Recibido: 30/4/2023 / Aceptado: 22/5/2023
doi: https://doi.org/10.26439/interfases2023.n017.6361

Comparativa entre RESNET-50, VGG-16, Vision Transformer y Swin Transformer para el reconocimiento facial con oclusión de una mascarilla

Brenda Xiomara Tafur Acenjo
20172692@aloe.ulima.edu.pe
https://orcid.org/0000-0001-8022-3260
Universidad de Lima, Perú

Martin Alexis Tello Pariona
20163654@aloe.ulima.edu.pe
https://orcid.org/0009-0005-0933-2890
Universidad de Lima, Perú

Edwin Jhonatan Escobedo Cárdenas
eescobed@ulima.edu.pe
https://orcid.org/0000-0003-2034-513X
Universidad de Lima, Perú

RESUMEN

En la búsqueda de soluciones sin contacto físico en espacios cerrados para la verificación de identidad en el contexto de la pandemia por el SARS-CoV-2, el reconocimiento facial ha tomado relevancia. Uno de los retos en este ámbito es la oclusión por mascarilla, ya que oculta más del 50 % del rostro. La presente investigación evaluó cuatro modelos preentrenados por aprendizaje por transferencia: VGG-16, RESNET-50, Vision Transformer (ViT) y Swin Transformer, los cuales se entrenaron en sus capas superiores con un conjunto de datos propio. Para el entrenamiento sin mascarilla, se obtuvo un *accuracy* de 24 % (RESNET-50), 25 % (VGG-16), 96 % (ViT) y 91 % (Swin). En cambio, con mascarilla se obtuvo un *accuracy* de 32 % (RESNET-50), 53 % (VGG-16), 87 % (ViT) y 61 % (Swin). Estos porcentajes de *testing accuracy* indican que las arquitecturas más modernas como los *transformers* arrojan mejores resultados en el reconocimiento con mascarilla que las CNN (VGG-16 y RESNET-50). El aporte de la investigación recae en la experimentación con dos tipos de arquitecturas: CNN y *transformers*, así como en la creación del conjunto de datos público que se comparte a la comunidad

científica. Este trabajo robustece el estado del arte de la visión computacional en el reconocimiento facial por oclusión de una mascarilla, ya que ilustra con experimentos la variación del *accuracy* con distintos escenarios y arquitecturas.

PALABRAS CLAVE: reconocimiento facial, RESNET-50, VGG-16, Vision Transformer, Swin Transformer

COMPARATIVE BETWEEN RESNET-50, VGG-16, VISION TRANSFORMER AND SWIN TRANSFORMER FOR FACIAL RECOGNITION WITH MASK OCCLUSION

ABSTRACT

Face recognition has become relevant in the search for non-physical contact solutions in enclosed spaces for identity verification in the context of the SARS-CoV-2 pandemic. One of the challenges of face recognition is mask occlusion which hides more than 50 % of the face. This research evaluated four models pre-trained by transfer learning: VGG-16, RESNET-50, Vision Transformer (ViT), and Swin Transformer, trained on their upper layers with a proprietary dataset. The analysis obtained an accuracy of 24 % (RESNET-50), 25 % (VGG-16), 96 % (ViT), and 91 % (Swin) with unmasked subjects. While with a mask, accuracy was 32 % (RESNET-50), 53 % (VGG-16), 87 % (ViT), and 61 % (Swin). These percentages indicate that modern architectures such as the Transformers perform better in mask recognition than the CNNs (VGG-16 and RESNET-50). The contribution of the research lies in the experimentation with two types of architectures: CNNs and Transformers, as well as the creation of the public dataset shared with the scientific community. This work strengthens the state of the art of computer vision in face recognition by mask occlusion by illustrating with experiments the variation of accuracy with different scenarios and architectures.

KEYWORDS: face recognition, RESNET-50, VGG-16, Vision Transformer, Swin Transformer

1. INTRODUCCIÓN

La propagación del virus SARS-CoV-2 ha causado una crisis sin precedentes a nivel mundial, que a su vez ha provocado varios problemas en los sectores de salud, economía, transporte, seguridad, etcétera. La rápida proliferación de este virus y el surgimiento de nuevas variantes ha tenido como consecuencia altos índices de contagio, porque su propagación ocurre por contacto físico y también por superficies contaminadas. Ante esta situación, se elevó la demanda de espacios que requieren utilizar métodos de verificación biométrica libres de contacto físico, confiables y eficaces. Se ha planteado el uso del reconocimiento facial como principal medio de identificación. Sin embargo, debido a la medida impuesta del uso de mascarillas faciales, se genera la oclusión de gran parte del rostro, con lo que el *accuracy* de los modelos de reconocimiento facial se ha reducido notoriamente (Damer et al., 2020). Así, se puede notar que existe demanda por sistemas de reconocimiento facial que no supongan contacto físico y que sean lo suficientemente robustos para identificar la identidad de sujetos portando mascarillas faciales.

La oclusión es un tema que se viene investigando hace años. Los casos de oclusiones más comunes son el uso de gafas de sol, bufandas, cabello en el rostro, envejecimiento, entre otros (Sáez Trigueros et al., 2018). Sin embargo, en comparación con otros tipos de oclusiones, la oclusión facial es el menos estudiado entre todos. Actualmente, existen pocas investigaciones que se centren en la aplicación de las arquitecturas *transformers* hacia el reconocimiento facial, menos aún si le sumamos la oclusión facial (Tran et al., 2022). El uso más común que se viene dando al reconocimiento facial se encuentra en la videovigilancia o el distanciamiento social debido al COVID-19 (Meena & Meena, 2022).

La influencia de la oclusión facial en soluciones tecnológicas relacionadas con el reconocimiento facial se produce por la extracción de puntos clave, como los que están en la "zona T" (ojos, nariz y boca), para identificar a una persona; estos son un factor determinante para el éxito o fallo del algoritmo. En ambientes dinámicos con muchas interferencias externas, los modelos tradicionales para el reconocimiento facial fallarán. El uso de ventanas desplazadas hace más eficiente el procesamiento de información al limitar el análisis a pequeñas secciones que no se superponen, pero que aun así se conectan entre ellas. Las principales dificultades para el reconocimiento facial causadas por la oclusión son pérdida de rasgos, error de alineación y *aliasing* local en la imagen (Cheng & Pan, 2022).

La oclusión facial parcial sigue siendo un problema que presenta limitaciones en el reconocimiento facial, el cual empeora mucho más con el uso de una mascarilla, siendo este caso el más difícil de todos (Hariri, 2022). El uso de una mascarilla cubre el 50 % de la zona frontal del rostro; por ello, el porcentaje de efectividad es reducido y hace a estos sistemas no confiables. La empresa tecnológica líder en inteligencia artificial SenseTime

Technology reportó que la tasa de efectividad de un sistema de reconocimiento facial se puede reducir hasta en un 10 % cuando la persona tiene 50 % de la nariz expuesta al usar una mascarilla; este porcentaje puede llegar a disminuir mucho más si se tiene oculta la mitad del rostro (Wang et al., 2023). Pese a la reducción de la tasa de *accuracy* de los sistemas de reconocimiento facial, su uso se ha incrementado debido a que los sistemas biométricos de contacto físico, como la digitación de contraseñas o huellas dactilares, generan más riesgo de contagio del coronavirus SARS-CoV-2 (Hariri, 2022).

