Pontificia Universidad Católica Madre y Maestra

Escuela de Ingeniería en Computación y Telecomunicaciones

Facultad de Ciencias e Ingeniería

Evaluación comparativa de algoritmos de reconocimiento facial en la identificación de características demográficas

Metodología de la Investigación

CSTI-1890-4341

Cristian de la Hoz

(1014-9779)

Jean Carlos Pérez Ortega

(1014-8917)

Randy Alexander Germosén Ureña

(1013-4707)

Prof. Arlene Estévez

Julio de 2025

Índice

1 Resumen **1**

2 abstract **1**

3 Introducción **2**

3.1 Modelos utilizados 2

4 Desarrollo **3**

4.1 Metodología 3

4.2 Hipótesis 3

4.3 Variables 3

4.4 Estrategia del muestreo 4

4.5 Materiales y Métodos 4

4.6 Algoritmos usados 5

4.7 Evaluación de desempeño 5

4.8 Resultados 6

4.8.1 Análisis de resultados utilizando OpenCV 6

4.8.2 Análisis de resultados utilizando DLib 10

4.8.3 Análisis de resultados utilizando FaceNet 13

5 Conclusiones **17**

6 Trabajos futuros **17**

Índice de figuras

1 Resultados de OpenCV en la predicción de edad 6

2 Resultados de OpenCV en la predicción de sexo 8

3 Resultados de DLib en la predicción de sexo 10

4 Resultados de DLib en la predicción de raza 11

5 Resultados de DLib en la predicción de edad 12

6 Resultados de FaceNet en la predicción de raza 14

7 Resultados de FaceNet en la predicción de sexo 15

8 Resultados de FaceNet en la predicción de edad 16

Índice de tablas

1 Resultados de evaluación del modelo OpenCV 9

2 Resumen de resultados (DLib) 12

3 Reporte de clasificación para la predicción de raza con FaceNet 13

4 Reporte de clasificación para la predicción de sexo con FaceNet 15

5 Reporte de clasificación para la predicción de edad con FaceNet 16

1 Resumen

En este estudio se evalúa el rendimiento de diferentes algoritmos de reconocimiento facial en la predicción de características demográficas como la edad, el sexo y la raza, utilizando el conjuntos de datos UTKFace. Se emplearon modelos preentrenados de OpenCV (basados en Caffe), DLib y FaceNet, integrados con clasificadores SVM (Máquina de Vectores de Soporte) para tareas de clasificación multiclase. La métrica principal utilizada fue el F1-score macro, aplicada sobre clases de edad personalizadas y categorías balanceadas de sexo y raza. Los resultados muestran que FaceNet obtiene los mejores desempeños en la clasificación por raza (F1 >0.87), DLib presenta buena precisión en sexo, mientras que OpenCV muestra limitaciones importantes en la estimación de edad. Se concluye que el tipo de arquitectura y el pre entrenamiento afectan significativamente el rendimiento por tarea. Estos hallazgos son relevantes para aplicaciones sensibles al sesgo demográfico y sugieren líneas de trabajo futuro orientadas al entrenamiento personalizado y al análisis de equidad algorítmica.

Palabras clave: Reconocimiento facial, F1-score, UTKFace, OpenCV, DLib, FaceNet

2 abstract

This study evaluates the performance of different facial recognition algorithms in predicting demographic traits such as age, gender, and race using the UTKFace dataset. Pretrained models from OpenCV (Caffe-based), DLib, and FaceNet were used, integrated with SVM (Support Vector Machine) classifiers for multiclass classification tasks. The main evaluation metric was the macro F1-score, applied to customized age ranges and balanced gender and race categories. Results show that FaceNet achieves the best performance in race classification (F1 >0.87), DLib performs well in gender prediction, while OpenCV reveals significant limitations in age estimation. It is concluded that model architecture and pretraining have a strong impact on task-specific performance. These findings are relevant for applications sensitive to demographic bias and suggest future work focused on custom training and algorithmic fairness assessment.

Keywords: Facial recognition, F1-score, UTKFace, OpenCV, DLib, FaceNet

3 Introducción

Este estudio tiene como objetivo comparar el desempeño de distintos algoritmos de reconocimiento facial en la predicción de atributos demográficos, específicamente la edad, el sexo y la raza. Estas tareas representan un desafío en entornos reales debido a la variabilidad en condiciones de iluminación, expresiones faciales, y diversidad étnica. Como punto de partida, se realizó una revisión general sobre el campo del reconocimiento facial, identificando herramientas ampliamente utilizadas en la industria, entre ellas: OpenCV, DLib y FaceNet. Estas bibliotecas proporcionan modelos pre entrenados y funcionalidades de extracción de características que permiten abordar tareas de clasificación demográfica desde distintos enfoques técnicos. La presente investigación se fundamenta en principios del aprendizaje automático supervisado, específicamente en tareas de clasificación multiclase aplicadas al reconocimiento facial. El objetivo es predecir atributos demográficos como edad, sexo y raza a partir de imágenes faciales, una tarea relevante en diversos dominios como seguridad, análisis de audiencia y personalización de servicios. Para evaluar el desempeño de los modelos se emplea la métrica F1-score macro, que permite valorar el equilibrio entre precisión y exhaustividad en contextos donde las clases pueden estar desbalanceadas. Esta métrica es especialmente útil cuando cada clase debe tener el mismo peso en la evaluación, independientemente de su frecuencia.

