia de problemas.

Bibliografía

[1] P Vallimeena, Uma Gopalakrishnan, Bhavana B Nair, and Sethuraman N Rao. Cnn algorithms for detection of human face attributes–a survey. In *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*, pages 576–581. IEEE, 2019.

[2] Sarah Vluymans. *Dealing with imbalanced and weakly labelled data in machine learning using fuzzy and rough set methods*. Springer, 2019.

[3] Andreas Kafkalias, Stylianos Herodotou, Zenonas Theodosiou, and Andreas Lanitis. Bias in face image classification machine learning models: The impact of annotator's gender and race. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pages 89–100. Springer, 2022.

[4] Osman Aka, Ken Burke, Alex Bauerle, Christina Greer, and Margaret Mitchell. Measuring model biases in the absence of ground truth. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 327–335, 2021.

[5] Jia Li, Yafei Song, Jianfeng Zhu, Lele Cheng, Ying Su, Lin Ye, Pengcheng Yuan, and Shumin Han. Learning from large-scale noisy web data with ubiquitous reweighting for image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(5):1808–1814, 2019.

[6] Gregory S Nelson. Bias in artificial intelligence. *North Carolina medical journal*, 80(4):220–222, 2019.

[7] Shigang Liu, Jun Zhang, Yang Xiang, and Wanlei Zhou. Fuzzy-based information decomposition for incomplete and imbalanced data learning. *IEEE Transactions on Fuzzy Systems*, 25(6):1476–1490, 2017.

[8] Justin M Johnson and Taghi M Khoshgoftaar. Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1):1–54, 2019.

[9] Liu Tianyu, Li Fei, and Wang Rui. Human face gender identification system based on mb-lbp. In *2018 Chinese Control And Decision Conference (CCDC)*, pages 1721–1725. IEEE, 2018.

[10] Mohammad Rahimzadeh and Abolfazl Attar. A modified deep convolutional neural network for detecting covid-19 and pneumonia from chest x-ray images based on the concatenation of xception and resnet50v2. *Informatics in medicine unlocked*, 19:100360, 2020.

[11] Olatunbosun Agbo-Ajala and Serestina Viriri. Face-based age and gender classification using deep learning model. In *Pacific-Rim Symposium on Image and Video Technology*, pages 125–137. Springer, 2019.

[12] Byungok Han, Woo-Han Yun, Jang-Hee Yoo, and Won Hwa Kim. Toward unbiased facial expression recognition in the wild via cross-dataset adaptation. *IEEE Access*, 8:159172–159181, 2020.

[13] S Poornima, N Sripriya, S Preethi, and Saanjana Harish. Classification of gender from face images and voice. In *Intelligence in Big Data Technologies—Beyond the Hype*, pages 115–124. Springer, 2021.

[14] Wenying Wu, Pavlos Protopapas, Zheng Yang, and Panagiotis Michalatos. Gender classification and bias mitigation in facial images. In *12th acm conference on web science*, pages 106–114, 2020.

[15] Christina Wadsworth, Francesca Vera, and Chris Piech. Achieving fairness through adversarial learning: an application to recidivism prediction. *arXiv preprint arXiv:1807.00199*, 2018.

[16] Matthew J Vowels, Necati Cihan Camgoz, and Richard Bowden. Nestedvae: Isolating common factors via weak supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9202–9212, 2020.

[17] Yan Yan, Ying Huang, Si Chen, Chunhua Shen, and Hanzi Wang. Joint deep learning of facial expression synthesis and recognition. *IEEE Transactions on Multimedia*, 22(11):2792–2807, 2019.

[18] Tzung-Pei Hong, Wei-Chun Peng, Ja-Hwung Su, and Shyue-Liang Wang. Fuzzy adaptive focal loss for imbalanced datasets. In *2021 IEEE International Conference on Imaging Systems and Techniques (IST)*, pages 1–5. IEEE, 2021.

[19] Ali Tunç, Sakir Taşdemir, and Murat Köklü. Fuzzy based noise removal, age group and gender prediction with cnn. In *International Conference on Intelligent and Fuzzy Systems*, pages 204–212. Springer, 2020.

