Sesgo de género en el reconocimiento de imágenes: El análisis de los datos

Andoni Aizpuru
andoni.aizpuru@estudiante.uam.es
Universidad Autónoma de Madrid
Madrid, Madrid, España

Salvador Martín Barcia
salvador.amaya@estudiante.uam.es
Universidad Autónoma de Madrid
Madrid, Madrid, España

Álvaro Sánchez Romero
alvaro.sanchezromero@estudiante.uam.es
Universidad Autónoma de Madrid
Madrid, Madrid, España

Nicolás Serrano Salas
nicolas.serranos@estudiante.uam.es
Universidad Autónoma de Madrid
Madrid, Madrid, España

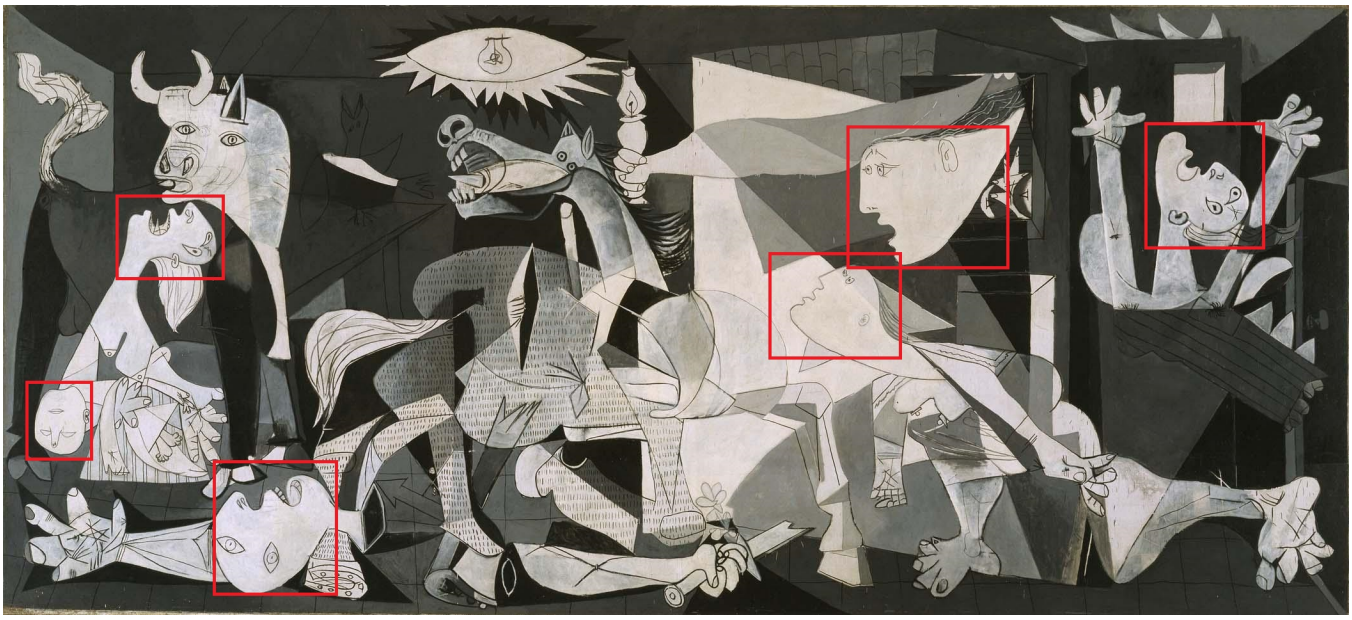

Figure 1: Reconocimiento de caras en el Guernica.

ABSTRACT

En la actualidad, una problemática muy común en el reconocimiento de género en imágenes faciales es el sesgo encontrado tanto en algoritmos como en datos. En este artículo se va a realizar un estudio del sesgo con el objetivo de mejorar los resultados globales de los modelos, en concreto, se estudiará el sesgo relacionado con los datos. En primer lugar, se observa cómo son utilizados los datos por los algoritmos de reconocimiento facial, seguido por una revisión de los principales datasets de imágenes faciales: FERET, GENDER-FERET, UTKFace y FairFace. A continuación, se tratará de estudiar los orígenes de estos sesgos en los datos, lo cual ayudará a descubrir estrategias para evitarlos. Para obtener mejores resultados generaremos una estrategia con el objetivo de crear datasets no sesgados. Esta estrategia seguirá una serie de pautas que utilizaremos en la creación de un nuevo dataset justo, utilizando imágenes combinadas de los dataset estudiados junto a imágenes generadas sintéticamente. Por último, se realizará una comparación entre datasets valorando cual da mejores resultados para diferentes algoritmos, comparándolos tanto de manera general como dividiendo los resultados por etnia y género. Esto nos permitirá observar los sesgos existentes.

KEYWORDS

Datos, Sesgo, Reconocimiento de Imágenes, Género

1 INTRODUCCIÓN

El sesgo de los algoritmos de aprendizaje automático a la hora de clasificar a las personas por género es, al igual que el sesgo que aparece en general con minorías y grupos sociales desfavorecidos, una problemática ampliamente documentada [15, 18, 28] y preocupante tanto desde el punto de vista técnico, pues obstaculiza las aplicaciones prácticas que puedan estar relacionadas con ello, como desde el punto de vista ético.

Un buen ejemplo de este tipo de sesgos aparece en [28], donde se busca clasificar mediante aprendizaje automático las temáticas de

una serie de cuadros. Cuando en los cuadros aparecen, por ejemplo, hombre con kimono o toga, el algoritmo los clasifica como mujeres.

La extensa literatura al respecto [1, 14, 15, 18, 28, 36] habla en general de dos causas posibles para este tipo de sesgo: el sesgo latente en los propios datos utilizados para el aprendizaje y el sesgo introducido (o amplificado) por los algoritmos que se utilizan para dicho aprendizaje. Además, también existe la posibilidad de que, por razones fisiológicas, ciertos grupos de personas resulten más difíciles de identificar en ciertos casos, tal como se discute en [15].

En este artículo se tratará de estudiar la primera posible causa de este sesgo, es decir, que venga del propio sesgo en la toma de los datos. En trabajos futuros se analizará el sesgo creado y amplificado por los propios algoritmos y posibles soluciones.

Una posible solución para el problema del sesgo de los datos provenientes de un sesgo humano sería usar datos dados por marcadores puramente biométricos, como huellas dactilares o imágenes de rayos X [3, 24], sin embargo en este estudio nos vamos a centrar principalmente en imágenes de caras.