En el entrenamiento de estos sistemas, que generalmente están basados en aprendizaje profundo (DL, por sus siglas en inglés), la tasa del *accuracy* depende de la cantidad de imágenes que se tenga y puedan servir para el entrenamiento de estos modelos (Sáez Trigueros et al., 2018). Esto, sumado a la variedad de modelos de algoritmos que existen para el reconocimiento facial, conduce a que aún no se cuente con información de estudios comparativos que analicen diferentes escenarios, a fin de cuantificar la tasa de reducción en el *accuracy* ante la oclusión de una mascarilla. Además, otra limitación es la poca cantidad de los conjuntos de datos disponibles con imágenes de sujetos que usan mascarilla y sin mascarilla. A pesar de que existen conjuntos de datos de rostros enmascarados, como Real-World Fasked Face Recognition Dataset (RMFRD), este solo está disponible para la industria y la academia. Así, el público en general no cuenta con libre disponibilidad para su uso e investigación (Laxminarayanamma et al., 2021).

A partir de lo expuesto anteriormente, se ha propuesto realizar un estudio comparativo entre las arquitecturas CNN y *transformer*, ambas preentrenadas sobre el conjunto de datos ImageNet21k, con el fin de cuantificar su *testing accuracy* en distintos escenarios, ya sea con imágenes no ocluidas u ocluidas. Para esto se creó una base de datos propia compuesta por 30 clases. A continuación, el trabajo presenta una revisión de literatura, así como la exposición de los principales fundamentos teóricos de los modelos utilizados. Además, se propone una metodología y se documenta todo el proceso de experimentación, discusión y resultados. Finalmente, se mencionan las conclusiones y los trabajos futuros.

La contribución de nuestra investigación se resume en los siguientes puntos:

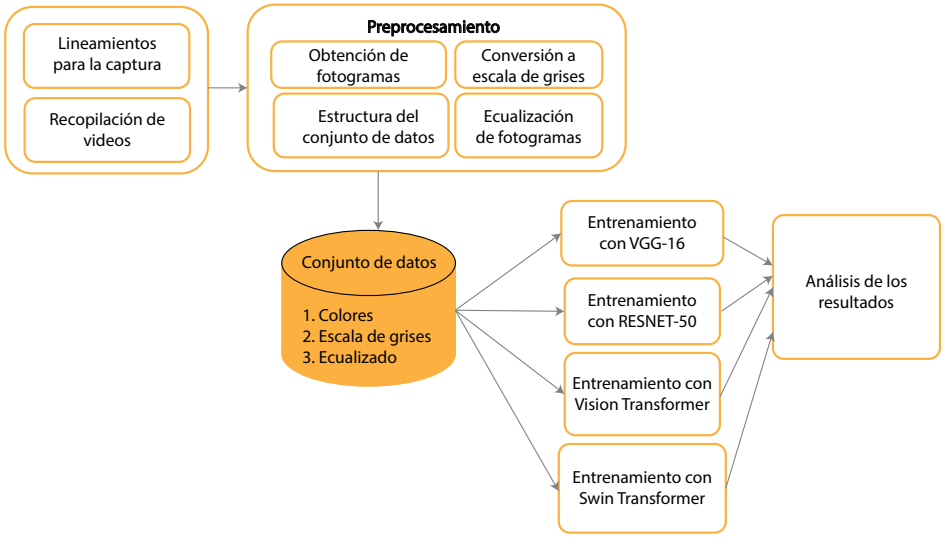
- Crear un nuevo conjunto de datos con imágenes faciales del rostro con y sin oclusión por mascarilla. Consta de 30 clases, las cuales tienen imágenes de personas desde un ángulo frontal y lateral.
- Cuantificar la reducción de la métrica *testing accuracy*, en adelante *accuracy*, en el reconocimiento facial por oclusión para las arquitecturas CNN y *transformer*.
- Aportar en la investigación de las arquitecturas *transformers* en el campo del reconocimiento facial con oclusión.

2. METODOLOGÍA

La investigación se inició con la definición de los lineamientos para la captura de videos. Una vez recolectados, comenzó la etapa del preprocesamiento de imágenes, donde se descompuso los videos en fotogramas. Luego, se dividió los tres conjuntos de datos que se tenía: imágenes a colores, a escalas de grises y ecualizadas[1]. Después de ello, se seleccionó solo el conjunto de datos con imágenes a colores para realizar el entrenamiento de dos redes neuronales convolucionales: VGG-16 y RESNET-50, un modelo de Vision Transformer y un modelo de Swin Transformer. Una vez que se obtuvieron los resultados de los modelos, se realizó un análisis comparativo. En la Figura 1 se puede apreciar cada una de las etapas.

Figura 1

Etapas de la metodología



Para la creación de la base de datos, se han utilizado tres técnicas. La primera es la descomposición en *frames.* La segunda son las redes neuronales convolucionales en cascada multitarea (MTCNN, por sus siglas en inglés), como identificador y extractor de datos. Por último, tenemos el aumento de datos (*data augmentation*, por su traducción en inglés). Estas imágenes fueron analizadas a color y, posteriormente, convertidas en un formato de escala de grises. Además, se aplicó ecualización de histogramas.

1 El conjunto de datos de personas con y sin mascarilla facial utilizado para la investigación se encuentra publicado en el Repositorio Institucional de la Universidad de Lima (https://hdl.handle.net/20.500.12724/18500).

Como primer paso, se establecieron los lineamientos para la captura de los videos:

1. Los videos serían grabados en un ambiente cerrado con luz adecuada, ya sea natural o artificial, iluminando el rostro de los voluntarios.
2. El voluntario debería mostrar las orejas en el cuadro de video.
3. El voluntario no debe portar ningún objeto que ocluya la imagen, como lentes, bufandas, entre otros.

De esta manera, se realizó la recopilación de videos de 30 voluntarios para la creación de un conjunto de datos propio. Estos videos fueron grabados con un celular Samsung Galaxy A32. En total, se obtuvieron cuatro videos de 40 segundos aproximadamente por persona, dos de ellos portando mascarilla y los otros dos sin mascarilla. Además, uno de ellos debía estar a 40 centímetros de distancia y el otro a un metro y medio. Cabe resaltar que no hay restricción acerca del tipo de mascarilla, que puede ser KN95, quirúrgica o de tela, entre otras. De igual forma, se consideraron algunas variantes en los videos, como girar hacia los lados laterales, tanto derecha como izquierda; así como realizar un movimiento circular con la cabeza, simulado con el fin de obtener todas las características tanto desde una perspectiva alta como baja. Los diferentes escenarios que se plantean para la recolección de videos tienen el fin de brindar robustez a los modelos en cuanto al reconocimiento facial (Damer et al., 2020).

Luego, se llevó a cabo la etapa de preprocesamiento. Para esto, se efectuó el tratamiento de imágenes con las redes convolucionales en cascada multitarea (MTCNN, por sus siglas en inglés), seguido del diseño de la estructura del conjunto de datos y la obtención de fotogramas. En total, se consiguió, aproximadamente, 1200 imágenes por video por persona (Yanai & Kawano, 2015). En los pasos posteriores, se usó el método cvtColor, el cual se implementó en una librería de visión computacional OpenCV para obtener las imágenes a escala de grises. Además, mediante la aplicación de ecualización de fotogramas a las imágenes en escala de grises, se consiguió que las imágenes tuvieran un mayor contraste. Finalmente, se obtuvieron tres tipos de conjuntos de datos: a color, a escala de grises y con ajuste por ecualización de histogramas. Esto último se realizó con la función "equalizeHist" de OpenCV. Originalmente, los modelos preentrenados se estrenarán solo con las imágenes a color. Adicional a ello, se utilizarán las imágenes ecualizadas y a escala de grises para experimentar un escenario específico y contrastar cómo varía contra las imágenes a color.