[20] Kamlesh Upadhyay, Prabhjot Kaur, SVAV Prasad, and Lingayas Vidyapeeth. State of the art on data level methods to address class imbalance problem in binary classification. *GIS Science Journal*, 2021.

[21] Nasik Muhammad Nafi and William H Hsu. Addressing class imbalance in image-based plant disease detection: Deep generative vs. sampling-based approaches. In *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)*, pages 243–248. IEEE, 2020.

[22] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[23] Salman H Khan, Munawar Hayat, Mohammed Bennamoun, Ferdous A Sohel, and Roberto Togneri. Cost-sensitive learning of deep feature representations from imbalanced data. *IEEE transactions on neural networks and learning systems*, 29(8):3573–3587, 2017.

[24] Barath Narayanan Narayanan, Russell C. Hardie, Vignesh Krishnaraja, Christina Karam, and Venkata Salini Priyamvada Davuluru. Transfer-to-transfer learning approach for computer aided detection of covid-19 in chest radiographs. *AI*, 1(4):539–557, 2020. ISSN 2673-2688. URL https://www.mdpi.com/2673-2688/1/4/32.

[25] EK Loo, TS Lim, LY Ong, and CH Lim. The influence of ethnicity in facial gender estimation. In *2018 IEEE 14th International Colloquium on Signal Processing & Its Applications (CSPA)*, pages 187–192. IEEE, 2018.

[26] Yongjing Lin and Huosheng Xie. Face gender recognition based on face recognition feature vectors. In *2020 IEEE 3rd International Conference on Information Systems and Computer Aided Education (ICISCAE)*, pages 162–166. IEEE, 2020.

[27] Vladimir Khryashchev, Lev Shmaglit, Andrey Priorov, and Andrey Shemyakov. Adaptive feature extraction for gender classification of human faces. In *Grafikon (Russia)'2013*, pages 71–74, 2013.

[28] Md Mahbubul Islam, Nusrat Tasnim, and Joong-Hwan Baek. Human gender classification using transfer learning via pareto frontier cnn networks. *Inventions*, 5(2):16, 2020.

[29] Avishek Garain, Biswarup Ray, Pawan Kumar Singh, Ali Ahmadian, Norazak Senu, and Ram Sarkar. Gra_net: A deep learning model for classification of age and gender from facial images. *IEEE Access*, 9:85672–85689, 2021.

[30] Olatunbosun Agbo-Ajala and Serestina Viriri. Deeply learned classifiers for age and gender predictions of unfiltered faces. *The Scientific World Journal*, 2020, 2020.

[31] Tahmina Akter Sumi, Mohammad Shahadat Hossain, Raihan Ul Islam, and Karl Andersson. Human gender detection from facial images using convolution neural network. In *International Conference on Applied Intelligence and Informatics*, pages 188–203. Springer, 2021.

[32] Xiaoxiong Zhang, Sajid Javed, Ahmad Obeid, Jorge Dias, and Naoufel Werghi. Gender recognition on rgb-d image. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 1836–1840. IEEE, 2020.

[33] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. Ieee, 2005.

[34] Jayaprada S Hiremath, Shantakumar B Patil, and Premjyoti S Patil. Human age and gender prediction using machine learning algorithm. In *2021 IEEE International Conference on Mobile Networks and Wireless Communications (ICMNWC)*, pages 1–5. IEEE, 2021.

[35] Vivek Kumar Verma, Sumit Srivastava, Tarun Jain, and Ashish Jain. Local invariant feature-based gender recognition from facial images. In *Soft computing for problem solving*, pages 869–878. Springer, 2019.

[36] Selim Yilmaz and CEMİL ZALLUHOĞLU. An evolutionary-based image classification approach through facial attributes. *Turkish Journal of Electrical Engineering and Computer Sciences*, 29(2): 860–874, 2021.

[37] Tawsin Uddin Ahmed, Sazzad Hossain, Mohammad Shahadat Hossain, Raihan ul Islam, and Karl Andersson. Facial expression recognition using convolutional neural network with data augmentation. In *2019 Joint 8th International Conference on Informatics, Electronics & Vision (ICIEV) and 2019 3rd International Conference on Imaging, Vision & Pattern Recognition (icIVPR)*, pages 336–341. IEEE, 2019.