Para ello estudiaremos algunos de los datasets públicos de caras más utilizados y compararemos el sesgo que se da con cada uno de ellos analizando si alguno presenta un sesgo menor. Tras ello se utilizará lo aprendido para tratar de crear unas pautas para generar un dataset equilibrado y poco sesgado.

2 ESTADO DEL ARTE

2.1 Algoritmos de reconocimiento facial

La detección y el análisis de las caras es un desafío en los problemas de visión artificial. La arquitectura tradicional para el reconocimiento de caras se basa en varias etapas. La primera, la detección de las caras en las imágenes, después alineación de las caras basándose en puntos clave de las caras, continuando por el procesado de las caras y finalizando con extracción de características para poder por último clasificar. Algunos modelos basados en redes convolucionales profundas, unifican este proceso [25] y algunos otros, además, se centran en la clasificación de estas imágenes faciales por género [7].

Las arquitecturas de estos modelos se basan en el Transfer Learning recortando capas de un modelo preentrenado y ajustándolas a los datos con los que se va trabajar [32]. Estos modelos utilizan implementaciones de arquitecturas nuevas de CNN que consiguen mejores resultados basándose en bloques de Inception que capturan características locales para diferentes tamaños de imagen y conexiones residuales que aceleran el entrenamiento de las redes, permitiendo entrenar modelos más profundos y con mejores resultados [11, 15, 29].

2.2 Datasets

Como ya se ha mencionado en la introducción, pueden utilizarse distintos tipos de datos para la clasificación con género, cada uno con sus propios sesgos y problemas éticos. Sin embargo, el tipo de datos más común que se utiliza en la literatura son imágenes de caras, de manera que este artículo se va a centrar en datasets formados por este tipo de datos. A continuación se detallan los orígenes de los datasets que vamos a utilizar en el trabajo:

2.2.1 FERET. FERET [23] es una base de datos formada por 14126 fotos a color de caras de 1199 individuos distintos, es en general una de las bases de datos más utilizadas en el ámbito del reconocimiento de caras y en concreto en la clasificación por género [6, 13, 19, 30], debido en parte a que es históricamente relevante, al haber sido publicada originalmente en 1996, cuando las bases de datos de esta magnitud escaseaban. Por ello, ha sido elegido como uno de los objetos de nuestro estudio.

Además, existe un dataset derivado llamado GENDER-FERET, utilizado en [7], que contiene 946 imágenes de caras en blanco y negro y que en teoría está equilibrado específicamente para la clasificación por género, de manera que comprobaremos si este equilibrio funciona para evitar el sesgo, y si es así qué podemos aprender de él.

En la Figura 2 podemos ver un ejemplo de imágenenes del dataset FERET. En este caso se trata de hacer detección facial de las caras mediante el cuadrado verde. Es interesante que en este dataset está disponible la misma cara desde distintos ángulos, lo cual va a permitir mejorar la predicción de los modelos entrenados.

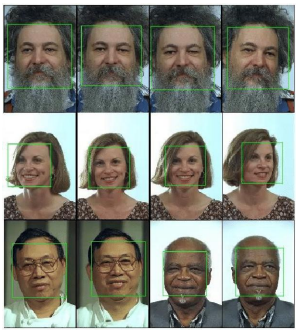

Figure 2: Ejemplo de imágenes de caras en FERET.

2.2.2 UTKFace. UTKFace es otra base de datos de imágenes de caras muy utilizada [8, 9, 16] y que resulta interesante debido a la gran cantidad de imágenes que contiene (más de 20000) y a que estas son muy variadas, tanto respecto a los sujetos de las imágenes, que son de todos los géneros, etnias y edades, como respecto a las propias características de las imágenes, como la iluminación, el color y resolución. Utilizaremos este dataset en el trabajo debido a que la gran variedad en las imágenes añaden ruido y lo acercan más a un caso de uso real, frente a otros datasets que utilizan fotografías "de estudio".

Figure 3: Ejemplo de imágenes de caras en UTKFace.

2.2.3 FairFace. FairFace [17] es una base de datos de imágenes de caras balanceada en cuanto a género, raza y edad, utilizada en [16]. Comprobaremos si este equilibrio en varias características funciona para reducir el sesgo en la clasificación por género, y si es así qué podemos aprender de él. En la siguiente gráfica (Figura 4) del artículo [17] se representa el t-distributed Stochastic Neighbor Embedding (t-SNE) de los datos del dataset. t-SNE es un algoritmo estocástico de representación de datos de alta dimensionalidad como pueden ser en nuestro caso imágenes. Se tiende a usar previamente PCA para el mejor funcionamiento del algoritmo. En la imagen se puede ver que los datos son separables en su mayoría y están bien agrupados, mientras que con otras representaciones de otros datasets que se analizan en el artículo no se produce esto.

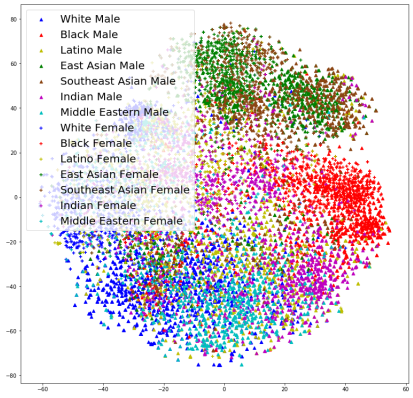

Figure 4: Representación t-SNE del dataset FairFace.

Por ejemplo, en la Figura 5 podemos ver que pese a estar bien distribuida hay un número mucho menor de etnias tenidas en cuenta. Además no se tiene una estructura circular, ni organización tan buena como la que aparecía para FairFace.

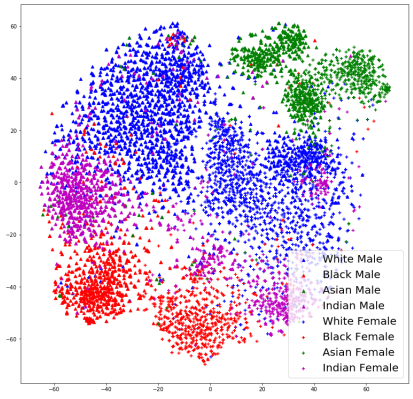

Figure 5: Representación UTKFace del dataset FairFace.