Con la información recolectada, se prepararon los hiperparámetros para la ejecución del modelo preentrenado VGG-16, tales como un *size* de imagen de 224 × 224 píxeles, 100 épocas y un *batch size* igual a 128. Debido a que el modelo ya estaba previamente entrenado, se aplicó aprendizaje por transferencia. Lo mismo se hizo con el modelo RESNET-50, usando como hiperparámetros un tamaño de imagen de

128 × 128 píxeles, 100 épocas y un *batch size* de 128 (Mandal et al., 2021). Además, se realizó el entrenamiento de un Vision Transformer, que es un modelo distinto de las redes neuronales convolucionales entrenadas anteriormente (Zhong & Deng, 2021). Igualmente, se entrenó un Swin Transformer, que es un transformador que utiliza ventanas desplazadas, las cuales mejoran la eficiencia al restringir el cálculo de autoatención a ventanas locales que no se solapan, al mismo tiempo que permite la interconexión entre estas ventanas (Liu et al., 2021).

3. RESULTADOS

Para la presente investigación, se entrenaron cuatro modelos preentrenados: VGG-16, RESNET-50, Vision Transformer y Swin Transformer. Se analizó los escenarios (a) con mascarilla y (b) sin mascarilla. Asimismo, el entrenamiento se realizó sobre el conjunto de datos completo, es decir, con imágenes de 30 sujetos.

3.1 Entrenamiento de modelos CNN

En el entrenamiento de VGG-16 sin mascarilla, se obtuvo un 25 % de *testing accuracy.* Además, la matriz de confusión muestra que el modelo solo se equivoca en algunos casos específicos, como con "Sujeto 8" y "Sujeto 27" (Figura 2). En el modelo con mascarilla, arrojó un *testing accuracy* de 53 %. Además, la matriz de confusión muestra que el modelo reconoce de forma consistente a la mayoría de los sujetos (Figura 3).

En el entrenamiento de RESNET-50 sin mascarilla, se obtuvo un 24 % de *testing accuracy.* Este valor es similar al alcanzado con el otro modelo de red neuronal convolucional, VGG-16. Además, la matriz de confusión muestra que el modelo confunde a la mayoría de los sujetos con "Sujeto 4" (Figura 4). En el modelo con mascarilla, arrojó un *testing accuracy* de 32 %. La matriz de confusión señala que el modelo suele predecir a los sujetos de manera incorrecta con "Sujeto 16" y "Sujeto 27" (Figura 5). A pesar de ello, se puede ver que logra una mejor consistencia de reconocimiento en comparación con el modelo sin mascarilla.