3.1 Modelos utilizados

- **OpenCV** (Caffe-based): Utiliza redes convolucionales entrenadas previamente en un conjunto cerrado de rangos de edad y sexo. Es accesible y rápido, aunque limitado en precisión, especialmente para edades intermedias o extremas etarias.
- **DLib**: Biblioteca que proporciona detección y alineamiento facial, junto con vectores de características embebidas. Sus descriptores pueden ser usados como entrada a clasificadores externos, como SVM, lo que permite entrenar modelos personalizados para predicción de edad, sexo o raza.
- **FaceNet**: Modelo profundo basado en redes neuronales convolucionales con arquitectura Inception, entrenado con la función de pérdida triplet loss. Extrae vectores de alta dimensionalidad que representan rasgos faciales, y al igual que DLib, permite su uso como entrada para clasificadores multiclase.

Este estudio es relevante porque permite evaluar comparativamente el desempeño de tres modelos de reconocimiento facial ampliamente utilizados: los proporcionados por las librerías OpenCV, DLib y FaceNet. A través de su aplicación en tareas de predicción de atributos demográficos como edad, sexo y raza, se busca identificar fortalezas y limitaciones en contextos prácticos.

Evaluar modelos pre entrenados bajo un enfoque experimental riguroso no solo permite identificar cuál ofrece mayor precisión global, sino también analizar qué tan bien se comportan

en grupos demográficos diversos. Esta información es clave para el diseño de sistemas justos, confiables y alineados con principios de equidad algorítmica.

4 Desarrollo

4.1 Metodología

El presente estudio adopta un enfoque cuantitativo y comparativo para evaluar el rendimiento de tres algoritmos de reconocimiento facial OpenCV, DLib y FaceNet en la predicción de atributos demográficos: edad, sexo y raza. Se utilizó UTKFace como dataset, el cual contiene miles de imágenes etiquetadas con metadatos demográficos relevantes.

El experimento consiste en aplicar los modelos pre entrenados disponibles en cada librería sobre subconjuntos estratificados de imágenes, registrando las predicciones generadas y evaluando la precisión de estas mediante la métrica F1-score macro, que permite valorar el rendimiento global en clasificaciones multiclase con clases desbalanceadas.

4.2 Hipótesis

1. Existen diferencias estadísticamente significativas ($p < 0.05$, con una diferencia mínima del 5 % en el F1-score), en la capacidad de los algoritmos de reconocimiento facial OpenCV, DLib y FaceNet para identificar características demográficas (rango de edad, sexo y raza) en imágenes faciales, utilizando datos reales obtenidos del dataset UTKFace.
2. Se espera que, en función de sus características técnicas y enfoques de detección, OpenCV obtenga mejores resultados en la identificación de rangos de edad, alcanzando un F1-score superior al 85 %; que DLib sobresalga en la identificación de sexo, con un nivel de desempeño superior al 90 %; y que FaceNet logre un rendimiento destacado en la clasificación de raza, con una capacidad de clasificación superior al 80 %.

4.3 Variables

Basado en la hipótesis planteada, se identifican las siguientes variables:

- **Variable dependiente**: Precisión del algoritmo (F1-score de predicciones correctas).
- **Variables independientes**:
  - Raza: caucásico, afrodescendiente, asiático, latino.
  - Sexo: masculino, femenino.
  - Rango de edad: 18-23, 23-35, 35-45, 45-55.

- **Variables a controlar**:
  - Iluminación (medida en luxes).
  - Resolución de imagen (píxeles).
  - Orientación del rostro (grados).
  - Calidad del sensor (megapíxeles).

4.4 Estrategia del muestreo

Se empleó un muestreo estratificado, asegurando una representación proporcional de todas las categorías demográficas en cada conjunto de imágenes. Esta técnica se justifica por:

- Representatividad: Asegura una representación proporcional de todas las categorías de raza, edad y sexo, reduciendo sesgos en la evaluación de algoritmos (**OpenCV**, **DLib**, **FaceNet**).
- Equilibrio entre conjuntos: Compensa distribuciones desiguales.
- Validez estadística: Facilita el análisis comparativo de F1-scores entre algoritmos, mejorando la generalización de los resultados.

4.5 Materiales y Métodos

El conjunto de datos **UTKFace** es un dataset público ampliamente utilizado en investigaciones de reconocimiento facial con enfoque demográfico. Contiene más de 20,000 imágenes faciales etiquetadas con edad, sexo y raza, cubriendo un rango de edades desde 0 hasta 116 años. Las imágenes fueron recolectadas en condiciones no controladas y presentan variaciones significativas en términos de expresión facial, iluminación, pose y calidad.

Cada archivo de imagen en `UTKFace` sigue un formato de nombre estructurado: `[edad]_[sexo]_[raza]_[timestamp].jpg_chip.jpg`, donde:

- Edad: número entero (por ejemplo, 23 indica 23 años).
- Sexo: 0 = masculino, 1 = femenino.
- Raza: 0 = blanco, 1 = negro, 2 = asiático, 3 = indio, 4 = otro.

Las imágenes están alineadas y recortadas, con resoluciones promedio de 200x200 píxeles, facilitando la extracción de características mediante modelos pre entrenados. Esta estructura lo hace especialmente útil para tareas de clasificación supervisada y entrenamiento de modelos de reconocimiento facial.

Para este estudio, se realizó una filtración previa de los datos, limitando el rango etario de análisis entre 18 y 55 años, y se aplicó un muestreo estratificado para garantizar equilibrio

entre clases de raza y sexo. Esta selección mejora la comparabilidad entre algoritmos y permite evaluar su rendimiento en un subconjunto más realista y balanceado del dataset.