[38] Noortaz Rezaoana, Mohammad Shahadat Hossain, and Karl Andersson. Detection and classification of skin cancer by using a parallel cnn model. In *2020 IEEE International Women in Engineering (WIE) Conference on Electrical and Computer Engineering (WIECON-ECE)*, pages 380–386. IEEE, 2020.

[39] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[40] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[41] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[42] Ayesha Gurnani, Kenil Shah, Vandit Gajjar, Viraj Mavani, and Yash Khandhediya. Saf-bage: Salient approach for facial soft-biometric classification-age, gender, and facial expression. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 839–847. IEEE, 2019.

[43] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

[44] Michael P Kim, Amirata Ghorbani, and James Zou. Multiaccuracy: Black-box post-processing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 247–254, 2019.

[45] Anoop Krishnan, Ali Almadan, and Ajita Rattani. Understanding fairness of gender classification algorithms across gender-race groups. In *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 1028–1035. IEEE, 2020.

[46] Machine learning and generative adversarial networks course. URL `https://developers.google.com/machine-learning/gan/training?hl=es-419`. Consultado: 12-04-2023.

[47] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

[48] Haibo He, Yang Bai, Edwardo A Garcia, and Shutao Li. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*, pages 1322–1328. IEEE, 2008.

[49] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[50] Getinet Yilma, Kumie Gedamu, Maregu Assefa, Ariyo Oluwasanmi, and Zhiguang Qin. Generation and transformation invariant learning for tomato disease classification. In *2021 IEEE 2nd International Conference on Pattern Recognition and Machine Learning (PRML)*, pages 121–128. IEEE, 2021.

[51] Luo Jiang, Juyong Zhang, and Bailin Deng. Robust rgb-d face recognition using attribute-aware loss. *IEEE transactions on pattern analysis and machine intelligence*, 42(10):2552–2566, 2019.

[52] Wei-Chao Lin, Chih-Fong Tsai, Ya-Han Hu, and Jing-Shang Jhang. Clustering-based undersampling in class-imbalanced data. *Information Sciences*, 409:17–26, 2017.

[53] Ilias Siniosoglou, Vasileios Argyriou, Stamatia Bibi, Thomas Lagkas, and Panagiotis Sarigiannidis. Unsupervised ethical equity evaluation of adversarial federated networks. In *The 16th International Conference on Availability, Reliability and Security*, pages 1–6, 2021.

[54] Yufei Zhao, Jinxin Yang, Jiangtao Du, Zhen Chen, and Wen-Chi Yang. A lightweight classifier for facial expression recognition based on evolutionary svm ensembles. In *2021 6th International Conference for Convergence in Technology (I2CT)*, pages 1–9. IEEE, 2021.

[55] Zhisheng Zhong, Jiequan Cui, Shu Liu, and Jiaya Jia. Improving calibration for long-tailed recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16489–16498, 2021.

[56] Poongjin Cho, Minhyuk Lee, and Woojin Chang. Instance-based entropy fuzzy support vector machine for imbalanced data. *Pattern Analysis and Applications*, 23(3):1183–1202, 2020.

[57] Harshita Patel and Ghanshyam Singh Thakur. Classification of imbalanced data using a modified fuzzy-neighbor weighted approach. *International Journal of Intelligent Engineering and Systems*, 10(1):56–64, 2017.

[58] Alberto Fernández, María Calderón, Edurne Barrenechea, Humberto Bustince, and Francisco Herrera. Solving multi-class problems with linguistic fuzzy rule based classification systems based on pairwise learning and preference relations. *Fuzzy sets and systems*, 161(23):3064–3080, 2010.

[59] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[60] Lijue Liu, Xiaoyu Wu, Shihao Li, Yi Li, Shiyang Tan, and Yongping Bai. Solving the class imbalance problem using ensemble algorithm: application of screening for aortic dissection. *BMC Medical Informatics and Decision Making*, 22(1):1–16, 2022.

[61] Hongyu Guo and Herna L Viktor. Learning from imbalanced data sets with boosting and data generation: the databoost-im approach. *ACM Sigkdd Explorations Newsletter*, 6(1):30–39, 2004.

[62] Ameet Annasaheb Rahane and Anbumani Subramanian. Measures of complexity for large scale image datasets. In *2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, pages 282–287. IEEE, 2020.