Figura 2

Matriz de confusión de VGG-16 con 30 sujetos sin mascarilla

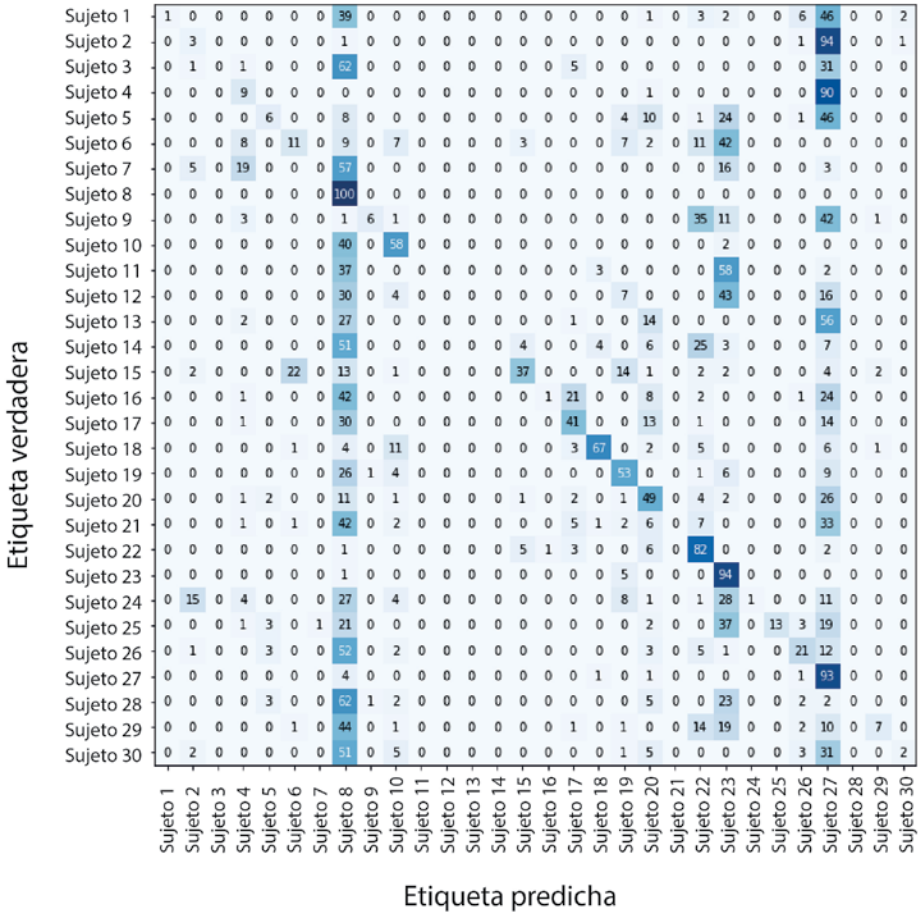


Figura 3

Matriz de confusión de VGG-16 con 30 sujetos con mascarilla

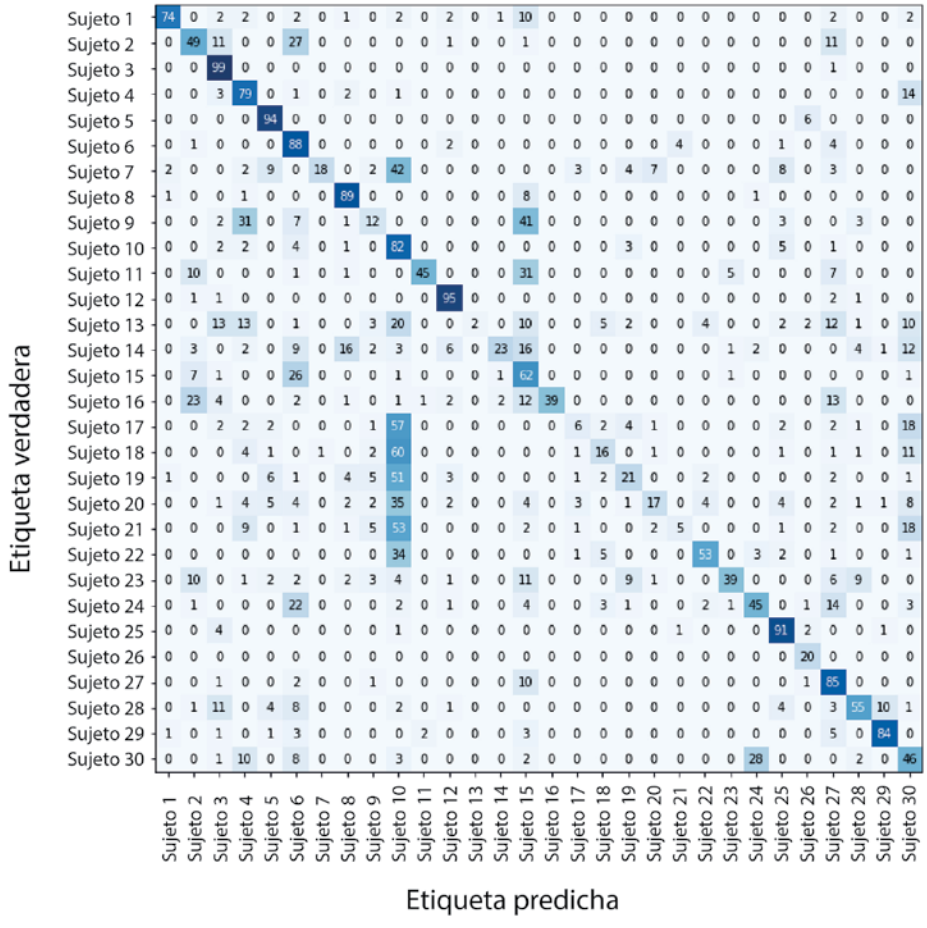


Figura 4

Matriz de confusión de RESNET-50 con 30 sujetos sin mascarilla

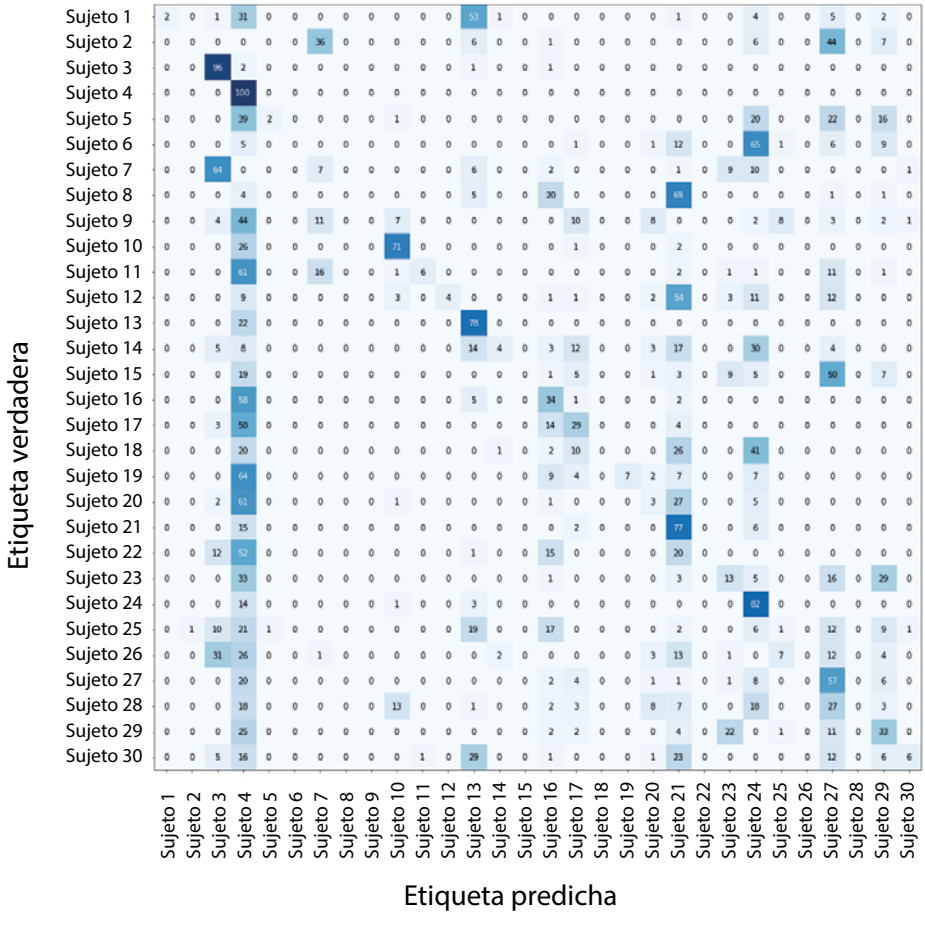
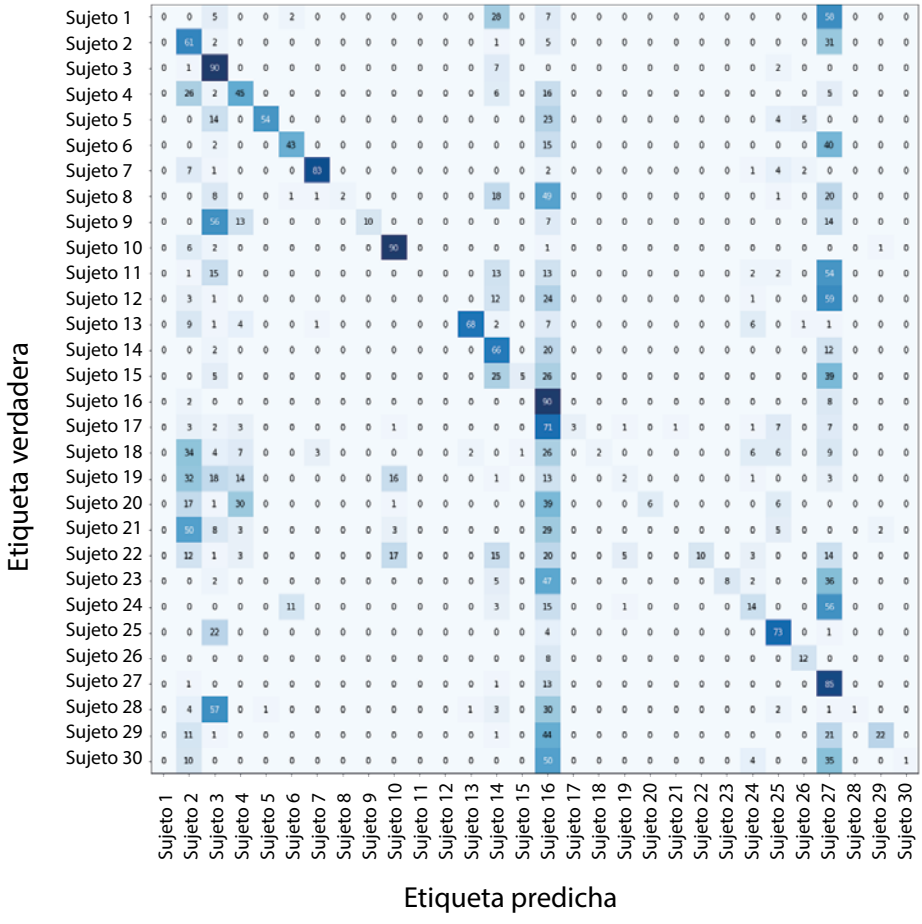


Figura 5

Matriz de confusión de RESNET-50 con 30 sujetos con mascarilla



3.2 Entrenamiento de modelos *transformers*

El segundo y último entrenamiento del Vision Transformer sin mascarilla arrojó 96 % de *testing accuracy*. Además, la matriz de confusión muestra que el modelo reconoce de manera correcta a la mayoría de los sujetos (Figura 6). En el modelo con mascarilla, arrojó un *testing accuracy* de 87 %. Asimismo, la matriz de confusión señala que el modelo se equivoca solo en dos casos específicos con "Sujeto 2" y con "Sujeto 5" (Figura 7). A pesar de mostrar errores, se puede apreciar que los casos de confusión no superan las 30 imágenes de 100.