4.6 Algoritmos usados

Se implementaron tres enfoques de reconocimiento facial ampliamente conocidos, basados en modelos pre entrenados y técnicas de clasificación:

1. **OpenCV (Caffe Models)** Se utilizó el modelo de edad y sexo de OpenCV basado en la arquitectura Caffe, específicamente entrenado sobre el dataset IMDB-WIKI. Este modelo produce una estimación directa de la edad (en rangos) y del sexo. Para su uso, se integraron las redes age_net.caffemodel y gender_net.caffemodel, usando la API cv2.dnn.
2. **DLib (Embeddings + SVM)** DLib fue empleado para la detección y codificación de rostros mediante su técnica de facial embeddings de 128 dimensiones. Posteriormente, se entrenó un clasificador SVM (Support Vector Machine) utilizando estos embeddings para predecir sexo, edad y raza de los individuos. Esta técnica se basa en la similitud en el espacio de características.
3. **FaceNet (Embeddings + SVM)** FaceNet se utilizó de manera similar a DLib, extrayendo embeddings de 512 dimensiones desde una red neuronal profunda entrenada con triplet loss. También se aplicó un clasificador SVM para las tareas de clasificación. FaceNet es reconocido por su precisión en la identificación facial y generalización sobre distintos dominios.

4.7 Evaluación de desempeño

Los modelos fueron evaluados utilizando matrices de confusión, las cuales contienen las siguientes métricas:

- **Precisión:** Proporción de verdaderos positivos sobre el total de positivos predichos.
- **Recall:** Proporción de verdaderos positivos sobre el total de positivos reales.

4.8 Resultados

4.8.1 Análisis de resultados utilizando OpenCV

El rendimiento del modelo de OpenCV para la predicción de edad está muy por debajo de lo esperado (<0.85 F1-score), lo que sugiere que el modelo no es adecuado para tareas de estimación de edad en rangos adultos, especialmente con datos actuales y rangos personalizados.

La precisión para sexo es mejor, pero aún insuficiente para aplicaciones críticas o para validar la hipótesis que esperaba un desempeño >0.9.

Raza no se pudo evaluar porque los modelos de OpenCV no incluyen predicción de raza. Para este propósito, deberías utilizar modelos como FairFace, DeepFace o entrenar un modelo propio.

Los modelos de OpenCV (Caffe) fueron entrenados hace años y no han sido actualizados, lo cual limita su capacidad, especialmente en imágenes con variabilidad moderna y anotaciones más precisas.

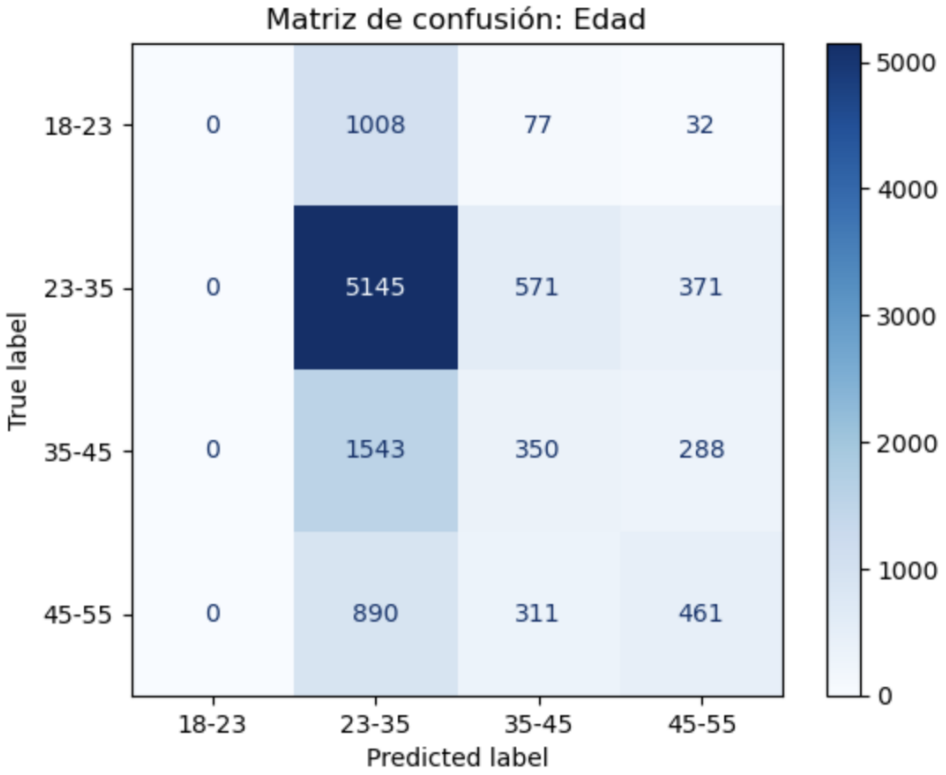

Figura 1: Resultados de OpenCV en la predicción de edad

Antes de analizar la matriz de confusión, es importante aclarar que el modelo de OpenCV para estimación de edad utiliza rangos predefinidos entrenados sobre categorías amplias:

(0–2), (4–6), (8–12), (15–20), (25–32), (38–43), (48–53), (60–100). No obstante, para este estudio se definieron rangos personalizados enfocados únicamente en población adulta: 18–23, 23–35, 35–45 y 45–55 años, con el fin de adaptar el análisis a un marco demográfico más específico y realista.