Es interesante señalar que FairFace es el dataset más completo a nivel racial dado que al contener información de 7 diferentes etnias permite hacer mejores predicciones de la etnia de una persona. Esto es debido a que en otros casos se usa el color de la piel para una distinguir la etnia de distintas personas. Esto puede fallar al tratarse de una variable dimensional (el color) y que además puede variar según la iluminación. Además, dentro del mismo grupo étnico no todos los individuos tienen el mismo color de piel, por ejemplo la gente del este del Asia tiene el mismo color de piel que la gente blanca pero no son de la misma etnia.

2.3 Sesgos

Se considera que un conjunto de datos tiene un sesgo o está sesgado cuando a raíz de estos se producen errores de predicción en los modelos de aprendizaje automático. Estos sesgos pueden venir por errores humanos al etiquetar, generar los datos (ruido) o por la propia naturaleza de los mismos. Dependiendo del tipo de sesgo estos podrán ser eliminados de una manera o de otra, siendo nuestro objetivo crear una metodología que permita generar un conjunto de imágenes que tengan el menor sesgo posible. Dado que queremos hacer las predicciones más justas y precisas posibles es fundamental eliminar los sesgos. Para lograr esto, el primer paso para ello es detectarlos:

2.3.1 Detección de sesgos: Algunos posibles sesgos o causas de sesgos son fácilmente detectables mediante una exploración de los datos. Es común tener datasets con clases muy desbalanceadas, lo que puede dar lugar a errores en la predicción. Pero hay otros sesgos que son más difíciles de detectar y no pueden ser directamente hallados mediante exploración, por ello se han creado métricas específicas para la detección de sesgos, como es el caso de la Relevance Mass Accuracy, Relevance Rank Accuracy y Area over the perturbation curve [22, 26, 27]. Estas métricas derivadas de la evaluación de mapas de atribución [34] (útiles en imágenes al explicar la influencia de los píxeles) pueden medir cómo de representativos son éstos teniendo en cuenta el sesgo. Existen otras métricas para medirlos pero están especializadas en el sesgo derivado de algoritmos (Selection Rate[32]) y en este artículo nos centraremos en el que proviene de los datos.

2.3.2 Sesgo por desbalanceo de datos: Una gran mayoría de los sesgos que se presentan en nuestros datos son debidos a la existencia de unos datos que no representan cuantitativamente la realidad o que tienen falta de representación para ciertos valores. Esta descompensación suele verse como un desbalanceo entre clases [21, 31, 37, 38] en el que una de ellas está poco representada, generando que se tienda a predecir más la clase más representada y se obtengan malos resultados para la que tiene pocas instancias. Pese a que existen algoritmos menos sensibles al desbalanceo de clases esto sigue siendo un gran problema.

Una solución muy usada en clasificación de imágenes es la técnica de data augmentation mediante la cual se generan nuevas instancias de las antiguas a partir de rotaciones y operaciones en las anteriores imágenes. A parte de esto es muy común el uso de Generative Adversarial Networks (GANs) para la creación sintética de datos. Por ejemplo, se usan para la generación imágenes de gente que no existe. Estas imágenes pueden servir para incrementar los datos que introducimos a nuestra red.

La arquitectura básica de una GAN puede verse en la Figura 6. Las redes adversarias están formadas por dos submodelos, el generador y el discriminador. El generador está encargado de generar nuevas

instancias que sean plausibles dentro de nuestro dominio, es decir, en nuestro caso imágenes de personas. Este submodelo recibe de entrada vectores con ruido gaussiano que actúen como semilla de la imagen generada. Por otro lado, tenemos un submodelo que es el discriminador. Éste va a tratar de clasificar las imágenes como reales o falsas, tratando de predecir si la imagen que ha recibido ha sido creada por el generador. La idea detrás de las GANs es que el generador logre hacer imágenes que engañen al discriminador de manera que crea que son reales. El discriminador va a servir para entrenar al generador a generar imágenes que parezcan reales, de manera que cuando acabemos con el entrenamiento tan solo necesitaremos la parte del generador.

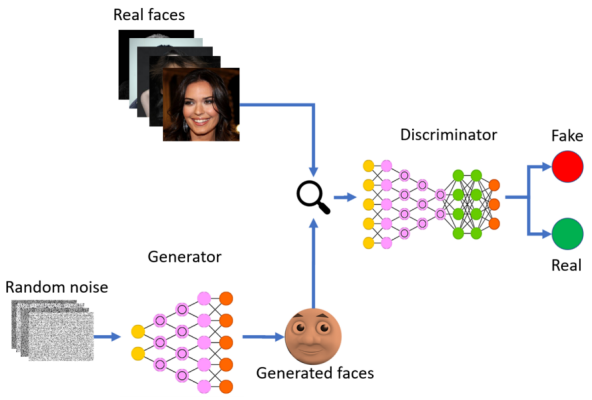

Figure 6: Esquema de una Generative Adversarial Networks.

Por otro lado, no debemos olvidar que para obtener la mejor predicción posible no solo las clases deben estar bien representadas sino cualquier atributo [15, 32]. Por ejemplo, de poco sirve si nuestro dataset tiene un buen número de imágenes de hombres y mujeres si no hemos representado adecuadamente a gente de distintas etnias. Cada etnia puede tener diferentes rasgos diferenciales entre géneros por lo que un dataset en el que prácticamente solo haya gente blanca no va a funcionar bien cuando aparezca gente de otras razas.

Al margen del marco étnico debemos pensar en cualquier otro grupo menos representados desde ancianos a gente del colectivo LGBT. Cuánto mejor representemos a estos grupos mejores serán nuestros resultados y a más países y contextos podrá ser aplicado nuestros datos.

En el contexto de las GANs lo que se acaba de mencionar de tener datasets con atributos balanceados es fundamental dado que si generamos instancias mediante un dataset con un claro desbalanceo o sesgo, nuestra red adversaria lo reproducirá. Por tanto es muy importante comprender con qué datos estamos generando nuestras nuevas instancias y para qué lo hacemos. En la siguiente subsección profundizaremos en este tema y técnicas para resolverlo, más específicamente en el contexto de imágenes faciales.