Figura 6

Matriz de confusión de Vision Transformer con 30 sujetos sin mascarilla

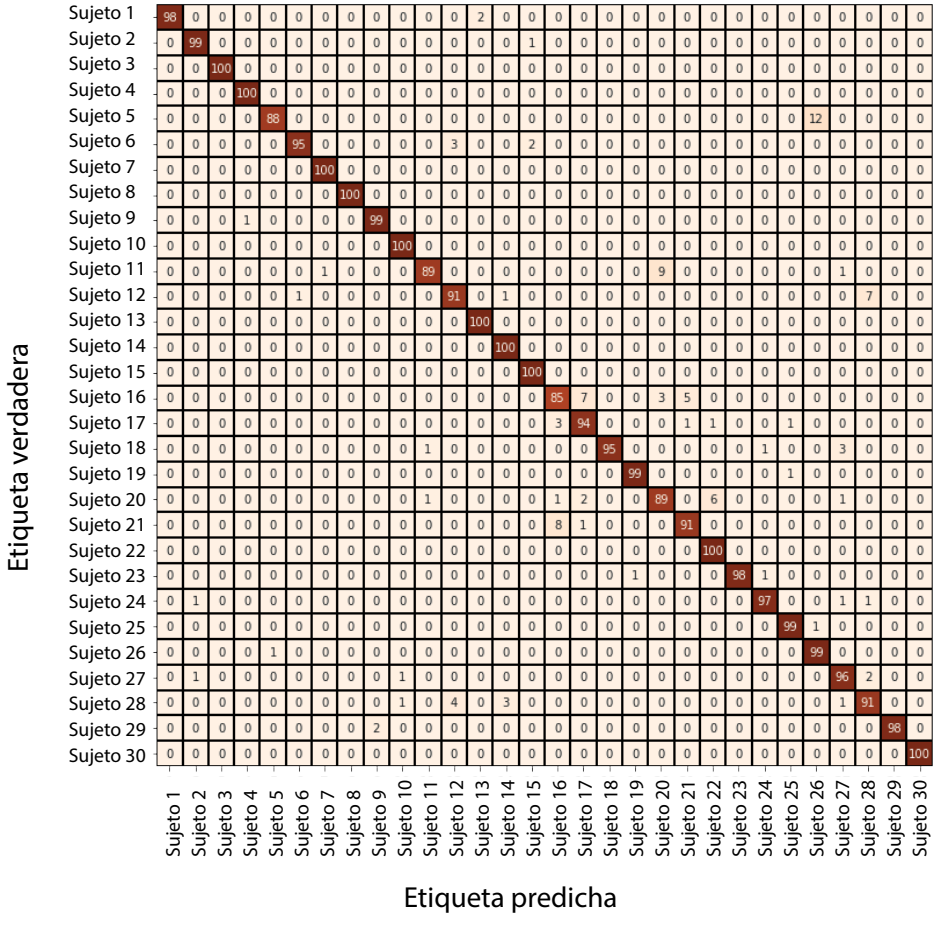
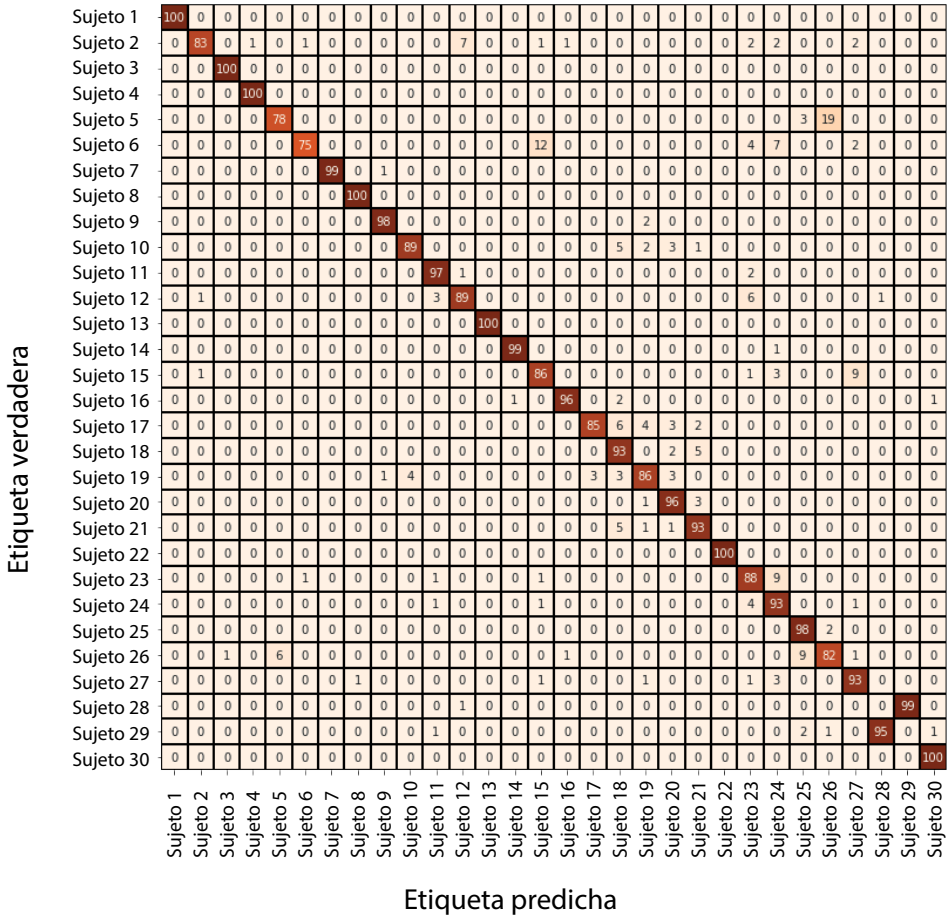


Figura 7

Matriz de confusión de Vision Transformer con 30 sujetos con mascarilla



El Swin Transformer se entrenó bajo los mismos parámetros del Vision Transformer, y en el modelo sin mascarilla, arrojó 91 % de *testing accuracy.* La matriz de confusión muestra que el modelo suele confundir a la mayoría de los sujetos con "Sujeto 29" y "Sujeto 30" (Figura 8). En cambio, en el modelo con mascarilla, arrojó un *testing accuracy* de 87 %. Además, la matriz de confusión señala que el modelo reconoce a la mayoría de los sujetos, salvo algunas confusiones con el "Sujeto 29" (Figura 9).