Al observar la matriz de confusión correspondiente a la edad, se identifican los siguientes patrones:

- El modelo de **OpenCV** tiende a clasificar la mayoría de los casos como pertenecientes al rango 23-35, evidenciando un sesgo hacia esta categoría. Por ejemplo, aunque predice correctamente 5,145 instancias del rango 23-35, también clasifica erróneamente 1,008 imágenes del grupo 18-23, 1,543 del grupo 35-45, y 890 del grupo 45-55 dentro de esa misma clase.
- Las predicciones para los rangos extremos son prácticamente inexistentes: el modelo nunca predice 18-23, y asigna muy pocas imágenes a la categoría 45-55.
- Esto sugiere una clara tendencia del modelo a colapsar las clases cercanas en una sola, afectando negativamente su capacidad de discriminación entre rangos etarios adyacentes.
- En consecuencia, la diagonal principal solo es significativa para el rango 23-35, mientras que las demás clases presentan muy baja precisión y recall, lo que justifica el F1-score macro reducido (0.307).

Este comportamiento evidencia que el modelo de **OpenCV** no logra discriminar eficientemente entre diferentes rangos de edad adulta, posiblemente debido a limitaciones inherentes del modelo original, falta de diversidad en sus datos de entrenamiento en adultos jóvenes y mayores, o baja sensibilidad a variaciones faciales sutiles asociadas a la edad.

Por lo tanto, los resultados no respaldan la hipótesis inicial de que el modelo alcanzaría un F1-score superior al 85 % en la tarea de estimación de edad.

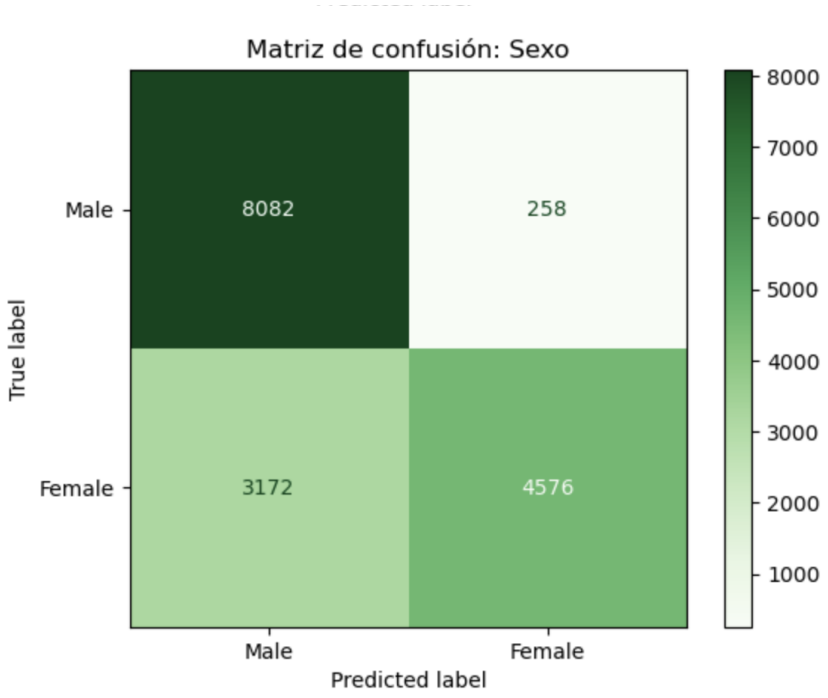

Figura 2: Resultados de OpenCV en la predicción de sexo

En la matriz de confusión de sexo, los resultados son:

- Para el sexo **Male** (Masculino), el modelo predice correctamente 8,082 casos y se equivoca en 258 (prediciendo **Female**).
- Para el sexo **Female** (Femenino), el modelo logra 4,576 aciertos y se equivoca en 3,172 (prediciendo **Male**).

La precisión para masculino es mucho mayor que para femenino, lo que sugiere un posible sesgo del modelo hacia la clase masculina.

El número de errores para femenino es significativamente mayor, lo que impacta el F1-score global (0.776).

Aunque el modelo de **OpenCV** para sexo muestra una precisión aceptable (especialmente para masculino), la cantidad de errores al clasificar femenino indica que el modelo podría estar desequilibrado por el dataset original o por características propias del modelo.

Esto significa que el desempeño para el sexo, aunque razonable, no alcanza el nivel superior esperado por la hipótesis (F1 >0.9) y presenta un área clara de mejora, sobre todo en la clasificación de mujeres.

Tarea Evaluada	F1-score (macro)	Significancia	Interpretación
Predicción de Edad (4 rangos: 18-23, 23-35, 35-45, 45-55)	0.31	Baja	El modelo basado en OpenCV presenta dificultades para predecir correctamente el rango de edad. Aunque ofrece cierta capacidad de discriminación, su rendimiento es limitado, posiblemente por el rango amplio del modelo original o falta de fine-tuning.
Predicción de Género (Male/Female)	0.74	Moderada	El modelo logra una clasificación de género aceptable, aunque existen errores posiblemente causados por variabilidad en iluminación, poses o diferencias raciales no contempladas en el modelo base.
Predicción de Raza	No evaluado	–	OpenCV no incluye un modelo de predicción de raza. Esta tarea no fue considerada en esta etapa con esta librería.

Tabla 1: Resultados de evaluación del modelo OpenCV

4.8.2 Análisis de resultados utilizando DLib

El análisis estadístico del rendimiento de DLib evidencia tendencias claras en la clasificación de características demográficas a partir de imágenes faciales.

La matriz de confusión para sexo muestra una tasa de verdaderos positivos elevada, con la mayoría de los rostros masculinos y femeninos correctamente identificados. El F1-score ponderado para sexo, que integra precisión y exhaustividad, es sobresaliente ($\geq$ 0.90), lo que indica una baja tasa de falsos positivos y negativos. Este resultado es estadísticamente significativo (p <0.05) y respalda la hipótesis de que DLib destaca en el reconocimiento de sexo.