2.3.3 Sesgo debido a técnicas de generación sintética de datos: Pese a la utilidad de estas técnicas, principalmente GANs, tienen el inconveniente que se puede producir un sesgo entre los datos reales y los sintéticos[35]. Una opción para tratar esto es el uso de redes adversarias como FESGAN que generen nuevas identidades con distintas expresiones y a las que le añadamos un distinto tipo de back-propagation (RDBP) al clasificar que disminuya las variaciones dentro de la misma clase. En la Figura 7 podemos ver el esquema de una Facial Expression Synthesis GAN (FESGAN) que a diferencia el de la GAN anteriormente observada, tiene dos disciminadores. Además, el generador tiene ahora una estructura de autoencoder que se va a encargar de aprender el estilo de contenido de imágenes de caras reales. El nuevo discriminador se encarga de determinar si la cara introducida tiene el gesto que tenía la cara original u otro diferente.

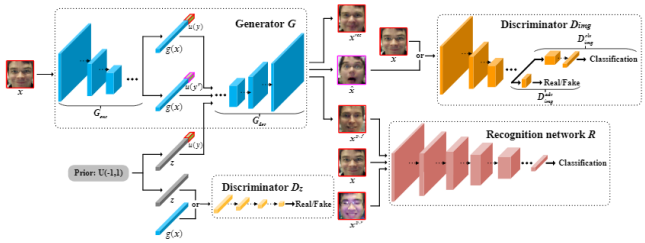

Figure 7: Esquema de una red FESGAN.

Otro problema es la generación de discrepancia entre instancias de atributos por realizar data augmentation o usar GANs [33]. Usando redes siamesas podemos minimizar las discrepancias en la distribución de variables de ambas fuentes anteriormente mencionadas.

2.3.4 Sesgo por etiquetado de datos: Al necesitar una cantidad tan ingente de datos nuestros algoritmos de aprendizaje automático, no siempre los etiquetados son correctos[10]. Estos errores previos al entrenamiento son importantes de evitar dado que si entrenamos nuestro modelo con datos erróneos vamos a obtener predicciones basadas en datos erróneos y por tanto, malas (Garbage in, garbage out). El sesgo por etiquetado puede ser debido a errores de anotación de atributos hasta clases. Hemos de considerar que los errores de etiquetado pueden ser a causa de personas o a posibles etiquetados automáticos.

Un caso importante a tener en cuenta en la predicción de género es que en prácticamente ningún dataset ni paper se tiene en cuenta que hay gente que no se identifica con ningún género [32]. Una posibilidad sería añadir un etiquetada como Otro o No binario, para funcionar de forma adecuada con esta gente.

2.3.5 Sesgo debido a ruido: Es muy probable que en nuestro conjunto de datos tengamos datos, en este caso, imágenes con ruido. Como es normal en el mundo real no todos nuestros datos de aprendizaje van a ser perfectos o tener una definición perfecta y eso puede provocar que nuestros modelos aprendan peor. Para atajar problemas de ruido se puede usar URNet [18] que permite dar diferentes pesos a cada instancia de entrenamiento según la confianza que tengamos en estas. Esto puede servir también para atajar otros sesgos como desbalanceo de datos o problemas de etiquetado.

2.3.6 Sesgo contextual: Este tipo de sesgo es aquel que hace que nuestras predicciones empeoren al tener en cuenta el contexto de la imagen [2, 4, 5, 12]. Por ejemplo, si entrenamos con imágenes de

hombres haciendo deporte, se producen errores al predecir mujeres que están realizando algún deporte. Debido a esta problemática se puede hacer preprocesamiento de los datos para detectar la cara o el cuerpo.

Otra solución en vez de hacer hacer un preprocesado de la imagen donde se seleccione la cara de la persona, puede consistir en el uso de otro modelo (Equalizer) y funciones de error [5]. El modelo Equalizer se caracteriza porque se concentra en la figura de la persona, ignorando el contexto. En la Figura 8 se puede ver en la izquierda cómo se ha predicho erróneamente al usar píxeles de la imagen que no pertenencen al cuerpo mientras que en la derecha mediante Equalizer se predice bien gracias a que se usan los píxeles correctos. Además, las dos funciones de error que se pueden añadir para mitigar el sesgo son la Appearance Confussion Loss y Confident Loss.

Figure 8: Comparativa de predicción con sesgo contextual y sin sesgo contextual.

3 PROPUESTA

En el estado del arte hemos visto distintas fuentes de sesgo que pueden ser causas del sesgo de género. Es por eso que hemos escogido cuatro datasets con características distintas en los que podremos medir las influencias de estos sesgos: FERET, un dataset generalista cuyas imágenes tienen poco ruido, UTKFace, un dataset generalista con más ruido, GENDER-FERET, un dataset creado de tal manera que sea balanceado en cuanto a géneros, y FairFace, que además de ser balanceado en cuanto a géneros también lo es en cuanto a etnias.

Por lo tanto, la primera parte de nuestra propuesta consiste en medir el sesgo de género existente en cada uno de los datasets y tratar de detectar los orígenes de dicho sesgo en cada uno de ellos.

Con estos resultados podremos identificar las fuentes de sesgo más importantes presentes en este tipo de datos y analizar, si es el caso, por qué algunos datasets presentan más sesgo que otros. En particular estudiaremos si GENDER-FERET y FairFace, datasets balanceados para la clasificación por género, es notablemente menos insesgado que el resto, lo que nos permitirá juzgar la importancia del sesgo por desbalanceo de datos en esta problemática, asimismo, la comparación entre FERET y UTKFace nos ayudará a comprender si el ruido en las imágenes juega un papel importante en el sesgo. Por último, también se medirá este sesgo separando las imágenes por etnias y grupos de edad, para observar si estas características contextuales de las personas afectan al sesgo de género.

La segunda parte de la propuesta consiste en crear una estrategia para generar datasets que minimicen los tipos de sesgos más importantes que se hayan encontrado anteriormente, con lo que generaremos una serie de pautas que ayuden a generar datasets insesgados, pautas que, por último, llevaremos a la práctica, utilizándolas para generar nuestro propio dataset combinando imágenes de los cuatro datasets que hemos escogido y creando nuevas imágenes sintéticas mediante GANs y FESGANs según se considere necesario, lo que permitirá evaluar si el sesgo de género se reduce considerablemente siguiendo estas pautas.

4 DISEÑO DE LA EXPERIMENTACIÓN

Primero, de acuerdo con la propuesta anterior, se hace un estudio de los diferentes sesgos sobre los diferentes datasets en los que se va a trabajar y compara directamente, sin ningún tipo de preprocesado y procesando mediante URNet atajando posibles problemas del sesgo debido a ruido. Para ello utilizamos los datasets UTKFace, FERET, GENDER-FERET y FairFace sobre diferentes algoritmos del estado del arte como son Inception-v4, ResNet-101 y Inception-ResNet. Para la comparación se utilizan la métricas de accuracy, recall y F1 como observamos en las tablas 1, 2, 3, 4, 5 y 6.