Figura 8

Matriz de confusión de Swin Transformer con 30 sujetos sin mascarilla

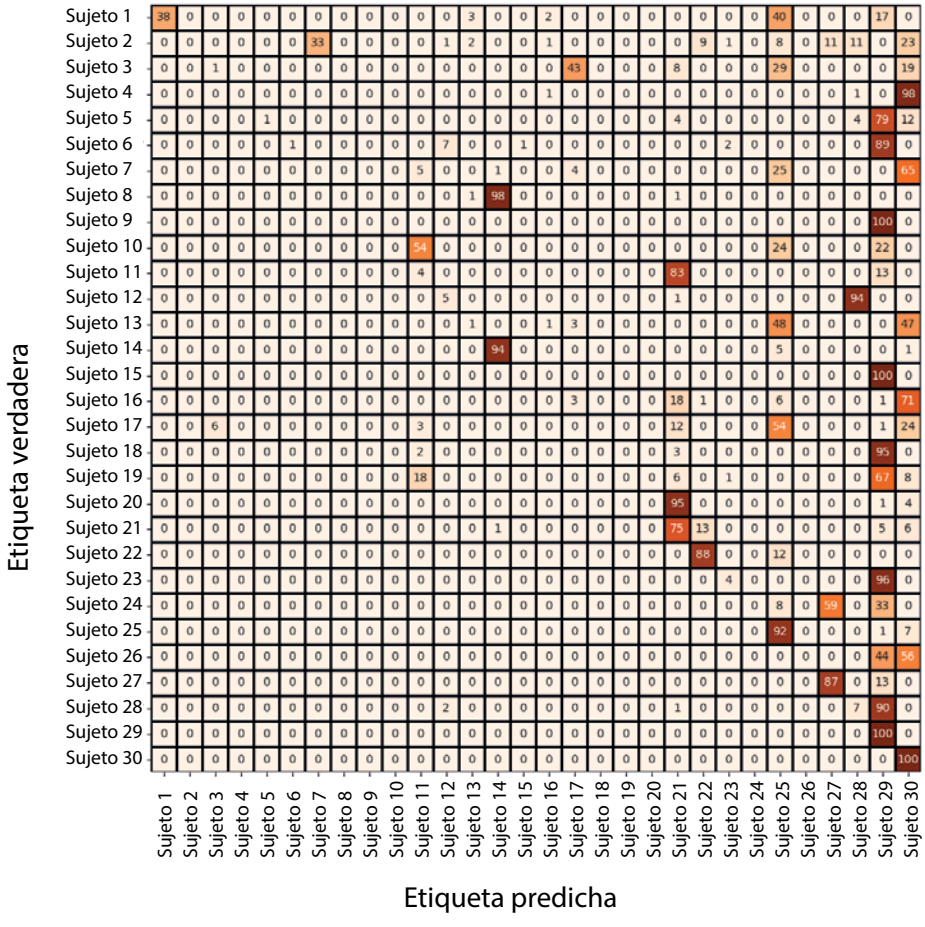
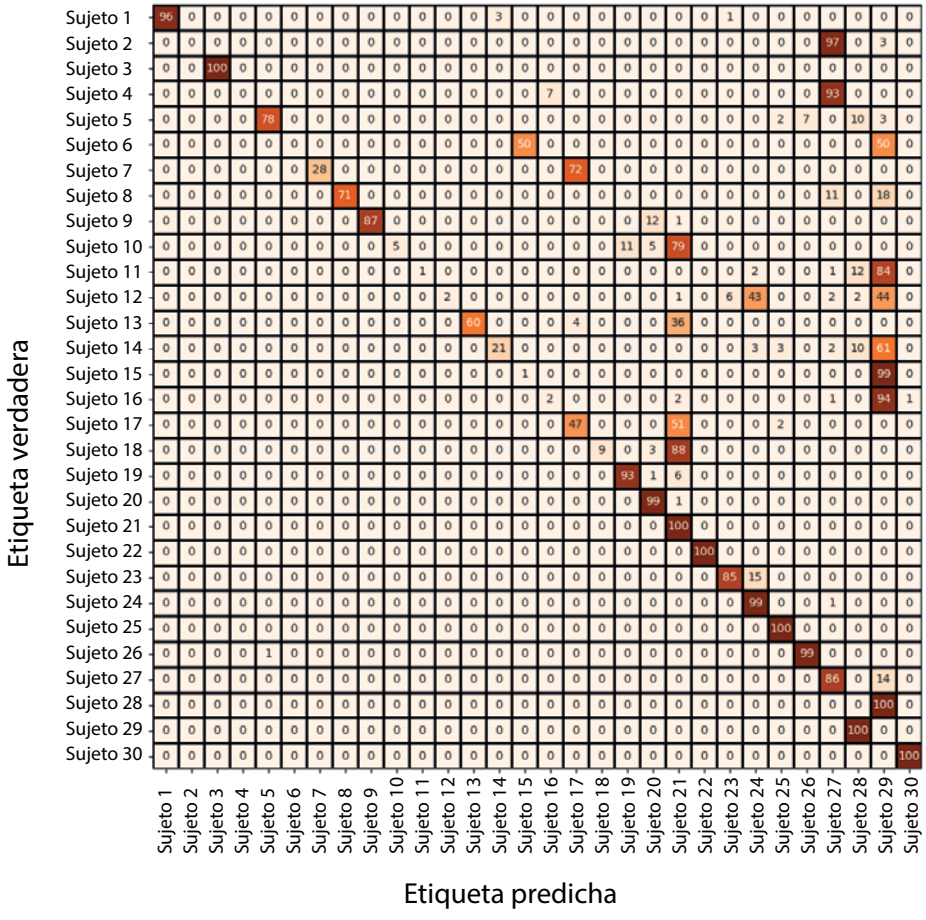


Figura 9

Matriz de confusión de Swin Transformer con 30 sujetos con mascarilla



3.2.1 Entrenamiento de modelos *transformers* en la base segmentada

En los modelos de Vision Transformer y Swin Transformer, podemos apreciar que la matriz de confusión del segundo arroja una mayor cantidad de sujetos identificados correctamente (Figura 11). En cambio, el modelo Vision Transformer tiende a confundir a varios sujetos con el "Sujeto 21" y solo reconoce bien a menos de la mitad (Figura 10).

Figura 10

Matriz de confusión de modelos transformers con 30 sujetos en base segmentada: Vision Transformer

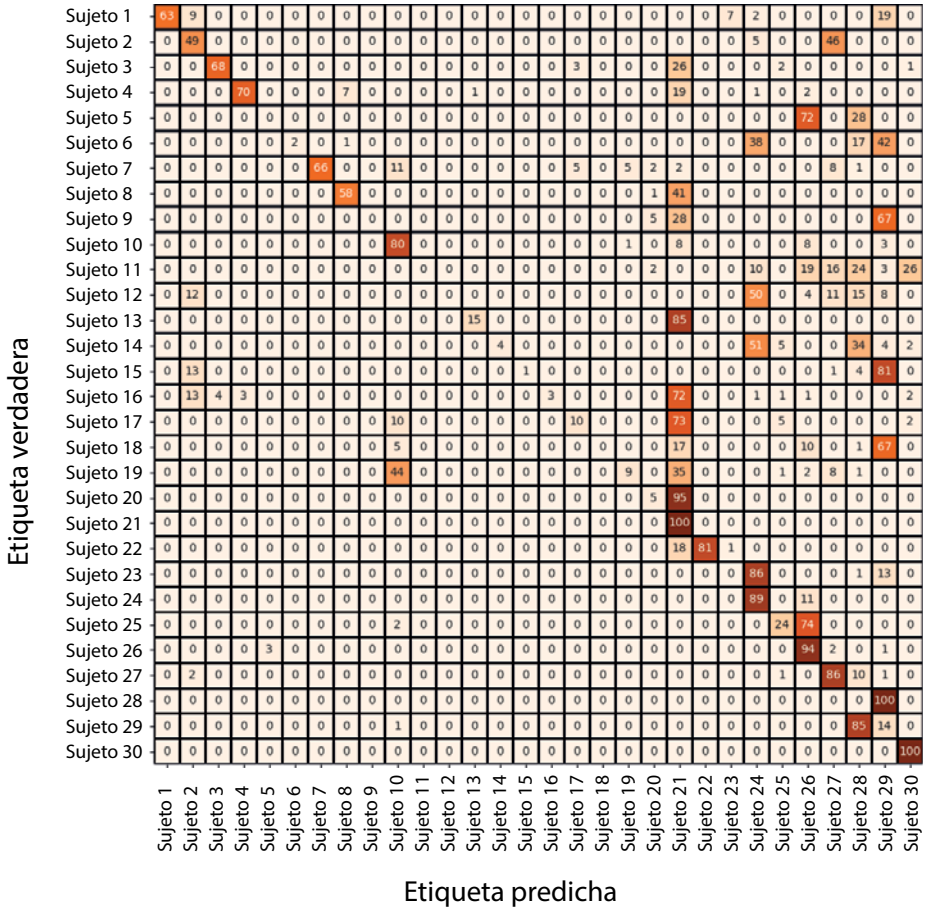
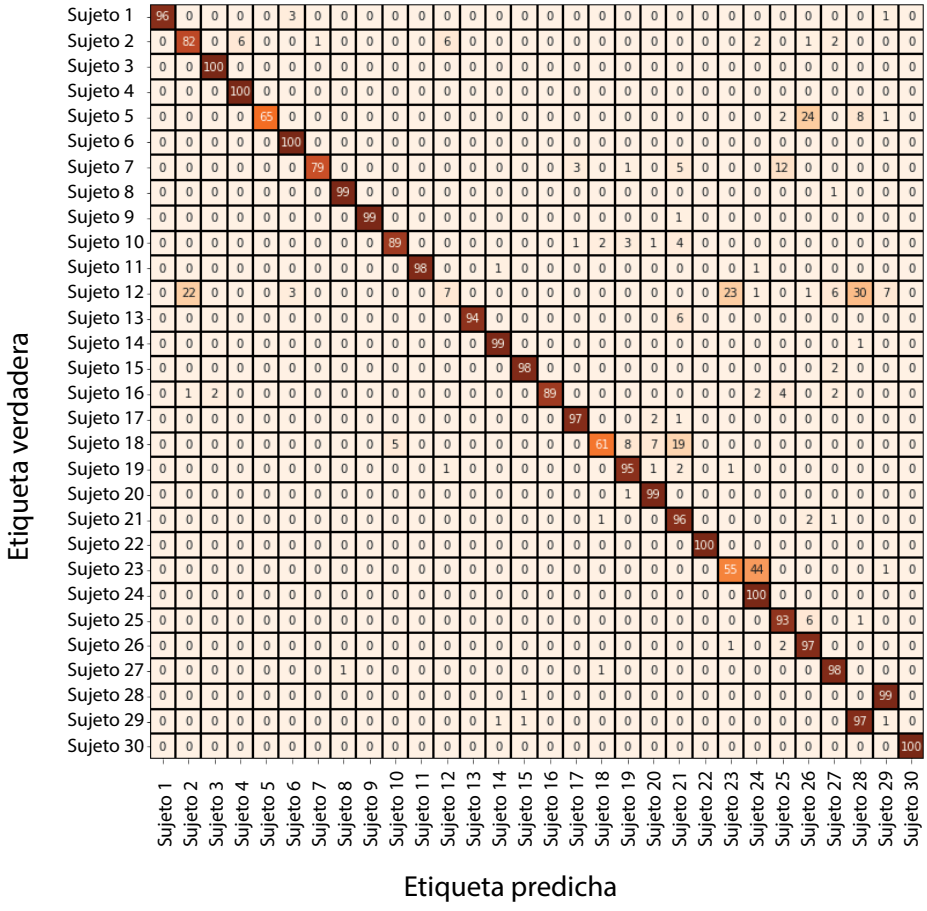