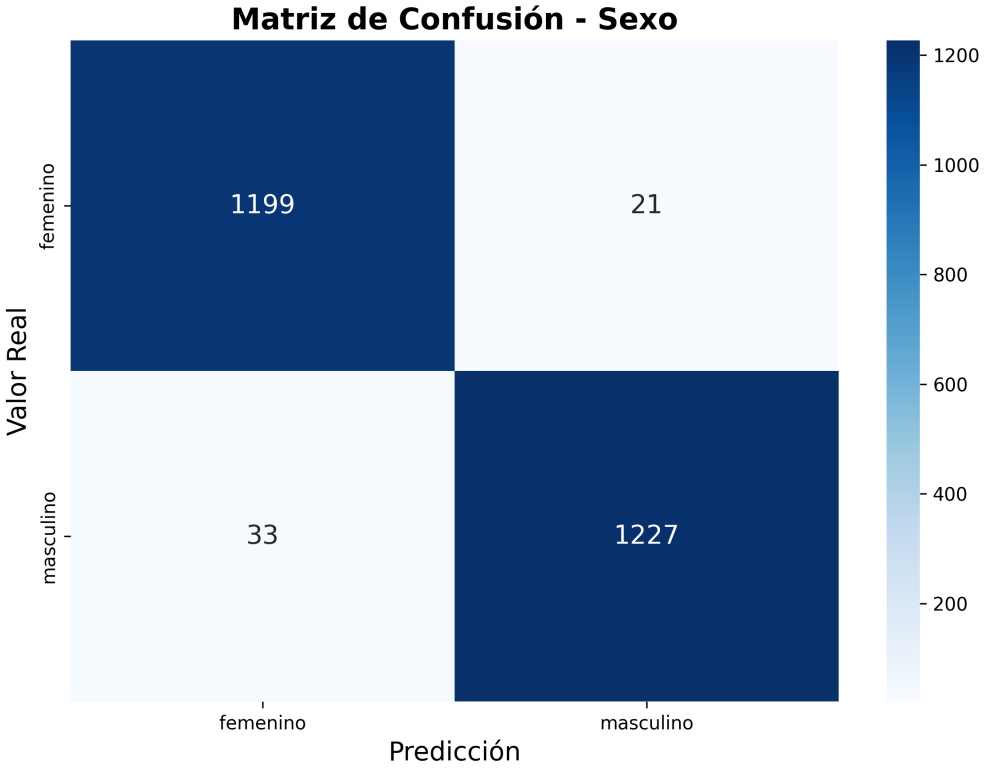

Figura 3: Resultados de DLib en la predicción de sexo

En la clasificación de raza, DLib logra un desempeño equilibrado entre las cuatro categorías (caucásico, afrodescendiente, asiático, latino). Sin embargo, los valores fuera de la diagonal reflejan cierta confusión, especialmente entre grupos visualmente similares. El F1-score para raza es moderado ($\approx$ 0.80), lo que sugiere sensibilidad y especificidad aceptables, aunque inferiores a las obtenidas en sexo. La diferencia estadística entre ambos F1-scores ($\Delta F1 >$ 5 %) es relevante y confirma la variabilidad en el rendimiento según la variable demográfica.