En la tabla 7, se observa la precisión en la clasificación de género divido entre los diferentes grupos étnicos de los datasets obtenidos para el mejor modelo que hemos obtenido. Estos grupos étnicos han sido extraídos de la división que hace el dataset FairFace. Así mismo en la tabla 8, se observa la precisión en la clasificación de género divido entre diferentes grupos de edades.

Por último en la tabla 9, se observa la precisión en la clasificación de género dividido en los distintos grupos étnicos y grupos de edades anteriores para nuestro dataset, evaluado en el mejor modelo.

Siguiendo con la propuesta, se crea un dataset extrayendo imágenes faciales de GENDER-FERET, UTKFace y FairFace. Así mismo se completa con datos sintéticos, utilizando redes siamesas, que reflejen distintos daños estéticos para rostros de diferentes etnias, edades y sexo. Esto permite evitar el sesgo contextual al utilizar únicamente los rostros y no el contexto de la imagen, sesgo de etiquetado de datos al utilizar datos que están perfectamente etiquetados provenientes de otro dataset, el sesgo de desbalanceo de datos al crear el dataset teniendo en cuenta este problema y minimizando el sesgo debido a técnicas de generación sintética de datos al minimizar las discrepancias en la distribución de variables con el uso de las redes siamesas. Así se utiliza la red FESGAN generando distintas expresiones faciales y minimizando más aun el sesgo debido a técnicas de generación sintética de datos.

Finalmente se estudia el sesgo del dataset creado siguiendo los mismos experimentos anteriormente mencionados y comparándolo UTKFace, FERET, GENDER-FERET y FairFace en las tablas 1, 2, 3, 4, 5, 6, 7, 8 y 9.

5 CONCLUSIONES

A pesar de sus grandes ventajas, los avances que se han dado en los últimos tiempos en el campo del Aprendizaje Automático han puesto en evidencia que estos tienden a reproducir sesgos injustos

	Accuracy	Recall	F1
UTKFace	0.762		
FERET			
GENDER-FERET			
FairFace	0.7148		
Nuestra			

Table 1: Precisión, Recall y F1 con Inception-v4, para la clasificación de género en distintos datasets de imágenes faciales, sin preprocesado.

	Accuracy	Recall	F1
UTKFace			
FERET			
GENDER-FERET			
FairFace			
Nuestra			

Table 2: Precisión, Recall y F1 con ResNet-101, para la clasificación de género en distintos datasets de imágenes faciales, sin preprocesado.

	Accuracy	Recall	F1
UTKFace			
FERET			
GENDER-FERET			
FairFace			
Nuestra			

Table 3: Precisión, Recall y F1 con Inception-ResNet, para la clasificación de género en distintos datasets de imágenes faciales, sin preprocesado.

	Accuracy	Recall	F1
UTKFace			
FERET			
GENDER-FERET			
FairFace			
Nuestra			

Table 4: Precisión, Recall y F1 con Inception-v4, para la clasificación de género en distintos datasets de imágenes faciales, tras preprocesado con URNet.

	Accuracy	Recall	F1
UTKFace			
FERET			
GENDER-FERET			
FairFace			
Nuestra			

Table 5: Precisión, Recall y F1 con ResNet-101, para la clasificación de género en distintos datasets de imágenes faciales, tras preprocesado con URNet.

	Accuracy	Recall	F1
UTKFace			
FERET			
GENDER-FERET			
FairFace			
Nuestra			

Table 6: Precisión, Recall y F1 con Inception-ResNet, para la clasificación de género en distintos datasets de imágenes faciales, tras preprocesado con URNet.

de origen humano, como es el caso del sesgo de género, que se estudia en este paper. Este trabajo pretende tratar de ayudar a entender y solucionar este sesgo centrándose en el papel que juegan datasets que se utilizan para entrenar los algoritmos en la reproducción del mismo, y desarrollando unas pautas para crear datasets que no lo reproduzcan.

Por ello, nos hemos centrado en cuatro datasets populares de imágenes de caras, dos de propósito general (FERET y UTKFace), y dos creados específicamente para ser balanceados en cuanto a género (GENDER-FERET y FairFace), para estudiar la prevalencia del sesgo con todos ellos y tratar de extraer que características hacen a unos más insesgados que a otros. Esta información se ha utilizado para crear un nuevo dataset más insesgado que ninguno de los cuatro originales que se han utilizado, lo cual lo convierte en un dataset de especial interés para trabajos futuros que traten de conseguir una clasificación de género con el menor sesgo posible y se ha dado la serie de pautas que se ha seguido para crear dicho dataset, lo cual facilitará a futuras investigaciones la creación de datasets insesgados, tanto en cuanto a género como en cuanto a otros sesgos de origen cultural, como pueden ser los relacionados con la etnia.

Para trabajos futuros sería especialmente interesante realizar una investigación similar para estudiar la otra posible fuente de sesgo: los algoritmos de clasificación; de forma que se compare el peso relativo del sesgo generado por estos con el generado por los datos, si existen unos algoritmos más claramente sesgados que otros y posibles estrategias para evitar estos sesgos. Además, podrán surgir lineas de investigación similares que estudien las fuentes de otros sesgos como los relacionados con la etnia o la edad.

Por último, puesto que la mayoría de estas clasificaciones en la actualidad se realizan utilizando modelos de caja negra como las redes neuronales, también podría ser interesante que trabajos futuros utilizaran Inteligencia Artificial Explicable (XAI) para tratar de entender las decisiones de los modelos que llevan a la reproducción del sesgo de género, lo cual podría resultar interesante tanto desde el punto de vista técnico, para tratar de evitar estos sesgos, como desde un punto de vista más teórico, para entender como los sesgos humanos se reproducen en los algoritmos de aprendizaje automático, además de poder realizar aportaciones a un campo con gran potencial pero que aun está en su infancia, como es el XAI.

REFERENCES

[1] Osman Aka, Ken Burke, Alex Bauerle, Christina Greer, and Margaret Mitchell. 2021. Measuring Model Biases in the Absence of Ground Truth. In *AIES 2021 - Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 327–335.