[63] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL `https://www.tensorflow.org/`. Software available from tensorflow.org.

[64] Amal S Ghanem, Svetha Venkatesh, and Geoff West. Learning in imbalanced relational data. In *2008 19th International Conference on Pattern Recognition*, pages 1–4. IEEE, 2008.

[65] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017.

[66] Robert Burduk. *Possibility of Use a Fuzzy Loss Function in Medical Diagnostics*, volume 47, pages 476–481. 09 2008. ISBN 978-3-540-68167-0. doi: 10.1007/978-3-540-68168-7_53.

[67] Ebrahim H Mamdani. Application of fuzzy algorithms for control of simple dynamic plant. In *Proceedings of the institution of electrical engineers*, volume 121, pages 1585–1588. IET, 1974.

[68] Anselm Kiefner. FuzzyLogic for Python, February 2022. URL `https://github.com/amogorkon/fuzzylogic`.

[69] Neelam Dwivedi and Dushyant Kumar Singh. Review of deep learning techniques for gender classification in images. In *Harmony Search and Nature Inspired Optimization Algorithms*, pages 1089–1099. Springer, 2019.

[70] Utkface - large scale face dataset. https://susanqq.github.io/UTKFace/. Consultado: 30-11-2022.

[71] Song Yang Zhang, Zhifei and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017.

[72] David Hughes, Marcel Salathé, et al. An open access repository of images on plant health to enable the development of mobile disease diagnostics. *arXiv preprint arXiv:1511.08060*, 2015.

[73] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[74] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[75] Imagenet large scale visual recognition challenge. https://image-net.org/challenges/LSVRC/. Consultado: 17-11-2022.

[76] R Alcalá, J Casillas, O Cordón, F Herrera, and JS Zwir. Techniques for learning and tuning fuzzy rule-based systems for linguistic modeling and their application knowledge engineering systems, techniques and applications. *Systems, Techniques and Applications*, 1999.

[77] Yunjeong Kang, Malrey Lee, Yongseok Lee, and Thomas M Gatton. Optimization of fuzzy rules: integrated approach for classification problems. In *Computational Science and Its Applications-ICCSA 2006: International Conference, Glasgow, UK, May 8-11, 2006, Proceedings, Part V 6*, pages 665–674. Springer, 2006.

[78] J-SR Jang. Anfis: adaptive-network-based fuzzy inference system. *IEEE transactions on systems, man, and cybernetics*, 23(3):665–685, 1993.

[79] AAA Esmin, AR Aoki, and G Lambert-Torres. Particle swarm optimization for fuzzy membership functions optimization. In *IEEE International Conference on Systems, Man and Cybernetics*, volume 3, pages 6–pp. IEEE, 2002.

[80] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[81] Jonas Oppenlaender. The creativity of text-to-image generation. In *Proceedings of the 25th International Academic Mindtrek Conference*, pages 192–202, 2022.

Apéndices

A

Consultas bibliográficas

Se adjuntan las consultas usadas para extraer bibliografía de las bases de datos académicas. Para realizar el estudio bibliográfico relacionado con este trabajo, se han buscado artículos por medio de los buscadores avanzados que ofrencen las bases de datos Web of Science (WoS), el IEEE y Scopus. Primero se han realizado consultas para investigar sobre el problema de la clasificación de género en imágenes; y después se ha explorado el problema del sesgo presente (y muy en especial el causado por desbalanceo de clases).

(“people” OR “person” OR “human”) AND
(“gender”) AND
(“image recognition” OR “image classification”) AND
(“data science” “OR deep learning” OR “machine learning OR “ML”) AND
(“algorithm” OR “technique”)