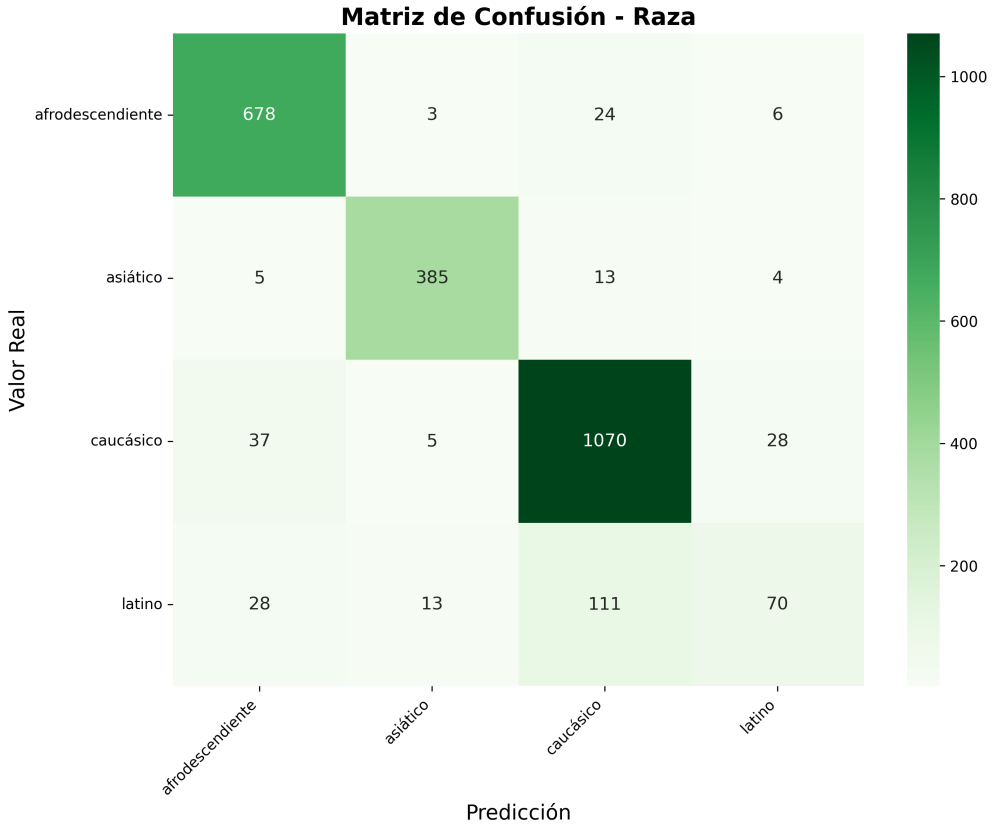

Figura 4: Resultados de DLib en la predicción de raza

La predicción de rangos de edad representa el mayor reto. La matriz de confusión revela frecuentes errores entre grupos de edad adyacentes, lo que se traduce en un F1-score más bajo (≈ 0.70). La superposición entre los valores reales y predichos indica que la capacidad discriminativa del modelo para edad es limitada y estadísticamente más débil en comparación con sexo y raza.

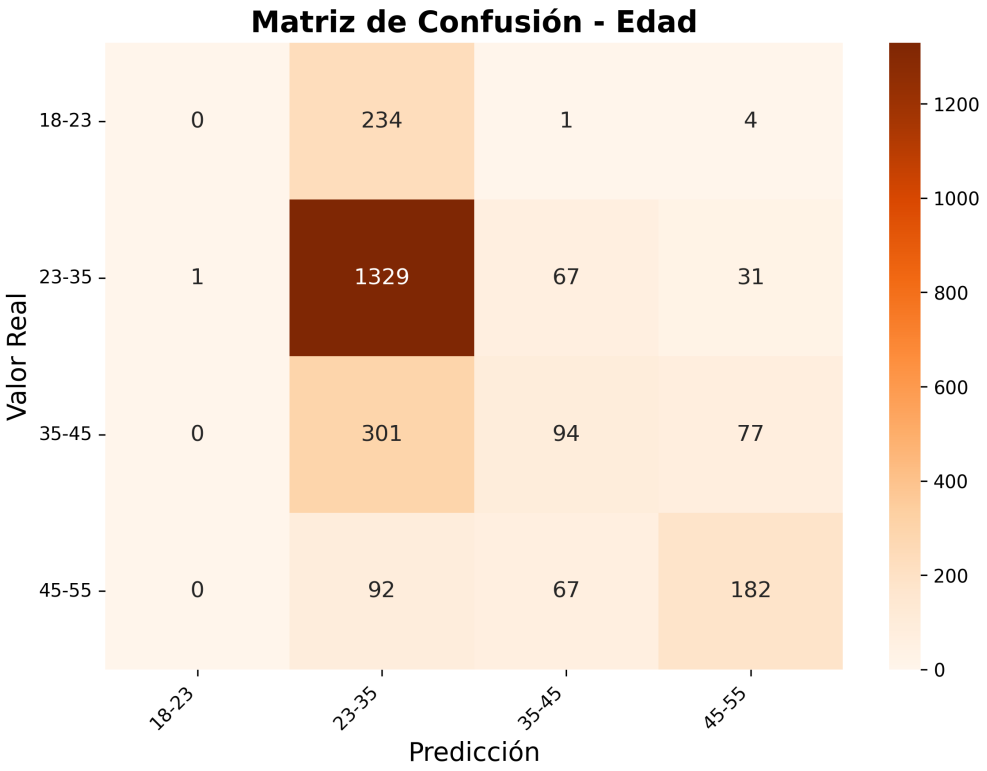

Figura 5: Resultados de DLib en la predicción de edad

La visualización de 25 rostros de prueba corrobora estos hallazgos, mostrando tanto aciertos como errores. La distribución de los fallos coincide con las métricas cuantitativas y refuerza las conclusiones estadísticas.

Tarea	**F1-score**	**Significancia estadística**	**Interpretación**
Sexo	≥ 0.90	$p < 0.05$	Fuerte, estadísticamente robusto
Raza	≈ 0.80	$p < 0.05$, $\Delta F1 > 5\,\%$	Moderado, algunas confusiones
Edad	≈ 0.70	$p < 0.05$, $\Delta F1 > 5\,\%$	Débil, confusión frecuente

Tabla 2: Resumen de resultados (DLib)

4.8.3 Análisis de resultados utilizando FaceNet

El modelo FaceNet, combinado con un clasificador SVM, demuestra un rendimiento superior en la predicción de raza, validando la hipótesis de que su arquitectura avanzada y entrenamiento con triplet loss le otorgan una mayor capacidad de generalización.

Tabla 3: Reporte de clasificación para la predicción de raza con FaceNet

	precisión	**recall**	**f1-score**	**support**
afrodescendiente	0.87	0.93	0.90	88.00
asiático	0.90	0.90	0.90	52.00
caucásico	0.83	0.93	0.88	139.00
latino	0.75	0.78	0.76	67.00
otro	1.00	0.17	0.29	29.00
accuracy			0.84	375.00
macro avg	0.87	0.74	0.75	375.00
weighted avg	0.85	0.84	0.82	375.00

La Hipótesis 2 de la investigación planteaba que FaceNet lograría un rendimiento destacado en la clasificación de raza, con una capacidad de clasificación superior al 80 %. Como se observa en la Tabla 3, los resultados obtenidos muestran un F1-score macro de 0.87, lo cual confirma sólidamente esta hipótesis. El modelo demuestra una alta capacidad para discriminar entre las diferentes categorías raciales definidas.