Etnia	Blanca		Oriente medio		Negra		Este Asiático		Sudeste Asiático		India		Latina	
Género	M	F	M	F	M	F	M	F	M	F	M	F	M	F
UTKFace														
FERET														
GENDER-FERET														
FairFace														
Nuestra														

Table 7: Precisión del género en los diferentes grupos étnicos según el mejor modelo, en los distintos datasets.

Edad	Infancia		Adolescencia		Juventud		Adultez		Vejez	
Género	M	F	M	F	M	F	M	F	M	F
UTKFace										
FERET										
GENDER-FERET										
FairFace										
Nuestra										

Table 8: Precisión del género en los diferentes grupos de edades según el mejor modelo, en los distintos datasets.

Etnia	Blanca		Oriente medio		Negra		Este Asiático		Sudeste Asiático		India		Latina	
Género	M	F	M	F	M	F	M	F	M	F	M	F	M	F
Infancia														
Adolescencia														
Juventud														
Adultez														
Vejez														

Table 9: Precisión del género en los diferentes grupos de etnias y edades según el mejor modelo.

[2] Luis Felipe de Araujo Zeni and Claudio Rosito Jung. 2018. Real-Time Gender Detection in the Wild Using Deep Neural Networks. In *2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*. 118–125. https://doi.org/10.1109/SIBGRAPI.2018.00022

[3] K. S Arun and K. S Sarath. 2011. A machine learning approach for fingerprint based gender identification. In *2011 IEEE Recent Advances in Intelligent Computational Systems*. IEEE, 163–167.

[4] Kaylee Burns, Lisa Anne Hendricks, Trevor Darrell, and Anna Rohrbach. 2018. Women also Snowboard: Overcoming Bias in Captioning Models. *CoRR* abs/1803.09797 (2018). arXiv:1803.09797 http://arxiv.org/abs/1803.09797

[5] Kaylee Burns, Lisa Anne Hendricks, Kate Saenko, Trevor Darrell, and Anna Rohrbach. 2018. Women also Snowboard: Overcoming Bias in Captioning Models. https://doi.org/10.48550/ARXIV.1803.09797

[6] Sahar Dammak, Hazar Mliki, and Emna Fendri. 2021. Gender effect on age classification in an unconstrained environment. *Multimedia tools and applications* 80, 18 (2021), 28001–28014.

[7] Neelam Dwivedi and Dushyant Kumar Singh. 2018. Review of Deep Learning Techniques for Gender Classification in Images. In *Harmony Search and Nature Inspired Optimization Algorithms*. Springer Singapore, Singapore, 1089–1099.

[8] Avishek Garain, Biswarup Ray, Pawan Kumar Singh, Ali Ahmadian, Norazak Senu, and Ram Sarkar. 2021. GRA_Net: A Deep Learning Model for Classification of Age and Gender From Facial Images. *IEEE access* 9 (2021), 85672–85689.

[9] Anirudh Ghildiyal, Sachin Sharma, Ishita Verma, and Urvi Marhatta. 2020. Age and Gender Predictions using Artificial Intelligence Algorithm. In *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)*. IEEE, 371–375.

[10] Byungok Han, Woo-Han Yun, Jang-Hee Yoo, and Won Hwa Kim. 2020. Toward Unbiased Facial Expression Recognition in the Wild via Cross-Dataset Adaptation. *IEEE Access* 8 (2020), 159172–159181. https://doi.org/10.1109/ACCESS.2020.3018738

[11] Md. Mahbubul Islam, Nusrat Tasnim, and Joong-Hwan Baek. 2020. Human Gender Classification Using Transfer Learning via Pareto Frontier CNN Networks. *Inventions (Basel)* 5, 2 (2020), 16.

[12] Luo Jiang, Juyong Zhang, and Bailin Deng. 2018. Robust RGB-D Face Recognition Using Attribute-Aware Loss. *CoRR* abs/1811.09847 (2018). arXiv:1811.09847 http://arxiv.org/abs/1811.09847

[13] V. V Khryashchev, L. A Shmaglit, A. L Priorov, and A. M Shemyakov. 2014. Extracting adaptive features for gender classification of human face images. *Programming and computer software* 40, 4 (2014), 215–221.

[14] Michael Kim, Amirata Ghorbani, and James Zou. 2019. Multiaccuracy: Black-Box Post-Processing for Fairness in Classification. In *Proceedings of the 2019 AAAI/ACM Conference on ai, ethics, and society (AIES '19)*. ACM, 247–254.

[15] Anoop Krishnan, Ali Almadan, and Ajita Rattani. 2020. Understanding Fairness of Gender Classification Algorithms Across Gender-Race Groups. *CoRR* abs/2009.11491 (2020). arXiv:2009.11491 https://arxiv.org/abs/2009.11491

[16] Anoop Krishnan, Ali Almadan, and Ajita Rattani. 2020. Understanding Fairness of Gender Classification Algorithms Across Gender-Race Groups. In *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 1028–1035.

[17] Kimmo Kärkkäinen and Jungseock Joo. 2019. FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age. https://doi.org/10.48550/ARXIV.1908.04913

[18] Jia Li, Yafei Song, Jianfeng Zhu, Lele Cheng, Ying Su, Lin Ye, Pengcheng Yuan, and Shumin Han. 2018. Learning from Large-scale Noisy Web Data with Ubiquitous Reweighting for Image Classification. *CoRR* abs/1811.00700 (2018). arXiv:1811.00700 http://arxiv.org/abs/1811.00700

[19] E. K Loo, T. S Lim, L. Y Ong, and C. H Lim. 2018. The influence of ethnicity in facial gender estimation. In *2018 IEEE 14th International Colloquium on Signal Processing Its Applications (CSPA)*. IEEE, 187–192.

[20] Oge Marques, Jhanon James, and Emilio Barcelos. 2018. Face-It-Up: a scientific app for face processing using mobile devices and machine learning APIs. In *Mobile Multimedia/Image Processing, Security, and Applications 2018*, Sos S. Agaian and Sabah A. Jassim (Eds.), Vol. 10668. International Society for Optics and Photonics, SPIE, 151 – 159. https://doi.org/10.1117/12.2307765

[21] Nasik Muhammad Nafi and William H. Hsu. 2020. 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)*. 243–248. https://doi.org/10.1109/IWSSIP48289.2020.9145239

[22] Ahmed Osman, Leila Arras, and Wojciech Samek. 2020. Towards Ground Truth Evaluation of Visual Explanations. *ArXiv* abs/2003.07258 (2020).