Tabla A.1: Consulta para papers de reconocimiento de género facial con el motor del IEEE.

```
(
TI="people" OR TS="people" OR AB="people" OR AK="people" OR
TI="person" OR TS="person" OR AB="person" OR AK="person" OR
TI="human" OR TS="human" OR AB="human" OR AK="human"
) AND (
TI="gender" OR TS="gender OR AB="gender" OR AK="gender
) AND (
TI="image recognition" OR TS="image recognition"
OR AK="image recognition" OR
TI="image classification" OR TS="image classification"
OR AK="image classification"
) AND (
TI="data science" OR TS="data science" OR AB="data science"
OR AK="data science" OR
TI="deep learning" OR TS="deep learning" OR AB="deep learning"
OR AK="deep learning" OR
TI="machine learning" OR TS="machine learning"
OR AB="machine learning" OR AK="machine leaning" OR
TI="ML" OR TS="ML" OR AB="ML" OR AK="ML"
) AND (
TI="algorithm" OR TS="algorithm" OR AB="algorithm"
OR AK="algorithm" OR
TI="technique" OR TS="technique" OR AB="technique"
OR AK="technique"
)
```

Tabla A.2: Consulta para papers de reconocimiento de género facial con el motor de Web of Science.

```
(
TITLE-ABS-KEY("people") OR
TITLE-ABS-KEY("person") OR
TITLE-ABS-KEY("human")
) AND (
TITLE-ABS-KEY("gender")
) AND (
TITLE-ABS-KEY("image recognition") OR
TITLE-ABS-KEY("image classfication")
) AND (
TITLE-ABS-KEY("data science") OR
TITLE-ABS-KEY("deep learning") OR
TITLE-ABS-KEY("machine leaning") OR
TITLE-ABS-KEY("ML")
) AND (
TITLE-ABS-KEY("algorithm") OR
TITLE-ABS-KEY("technique")
)
```

Tabla A.3: Consulta para papers de reconocimiento de género facial con el motor de Scopus.

(“bias”) AND (“image recognition” OR “image classification”) AND (“data science” OR “deep learning” OR “machine learning” OR “ML”) AND (“data” OR “dataset”)

Tabla A.4: Consulta para papers de sesgos en la clasificación de imágenes con el motor del IEEE.

(fuzzy OR class imabalance OR imbalance OR imbalanced) AND (bias) AND (data science OR deep learning OR machine learning OR ML) AND (data OR dataset) AND (imbalanced)

Tabla A.5: Consulta secundaria para papers de sesgos en la clasificación de imágenes con el motor del IEEE.

(TI=“bias” OR TS=“bias” OR AB=“bias” OR AK=“bias”) AND (TI=“image recognition” OR TS=“image recognition” OR AB=“image recognition” OR AK=“image recognition” OR TI=“image classification” OR TS=“image classification” OR AB=“image classification” OR AK=“image classification”) AND (TI=“data science” OR TS=“data science” OR AB=“data science” OR AK=“data science” OR TI=“deep learning” OR TS=“deep learning” OR AB=“deep learning” OR AK=“deep learning” OR TI=“machine learning” OR TS=“machine learning” OR AB=“machine learning” OR AK=“machine learning” OR TI=“ML” OR TS=“ML” OR AB=“ML” OR AK=“ML”) AND (TI=“data” OR TS=“data” OR AB=“data” OR AK=“data” OR TI=“dataset” OR TS=“dataset” OR AB=“dataset” OR AK=“dataset”)

Tabla A.6: Consulta para papers de sesgos en la clasificación de imágenes con el motor de Web of Science.

```
(
TITLE-ABS-KEY(“bias”)
)
AND
(
TITLE-ABS-KEY(“image recognition”) OR
TITLE-ABS-KEY(“image classification”)
)
AND
(
TITLE-ABS-KEY(“data science”) OR
TITLE-ABS-KEY(“deep learning”) OR
TITLE-ABS-KEY(“machine learning”) OR
TITLE-ABS-KEY(“ML”)
)
AND
(
TITLE-ABS-KEY(“data”) OR
TITLE-ABS-KEY(“dataset”)
)
```

Tabla A.7: Consulta para papers de sesgos en la clasificación de imágenes con el motor de Scopus.

```
(
TITLE-ABS-KEY ( “gender” )
)
AND
(
TITLE-ABS-KEY ( “bias” ) OR
TITLE-ABS-KEY ( “imbalance” ) OR
TITLE-ABS-KEY ( “imbalanced” ) OR
TITLE-ABS-KEY ( “class imbalance” ) OR
)
AND
(
TITLE-ABS-KEY ( “image recognition” ) OR
TITLE-ABS-KEY ( “image classification” )
)
AND
(
TITLE-ABS-KEY ( “data science” ) OR
TITLE-ABS-KEY ( “deep learning” ) OR
TITLE-ABS-KEY ( “machine learning” ) OR
TITLE-ABS-KEY ( “ML” )
)
AND
(
TITLE-ABS-KEY ( “data” ) OR
TITLE-ABS-KEY ( “dataset” )
)
```

Tabla A.8: Consulta secundaria para papers de sesgos en la clasificación de imágenes con el motor de Scopus.