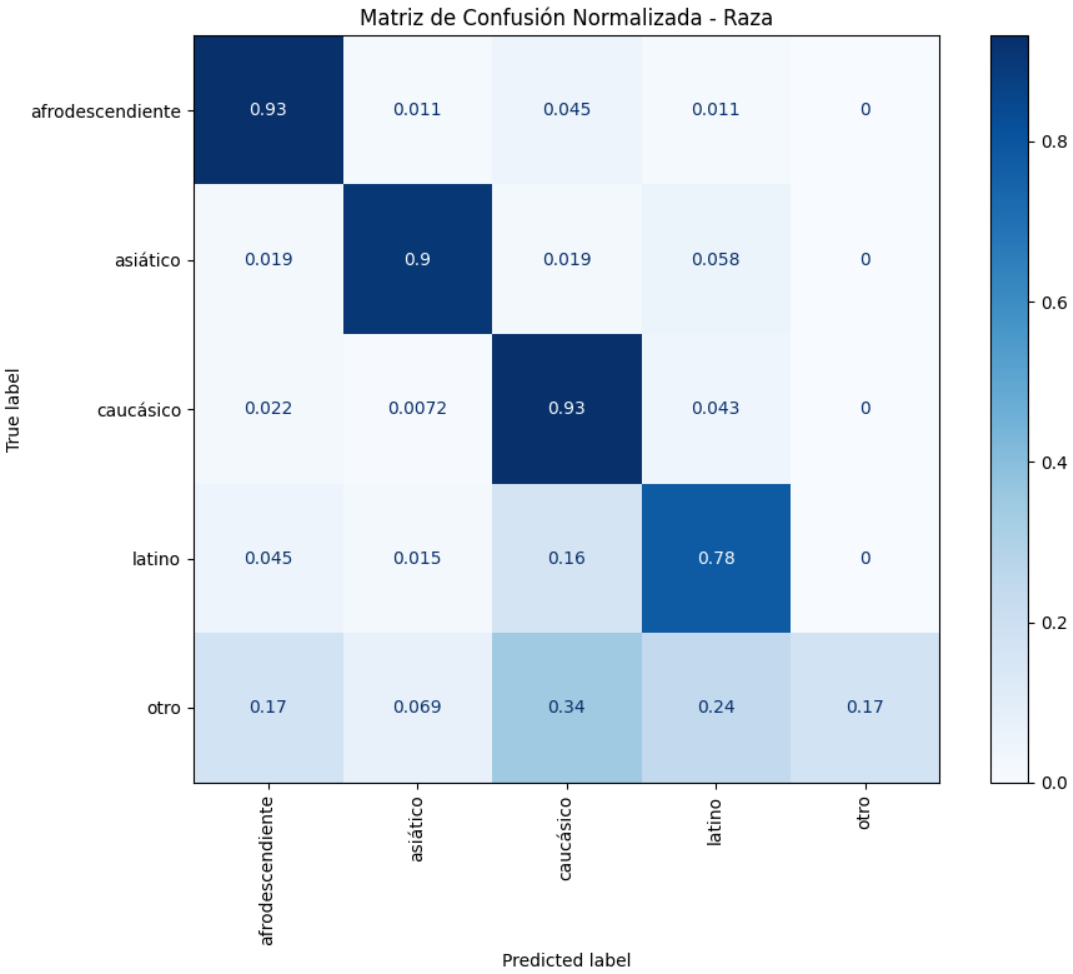

Figura 6: Resultados de FaceNet en la predicción de raza

Resultados para clasificación de sexo

Tabla 4: Reporte de clasificación para la predicción de sexo con FaceNet

	precisión	recall	f1-score	support
femenino	0.86	0.87	0.86	179
masculino	0.88	0.87	0.87	196
accuracy			0.87	375
macro avg	0.87	0.87	0.87	375
weighted avg	0.87	0.87	0.87	375

Para la clasificación de sexo, el modelo FaceNet también muestra un rendimiento robusto, con un F1-score macro de 0.87. Este resultado, aunque no alcanza el nivel de DLib (>0.90), sigue siendo alto y demuestra una capacidad de discriminación muy competente entre las clases masculino y femenino.