[23] P.J. Phillips, Hyeonjoon Moon, P. Rauss, and S.A. Rizvi. 1997. The FERET evaluation methodology for face-recognition algorithms. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 137–143. https://doi.org/10.1109/CVPR.1997.609311

[24] M.V Rajee and C Mythili. 2021. Gender classification on digital dental x-ray images using deep convolutional neural network. *Biomedical signal processing and control* 69 (2021), 102939.

[25] Rajeev Ranjan, Vishal M Patel, and Rama Chellappa. 2019. HyperFace: A Deep Multi-Task Learning Framework for Face Detection, Landmark Localization, Pose Estimation, and Gender Recognition. *IEEE transactions on pattern analysis and machine intelligence* 41, 1 (2019), 121–135.

[26] Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. 2017. Evaluating the Visualization of What a Deep Neural Network Has Learned. *IEEE Transactions on Neural Networks and Learning Systems* 28, 11 (2017), 2660–2673. https://doi.org/10.1109/TNNLS.2016.2599820

[27] Nina Schaaf, Omar de Mitri, Hang Beom Kim, Alexander Windberger, and Marco F. Huber. 2021. Towards Measuring Bias in Image Classification. *CoRR* abs/2107.00360 (2021). arXiv:2107.00360 https://arxiv.org/abs/2107.00360

[28] Sudeepti Surapaneni, Sana Syed, and Logan Yoonhyuk Lee. 2020. Exploring Themes and Bias in Art using Machine Learning Image Analysis. In *2020 Systems and Information Engineering Design Symposium (SIEDS)*. IEEE, 1–6.

[29] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alex Alemi. 2016. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. (2016).

[30] P Vallimeena, Uma Gopalakrishnan, Bhavana B Nair, and Sethuraman N Rao. 2019. CNN Algorithms for Detection of Human Face Attributes - A Survey. In *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*. IEEE, 576–581.

[31] Matthew J. Vowels, Necati Cihan Camgöz, and Richard Bowden. 2020. NestedVAE: Isolating Common Factors via Weak Supervision. *CoRR* abs/2002.11576 (2020). arXiv:2002.11576 https://arxiv.org/abs/2002.11576

[32] Wenying Wu, Pavlos Protopapas, Zheng Yang, and Panagiotis Michalatos. 2020. Gender Classification and Bias Mitigation in Facial Images. In *12th ACM Conference on Web Science (WebSci '20)*. ACM, 106–114.

[33] Yan Yan, Ying Huang, Si Chen, Chunhua Shen, and Hanzi Wang. 2020. Joint Deep Learning of Facial Expression Synthesis and Recognition. *CoRR* abs/2002.02194 (2020). arXiv:2002.02194 https://arxiv.org/abs/2002.02194

[34] Mengjiao Yang and Been Kim. 2019. BIM: Towards Quantitative Evaluation of Interpretability Methods with Ground Truth. *CoRR* abs/1907.09701 (2019). arXiv:1907.09701 http://arxiv.org/abs/1907.09701

[35] Getinet Yilma, Kumie Gedamu, Maregu Assefa, Ariyo Oluwasanmi, and Zhiguang Qin. 2021. Generation and Transformation Invariant Learning for Tomato Disease Classification. In *2021 IEEE 2nd International Conference on Pattern Recognition and Machine Learning (PRML)*. 121–128. https://doi.org/10.1109/PRML52754.2021.9520693

[36] Yi Zhang and Jitao Sang. 2020. Towards Accuracy-Fairness Paradox: Adversarial Example-based Data Augmentation for Visual Debiasing. In *MM 2020 - Proceedings of the 28th ACM International Conference on Multimedia*. 4346–4354.

[37] Yufei Zhao, Jinxin Yang, Jiangtao Du, Zhen Chen, and Wen-Chi Yang. 2021. A Lightweight Classifier for Facial Expression Recognition based on Evolutionary SVM Ensembles. In *2021 6th International Conference for Convergence in Technology (I2CT)*. 1–9. https://doi.org/10.1109/I2CT51068.2021.9417940

[38] Zhisheng Zhong, Jiequan Cui, Shu Liu, and Jiaya Jia. 2021. Improving Calibration for Long-Tailed Recognition. *CoRR* abs/2104.00466 (2021). arXiv:2104.00466 https://arxiv.org/abs/2104.00466

A TRABAJO BIBLIOGRÁFICO

A.1 Extracción de artículos de las bases de datos

Primero, para hacer la búsqueda bibliográfica en las diferentes bases de datos de documentos hemos dividido la tarea en dos consultas. La primera se centra en técnicas y algoritmos de reconocimiento de imágenes que se centren en el género de las personas, y la segunda se centra en el sesgo que se encuentra en el reconocimiento de imágenes intentando centrarnos en los datos. Hemos utilizado las bases de datos de Web of Science, Scopus e IEEE.

Para la primera consulta, hemos extraído las siguientes palabras clave: people, person, human, gender, "image recognition", "image classification", "data science", "deep learning", "machine learning", ML, algorithm y technique. Estas consultas se pueden observar en las figuras 9, 10 y 11.

```
(
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 learning" 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"
)
```

Figure 9: Consulta Web of Science, primera búsqueda

```
(
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 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("algorithm") OR
TITLE-ABS-KEY("technique")
)
```

Figure 10: Consulta Scopus, primera búsqueda

Para la segunda consulta hemos extraído las siguientes palabras clave: bias, "image recognition", "image classification", "data science", "deep learning", "machine learning", ML, data, dataset. Estas consultas se pueden observar en las figuras 12, 13 y 14.

Para ambas consultas nos hemos centrado en los últimos cinco años (2018-2022), y al ámbito de la ingeniería informática, ya que nos salían bastantes datos biomédicos que no nos interesaban debido a que está fuera de nuestro dominio. Tras todo este proceso y eliminando la intersección de artículos para la misma consulta en la misma base de datos nos hemos quedado con un total de 274 artículos de los cuales tendremos que hacer un prefiltrado para saber si son interesantes para el artículo que vamos a desarrollar. Creemos

```
(
"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"
)
```

Figure 11: Consulta IEEE, primera búsqueda

```
(
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"
)
```

Figure 12: Consulta Web of Science, segunda búsqueda

que es una preselección adecuada, pues leyendo por encima los resúmenes encontramos mucha variedad y completitud del estado del arte actual.