Figura 11

Matriz de confusión de modelos transformers con 30 sujetos en base segmentada: Swin Transformer



4. DISCUSIÓN

La creación de la base de datos ha permitido recopilar imágenes de 30 sujetos con y sin mascarilla en diferentes poses, que fueron de movimientos laterales y circulares. Ello se realizó a fin de ofrecer mayor robustez a los modelos de reconocimiento facial. Las imágenes obtenidas se dividieron en tres secciones: a colores, a escala de grises y aplicando ecualización de histogramas. No obstante, por los recursos computacionales disponibles, solo se usó la base de datos a color para el entrenamiento de los modelos, con un total de 30 sujetos. Luego, se realizó el entrenamiento de los cuatro modelos preentrenados: RESNET-50, VGG-16, Vision Transformer y Swin Transformer.

RESNET-50, al ser una arquitectura más profunda, extrae características más distintivas entre los sujetos y, al trabajar con bloques residuales, no se pierde información de las capas anteriores (Wu et al., 2019). Por ello, a diferencia de VGG-16, agrega *skip connections* y permite compartir desde la capa 1 todo el mapa de características hasta la capa 10. Así, la información que se ha ido perdiendo desde la capa 1 a la 9 se recupera en la capa 10. Los resultados mostraron que el modelo VGG-16 sin mascarilla confundía a la mayoría de los sujetos con "Sujeto 8" (Figura 12.d) y "Sujeto 27" (Figura 12.b). En cuanto a RESNET-50, cuando se evaluó sin mascarilla, la mayoría de las predicciones eran erróneas y las clasificaba como "Sujeto 4" (Figura 12.e). Sin embargo, cuando se evaluó este mismo modelo con mascarilla, reconocía a la mayoría de los sujetos como "Sujeto 16" (Figura 12.f). Además, al igual que en el caso anterior, el "Sujeto 16" comparte características como el color de piel y de cabello. Asimismo, las orejas no se aprecian y mantiene la misma dirección de mirada.

Figura 12

Sujetos de la base de datos propia

Al aplicar MTCNN, se ha reducido o quitado en su totalidad características como orejas, cabello y forma del rostro en las imágenes con las que se probaron los modelos. Esto ha disminuido el *accuracy* tanto de los modelos VGG-16 como de RESNET-50. En el caso de Vision Transformer, en el primer entrenamiento se obtienen bajos resultados de *accuracy*, casi como los de las redes neuronales convoluciones. Sin embargo, al realizar un segundo entrenamiento y descongelando el 40 % de la última capa, se consiguieron resultados que casi llegaron al 100 %: en el caso del modelo sin mascarilla, al 96 %; y con mascarilla, a un 87 % de *accuracy*. En el modelo de Swin Transformer, se pudo apreciar que los resultados obtenidos tienen un comportamiento similar al de Vision Transformer en cuanto a *accuracy.* Sin embargo, en la matriz de confusión de Swin Transformer, se puede observar una mayor consistencia en el modelo con mascarilla. Por el contrario, el modelo sin mascarilla suele confundir a la mayoría de los sujetos con "Sujeto 29" (Figura 13.b) y "Sujeto 30" (Figura 13.c).

Figura 13

Sujetos de la base de datos propia, con segmentación de la parte no ocluida

El descongelamiento se realizó a fin de realizar aprendizaje por transferencia del modelo preentrenado en bases de datos como ImageNet21K y CIFAR-10. Con ello se pudo ver que, al utilizar modelos preentrenados en bases de datos gigantes y con arquitecturas más complejas como *transformers*, mejoró en nivel de *accuracy* en el reconocimiento facial de nuestra base de datos propia. Adicionalmente, cabe destacar que este tipo de arquitecturas *transformers* está orientado a tener un entrenamiento más enfocado, lo cual permite que haya una iteración entre todas las partes participantes. Así, se aprende a tener una mejor distinción en la clasificación de los sujetos. Finalmente, es preciso mencionar que los modelos *transformers* requieren de mayores recursos computacionales frente a las arquitecturas de redes neuronales convolucionales.

5. CONCLUSIONES

La investigación se realizó en un contexto pospandemia por SARS-CoV-2, donde el uso de mascarillas se mantiene en la población. Además, el reconocimiento facial con oclusión por mascarilla para acceder a centros de estudio o trabajo se ha vuelto relevante, ya que, por medidas sanitarias, cierto sector de la población continúa usando mascarillas en espacios cerrados.

Se puede afirmar que se logró el objetivo de la investigación mediante la construcción de una base de datos propia y la comparación de arquitecturas tradicionales, como las redes neuronales convolucionales (CNN), frente a arquitecturas más modernas como los modelos *transformers*.

La base de datos creada contiene un total de 30 sujetos, lo cual nos permitió realizar los entrenamientos de las dos CNN y los dos modelos *transformers*. Esta base de datos recolecta imágenes con diversas poses de los sujetos, con lo que se agregó robustez a los modelos al momento de ser entrenados; cuenta con imágenes a color, ecualizadas y a escala de grises.

Además del entrenamiento con imágenes de rostros ocluidos y no ocluidos por mascarillas, se analizó un experimento en las arquitecturas *transformers* (ViT y Swin)

con imágenes segmentadas en las cuales se fuerza al algoritmo a enfocarse en extraer características específicas de la parte no ocluida del rostro, y de esta forma ignorar distintos factores que no sean relevantes para la identificación del rostro.