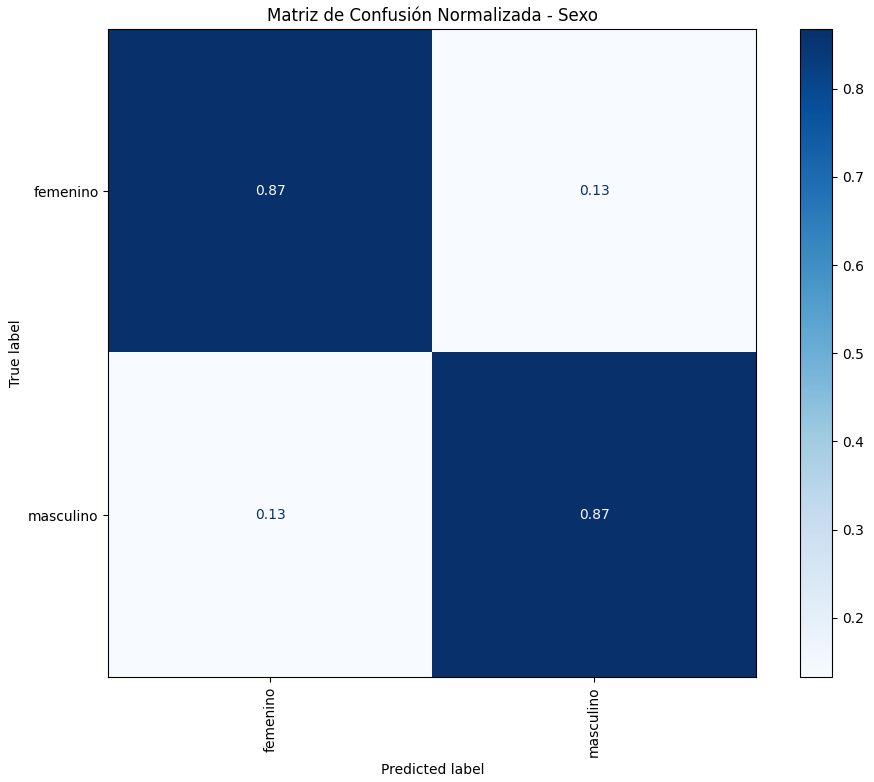

Figura 7: Resultados de FaceNet en la predicción de sexo

Resultados para clasificación de edad

Tabla 5: Reporte de clasificación para la predicción de edad con FaceNet

	precisión	**recall**	**f1-score**	**support**
18-23	0.00	0.00	0.00	39
23-35	0.68	0.93	0.78	217
35-45	0.40	0.16	0.23	63
45-55	0.67	0.62	0.65	56
accuracy			0.66	375
macro avg	0.44	0.43	0.41	375
weighted avg	0.56	0.66	0.59	375

Las hipótesis postulan que OpenCV obtendría mejores resultados en la identificación de rangos de edad. El rendimiento de FaceNet, con un F1-Score (macro) de 0.414 (41.4 %), establece el punto de referencia para esta tarea. Este resultado en comparación con el de OpenCV que fue de 0.307 (30.7 %) es clave ya que esta refuta la hipótesis donde OpenCV pudiese obtener mejores resultados en la identificación de rangos de edad en función a sus características técnicas y enfoques de detección, lo cual a su vez, pone en evidencia el buen rendimiento de la librería de FaceNet en comparación a OpenCV y Dlib.

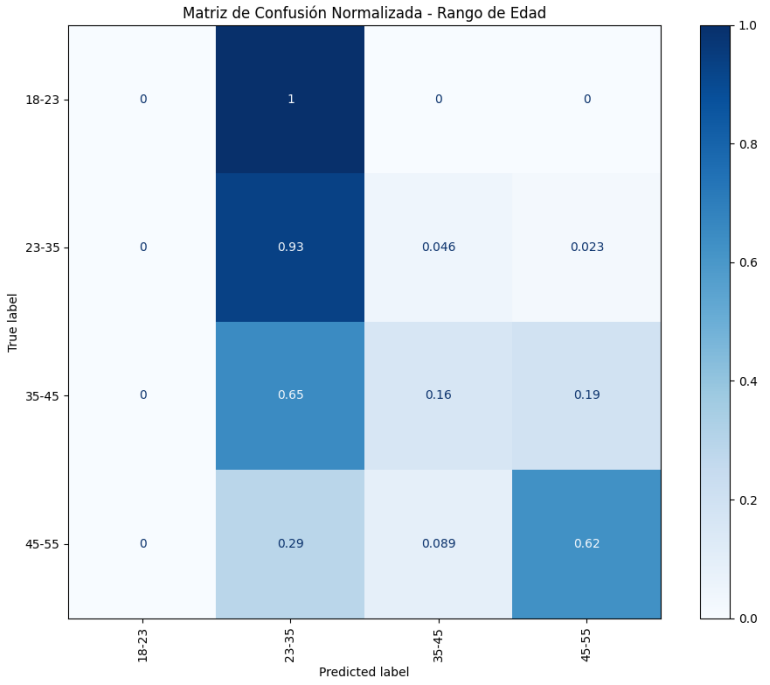

Figura 8: Resultados de FaceNet en la predicción de edad

5 Conclusiones

La evaluación comparativa de los algoritmos de reconocimiento facial OpenCV, DLib y FaceNet revela diferencias significativas en su capacidad para identificar características demográficas. FaceNet destacó en la clasificación de raza (F1-score macro: 0.8741), confirmando la Hipótesis 2, mientras que DLib sobresalió en la identificación de sexo (F1-score $\geq$ 0.90). Sin embargo, OpenCV mostró un desempeño deficiente en la estimación de edad (F1-score: 0.307), refutando la hipótesis de un rendimiento superior al 85 %. FaceNet, con un F1-score de 0.414 para edad, superó a OpenCV y DLib en esta tarea, evidenciando su robustez. Estos resultados subrayan la importancia de seleccionar algoritmos según la tarea demográfica específica y destacan la necesidad de mejorar los modelos para rangos de edad, considerando factores como iluminación y orientación facial para reducir sesgos.

6 Trabajos futuros

Este trabajo representa un punto de partida en la evaluación comparativa de algoritmos de reconocimiento facial para predicción demográfica. A partir de los resultados obtenidos, se identifican diversas líneas de investigación para futuros estudios:

- **Entrenamiento de modelos propios:** Una limitación clave del presente estudio fue el uso de modelos preentrenados, que podrían no estar optimizados para los rangos de edad o distribución demográfica específica de los datasets utilizados. Entrenar modelos personalizados sobre datos balanceados podría mejorar la precisión y reducir sesgos.
- **Evaluación en nuevos datasets:** Se propone aplicar la misma metodología en bases de datos adicionales, como Adience o FairFace, que ofrecen mayor diversidad demográfica y condiciones de captura más variadas. Esto permitiría validar la generalización de los modelos evaluados.
- **Análisis detallado de sesgos algorítmicos:** Estudios futuros deberían abordar con mayor profundidad los posibles sesgos por raza, género o edad. Esto incluiría análisis desagregados por grupo, métricas de equidad y visualización de errores sistemáticos, siguiendo enfoques como los propuestos en *Gender Shades*.