A.2 Filtrado de los artículos

En la hoja de cálculo adjunta, se muestra en las hojas "Consulta Gender" y "Consulta Bias" todos los artículos que hemos extraído a través de la consulta, de qué base de datos en concreto procede, si los hemos escogido como válidos o no, y el motivo de ello individualmente.

En la primera consulta hemos eliminado los artículos que nos salían que no tenían que ver con imágenes 2D estáticas, imágenes concretas de ámbitos médicos o biométricos que sean diferentes a la cara o cuerpo completo, como por ejemplo labios, palmas de las

```
(
   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")
)
```

Figure 13: Consulta Scopus, segunda búsqueda

```
(
   "bias"
)
AND
(
   "image recognition" OR
   "image classification"
)
AND
(
   "data science" OR
   "deep learning" OR
   "machine learning" OR
   "ML"
)
AND
(
   "data" OR
   "dataset"
)
```

Figure 14: Consulta IEEE, segunda búsqueda

manos o huellas dactilares e imágenes captadas fueras del espectro de la luz visible, como por ejemplo imágenes infrarrojas o ultravioletas. También hemos eliminado artículos que no trataban de estimar, identificar, reconocer o clasificar el género en las imágenes, ya que algunos solo se centraban en etnia o edad. No obstante, hemos cogido para leer con detenimiento aquellos que utilizan técnicas que aparte de clasificar género, clasifican por edad, etnia u otros factores. Por último también hemos eliminado los artículos que se centran solo en un conjunto de la población, por ejemplo únicamente en personas ancianas o únicamente en personas asiáticas.

En la segunda consulta hemos eliminado los artículos que no hablaban del sesgo, del reconocimiento de imágenes o que hablaban de alguno de los tipos de imágenes que hemos eliminado de la primera consulta como datos biomédicos. También hemos tenido que eliminar muchos artículos que lo único que mencionan respecto del sesgo es que su algoritmo tiene o no tiene ningún detalle más. Por último hemos eliminado los que tienen que ver con los sesgos inductivos que tienen los diferentes algoritmos.

Tras este filtrado de los 274 artículos, 114 para la primera consulta y 160 para la segunda, nos hemos quedado con un total de 59

artículos de los cuales 31 son de la primera consulta y 28 de la segunda.

A.3 Tabla de atributos

En la hoja de cálculo adjunta, en las hojas "Criterios Gender" y "Criterios Bias" se muestran las tablas a rellenar donde extraemos los atributos claves para la realización de nuestro artículo. Para la primera consulta los atributos a valorar son: Título, Año de publicación, ¿Accesible?, Propone dataset, Origen dataset, Dataset Pública / Privada, Link dataset, Tipo de dato (Cara, cuerpo), Supervisado / No supervisado, ¿Transfer learning?, Nombre transfer learning, Algoritmos Utilizados, ¿Algoritmo reproducible / usable?, ¿Preprocesado?, Métrica, Score, ¿Trata edad?, ¿Trata etnia?, ¿Género binario? y Resumen en un máximo de dos oraciones.

Para la segunda son: Título, Año de publicación, ¿Accesible?, Propone dataset, Origen dataset, Dataset Pública / Privada, Link dataset, Tipo de dato (Cara, cuerpo), Algoritmo Utilizado, Sesgo en datos o en algoritmo, Tipo de sesgo, ¿Propone como corregir el sesgo?, Cómo, Género, Minorías, Métrica, Score y Resumen en un máximo de dos oraciones.

Tras la lectura de los artículos habiendo recogido la información de estos, nos hemos encontrado algún problema puntual como que la publicación de alguno de ellos estaba bajo un repositorio el cual no teníamos acceso como estudiantes de la UAM de forma gratuita [20], ni accesible de ninguna otra forma, lícita o no. En estos casos hemos intentado extraer la mayor información posible del abstract del mismo, pues es la única información de la que disponíamos.

En resumen, primero para los artículos de la primera búsqueda y haciendo un resumen del estado del arte, encontramos que la mayoría de los artículos utilizan exclusivamente el uso de datasets donde los datos que se obtienen son caras humanas, sin embargo encontramos un par que utilizan el cuerpo de la persona o incluso manos y vestimenta. También hay gran variedad de las técnicas utilizadas dando todas buenos resultados, donde el accuracy es la métrica más utilizada. Se utiliza desde aprendizaje supervisado, semisupervisado y no supervisado donde encontramos algoritmos como CNN con y sin transfer learning, SVM, algoritmos genéticos, KNN o conjuntos de clasificadores. Destacamos que una gran parte de los artículos hacen un preprocesamiento previo de los datos, donde se extraen diferentes características. Observamos muchos artículos que junto con el reconocimiento del género clasifican también la edad, etnia o ambas de las personas, favoreciendo esto a tener mejores resultados en los diferentes conjuntos poblacionales, no obstante solo encontramos uno que no trate el género como binario. Por último podemos contrastar nuestros futuros resultados con la mayoría de propuestas que hemos leído, pues tenemos o acceso a los datasets los cuales se han utilizar para evaluar los modelos o tenemos la información para reproducir el modelo y el preprocesado del mismo.

Por otro lado, para los artículos de la segunda búsqueda de los 28 artículos observamos que 18 de ellos tratan de sesgo en los datos y 12 de ellos sesgo en los algoritmos. Se destaca sobretodo en muchos de ellos el sesgo producido por el desbalanceo de los datos, no obstante existe también sesgo debido al data augmentation, a la utilización de datos sintéticos, sesgo producido en el propio etiquetado de los datos o por las propias características locales de forma que extrae el algoritmo. Para corregir muchos de estos sesgos se utilizan diferentes técnicas de preprocesamiento de los datos, utilizar una función de perdida distinta que tenga en cuenta el desbalanceo de los datos, un mejor ajustado de los hiperparámetros, etc. Observamos que pocos de estos artículos, pero algunos, tratan sobre género o etnia de personas, pues no todos los datasets son de humanos, encontramos datasets de plantas, números o imágenes con miles de clases. Por último la mayoría de artículos usan el accuracy como métrica pero viendo los experimentos llevados a cabo no es muy representativo ya que cada uno hace medidas sobre datos y algoritmos muy distintos, sin embargo y viendo lo que utilizan otros artículos en menor medida creemos que es mejor la utilización de la métrica F1 o incluso la visualización de matrices de confusión cuando sea posible.