La realización de este tipo de experimentos con diferentes escenarios y arquitecturas es enriquecedora para la academia, puesto que permite realizar el contraste de los niveles de *accuracy* arrojados en cada escenario. Se pudo observar que las arquitecturas *transformers*, al tener una arquitectura más compleja y enfocada en el detalle, logran mejores resultados que las CNN en todos los casos que han sido simulados en los entrenamientos. Como punto importante, se observa que las redes CNN tienen una caída notable en su *accuracy* al ser entrenadas en una mayor cantidad de clases, mientras que en las arquitecturas *transformers* sucede todo lo contrario, ya que mantienen un alto nivel en su porcentaje de *accuracy*.

El aporte de la investigación recae en la experimentación con dos tipos de arquitecturas: CNN y *transformer*, así como en la creación del conjunto de datos público que se comparte a la comunidad científica. De igual manera, la experimentación con un modelo *transformer* permite comparar modelos tradicionales como las CNN frente a modelos modernos que se encuentran moldeando el estado del arte de la visión computacional. Cabe mencionar que actualmente existe un déficit en investigaciones que involucren las arquitecturas *transformers* en el ámbito del reconocimiento facial y mucho menos en el aspecto de oclusión.

Las experimentaciones mostraron la mejoría que resulta al contrastar la tarea del reconocimiento facial entre estos dos tipos de arquitecturas. Los resultados de esta investigación robustecen el estado del arte de la visión computacional en el reconocimiento facial por oclusión de una mascarilla, ya que ilustran con experimentos la variación del *accuracy* en distintos escenarios y usando dos tipos de arquitecturas diferentes. Ello contribuye a que se decida con evidencia cuáles son los modelos más adecuados para realizar la tarea de reconocimiento facial cuando la población usa mascarilla en espacios cerrados.

En trabajos futuros, se podrían incluir otras arquitecturas *transformers* modernas que vayan emergiendo del estado del arte, así como realizar la experimentación con nuevas arquitecturas híbridas que surgen de unir las redes CNN con la arquitectura *transformer*. Asimismo, se puede incrementar los sujetos para nuestra base de datos original, lo que permitiría aumentar el número total de imágenes que sirvan para los entrenamientos.

APÉNDICE

El conjunto de datos utilizado en esta investigación contiene imágenes de personas con y sin mascarillas faciales, y está disponible en: https://hdl.handle.net/20.500.12724/18500

El código fuente empleado en la investigación se encuentra disponible en el siguiente enlace: https://colab.research.google.com/drive/1T8n7ib--4b7QWWgAea86ZXXABfpuL6AZ?usp=sharing

REFERENCIAS

Cheng, P., & Pan, S. (2022). Learning from face recognition under occlusion. En *2022 International Conference on Big Data, Information and Computer Network* (BDICN) (pp. 721-727). IEEE. https://doi.org/10.1109/BDICN55575.2022.00140

Damer, N., Grebe, J. H., Chen, C., Boutros, F., Kirchbuchner, F., & Kuijper, A. (2020). *The effect of wearing a mask on face recognition performance: An exploratory study*. BIOSIG 2020 - Proceedings of the 19th International Conference of the Biometrics Special Interest Group, agosto. https://dl.gi.de/server/api/core/bitstreams/c3e8ae49-dde1-4b80-ad18-3d3536b1897b/content

Hariri, W. (2022). Efficient masked face recognition method during the COVID-19 pandemic. *Signal, Image and Video Processing*, *16*(3), 605-612. https://doi.org/10.1007/s11760-021-02050-w

Laxminarayanamma, K., Deepthi, V., Ahmed, M. F., & Sowmya, G. (2021). A real time robust facial recognition model for masked face images using machine learning model. En *2021 5th International Conference on Electronics, Communication and Aerospace Technology (ICECA)* (pp. 769-774). IEEE. https://doi.org/10.1109/ICECA52323.2021.9675936

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. En *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 10012-10022). IEEE. https://doi.org/10.1109/ICCV48922.2021.00986

Mandal, B., Okeukwu, A., & Theis, Y. (2021). *Masked face recognition using RESNET-50*. arXiv:2104.08997. https://doi.org/10.48550/arXiv.2104.08997

Meena, M. K., & Meena, H. K. (2022). A literature survey of face recognition under different occlusion conditions. En *2022 IEEE Region 10 Symposium (TENSYMP)* (pp. 1-6). IEEE. https://doi.org/10.1109/TENSYMP54529.2022.9864502

Sáez Trigueros, D. S., Meng, L., & Hartnett, M. (2018). Enhancing convolutional neural networks for face recognition with occlusion maps and batch triplet loss. *Image and Vision Computing*, *79*, 99-108. https://doi.org/10.1016/j.imavis.2018.09.011

Tran, C. P., Vu, A. K. N., & Nguyen, V. T. (2022). Baby learning with vision transformer for face recognition. En *2022 International Conference on Multimedia Analysis and Pattern Recognition (MAPR)* (pp. 1-6). IEEE. https://doi.org/10.1109/MAPR56351.2022.9924795

Wang, Z., Huang, B., Wang, G., Yi, P., & Jiang, K. (2023). Masked face recognition dataset and application. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, *5*(2), 298-304. https://doi.org/10.1109/TBIOM.2023.3242085

Wu, Z., Shen, C., & Van Den Hengel, A. (2019). Wider or deeper: Revisiting the RESNET model for visual recognition. *Pattern Recognition*, *90*, 119-133. https://doi.org/10.1016/j.patcog.2019.01.006

Yanai, K., & Kawano, Y. (2015). Food image recognition using deep convolutional network with pre-training and fine-tuning. En *2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)* (pp. 1-6). IEEE. https://doi.org/10.1109/ICMEW.2015.7169816

Zhong, Y., & Deng, W. (2021). *Face transformer for recognition*. arXiv:2103.14803. https://doi.org/10.48550/arXiv.2103.14803

ANEXOS

Anexo 1. Pseudocódigo para la separación de las imágenes en *frames*

```
IMPORTAR os
IMPORTAR cv2
FUNCIÓN FRAMES(FolderDestino,RutaVideo):
    CREAR os.FolderDestino
    VARIABLE contador
    LEER cv2.RutaVideo
    MIENTRAS(TRUE):
      VARIABLE frame
      CAPTURAR frame DE RutaVideo
      SI frame == NULO:
          BREAK
      SINO:
          GUARDAR cv2.frame{contador} EN FolderDestino
      contador = contador + 1
```

Anexo 2. Pseudocódigo para la conversión de las imágenes en escala de grises

```
IMPORTAR os
IMPORTAR cv2
FUNCIÓN BLACKWHITE(FilePathContent,RutaDestino:
```

```
CREAR os.FolderDestino EN RutaDestino
VARIABLE contador
FOR filename EN FilePathContent:
    SI filename TERMINA EN ".jpg" OR ".png":
        VARIABLE imagen
        LEER cv2.filename
        imagen = filename
        CONVERTIR img EN BGR2GRAY
        GUARDAR cv2.img{contador} EN FolderDestino
        contador = contador + 1
    SINO:
        Continúa
```

Anexo 3. Pseudocódigo para la implementación del MTCNN

```
IMPORTAR MTCNN
IMPORTAR os
IMPORTAR cv2
FUNCIÓN DETECTFACES(img_path,destino_path):
    img = LEERcv2.
    #REAJUSTAR TAMAÑO DE IMAGEN
    VARIABLE ancho
    VARIABLE alto
    VARIABLE contador
    img = cv2.RESIZE(img,(ancho,alto), INTERPOLATION = cv2.INTER_
    AREA)
    #INICIALIZAR MODELO MTCNN
    VARIABLE face_coord
    face_coord = mtcnn_face_detector_model.detect_faces(img)
```

```
#EXTRAER ROSTRO
FOR x EN face_coord:
        VARIABLE face
        face = cv2.RESIZE(face,(224,224),INTERPOLATION = cv2.
        INTER_AREA)
        cv2.GUARDAR(destino_path+face{contador}.jpg)
contador = contador + 1
```