B

Información adicional de los experimentos

Se ofrece la configuración de hiperparámetros óptima para cada experimento de los descritos en el Capítulo 4. Asimismo, se incluyen datos de tiempo de ejecución y consumo de memoria de GPU para cada uno, a fin de dar una visión más completa sobre el desarrollo de esta parte del trabajo. Para el experimento **WGAN**, se aportan dos mediciones separadas por una barra (/): la primera corresponde al entrenamiento de la WGAN para el aumento de datos, y la segunda, al entrenamiento del clasificador sobre la distribución aumentada.

	Recursos		Hiperparámetros óptimos		
Dataset	**GPU RAM (%)**	**Tiempo de ejecución**	**Épocas**	**batch_size**	**learning rate**
PlantVillage	60 %	36h	10	15	1.00E-03
UTKFaceBias	80 %	1sem	25	10	1.00E-03
UTKFaceFull	80 %	1sem	25	10	1.00E-04

Tabla B.1: Rendimiento e hiperparámetros óptimos para el experimento **Baseline**.

	Recursos		Hiperparámetros óptimos			
Dataset	**GPU RAM (%)**	**Tiempo de ejecución**	**Épocas**	**batch_size**	**learning rate**	**beta(Adam)**
PlantVillage	80 % / 60 %	24h / 36h	400 / 15	128 / 15	0.0002 / 0.001	0.5 / –
UTKFace	100 % / 80 %	36h / 1.5sem	400 / 25	128 / 10	0.0002 / 0.001	0.5 / –
UTKFaceFull	–	–	–	–	–	–

Tabla B.2: Rendimiento e hiperparámetros óptimos para el experimento **WGAN**.

	Recursos		Hiperparámetros óptimos			
Dataset	**GPU RAM (%)**	**Tiempo de ejecución**	**Épocas**	**batch_size**	**learning rate**	**gamma**
PlantVillage	60 %	36h	25	15	1.00E-04	5
UTKFaceBias	80 %	1.5sem	30	10	1.00E-04	2
UTKFaceFull	80 %	1.5sem	30	10	1.00E-04	2

Tabla B.3: Rendimiento e hiperparámetros óptimos para el experimento **Loss Focal**.

	Recursos		Hiperparámetros óptimos			
Dataset	**GPU RAM (%)**	**Tiempo de ejecución**	**Épocas**	**batch_size**	**learning rate**	**gamma**
PlantVillage	60 %	36h	10	15	1.00E-03	2
UTKFaceBias	80 %	2sem	25	10	1.00E-04	0.5
UTKFaceFull	80 %	2sem	30	10	1.00E-04	2

Tabla B.4: Rendimiento e hiperparámetros óptimos para el experimento **Loss Focal Difusa**.

	Recursos		Hiperparámetros óptimos		
Dataset	**GPU RAM (%)**	**Tiempo de ejecución**	**Épocas**	**batch_size**	**learning rate**
PlantVillage	80 %	48h	10	15	1.00E-03
UTKFaceBias	90 %	1.5sem	25	10	1.00E-04
UTKFaceFull	–	–	–	–	–

Tabla B.5: Rendimiento e hiperparámetros óptimos para el experimento **PRM-IM**.

	Recursos		Hiperparámetros óptimos		
Dataset	**GPU RAM (%)**	**Tiempo de ejecución**	**Épocas**	**batch_size**	**learning rate**
PlantVillage	60 %	36h	25	15	1.00E-03
UTKFaceBias	80 %	1sem	25	10	1.00E-04
UTKFaceFull	80 %	1.5sem	25	10	1.00E-04

Tabla B.6: Rendimiento e hiperparámetros óptimos para el experimento **CoSenCNN**.


UÁM
Universidad Autónoma
de Madrid