Modelos para predición de enfermedades pulmonares

Marlon Mora
Facultad de ingeniería
Universidad de San Buenaventura
Cali, Colombia
mdmorar@correo.usbcali.edu.co

Martín Correa
Facultad de ingeniería
Universidad de San Buenaventura
Cali, Colombia
mcorreav1@correo.usbcali.edu.co

***Resumen*—Este trabajo presenta un análisis comparativo del desempeño de tres arquitecturas de redes neuronales aplicadas a un conjunto de datos de radiografías de tórax. Se incluyen un modelo personalizado basado en redes neuronales convolucionales (CNN), y dos modelos de *Transfer Learning* basados en VGG16 y ResNet50. Los modelos fueron evaluados en términos de métricas relevantes como *accuracy*, *precision*, *recall* y *F1-score*. Además, se discuten las ventajas y limitaciones de cada enfoque, proporcionando una base para futuras investigaciones en el uso de inteligencia artificial en imágenes médicas.**

***Palabras clave*—Redes Neuronales Convolucionales, Transfer Learning, Clasificación de Imágenes Médicas**

I. Introducción

El diagnóstico basado en imágenes médicas ha sido un pilar fundamental en la identificación y tratamiento de enfermedades. Sin embargo, el aumento exponencial en la cantidad de datos médicos generados diariamente ha planteado desafíos significativos para los radiólogos y profesionales de la salud, incluyendo la necesidad de procesar grandes volúmenes de datos de manera precisa y eficiente [1]. En este contexto, el aprendizaje profundo (*Deep Learning*) ha emergido como una herramienta prometedora para automatizar tareas complejas, como la clasificación de imágenes médicas [2].

Entre las aplicaciones más relevantes, el análisis de radiografías de tórax ha ganado especial atención debido a su importancia en la detección de enfermedades respiratorias, incluyendo COVID-19. Estudios recientes han demostrado que el uso de modelos basados en aprendizaje profundo puede alcanzar niveles de precisión comparables, e incluso superiores, a los de especialistas humanos en tareas de clasificación de imágenes médicas [3]. La presente investigación aborda el problema de la clasificación de radiografías de tórax en cuatro categorías: imágenes normales, casos de COVID-19, opacidad pulmonar y neumonía viral. Para ello, se implementaron y compararon tres enfoques distintos: un modelo personalizado basado en redes neuronales convolucionales y dos modelos preentrenados (VGG16 y ResNet50), utilizando técnicas de *Transfer Learning*.

II. Metodología

II-A. Descripción del Dataset

El conjunto de datos utilizado en este trabajo es el *COVID-19 Radiography Dataset*, disponible en la plataforma Kaggle. Este dataset contiene un total de 21,165 imágenes de radiografías de tórax distribuidas en cuatro categorías: pulmones sanos, casos de COVID-19, opacidad pulmonar y neumonía viral. Cada imagen fue etiquetada de manera supervisada, lo que lo hace ideal para tareas de clasificación [4]. Las imágenes presentan diferentes resoluciones originales, por lo que se realizó un preprocesamiento para unificar su tamaño a 64×64 píxeles, facilitando su manejo durante el entrenamiento de los modelos.

II-B. División del Dataset

Para garantizar una evaluación adecuada del rendimiento de los modelos, el conjunto de datos se dividió en un 80 % para entrenamiento y un 20 % para validación. La división se realizó de manera estratificada, asegurando una distribución balanceada entre las cuatro categorías. Esta estrategia ayuda a evitar sesgos en los resultados y mejora la representatividad de las evaluaciones.

II-C. Preprocesamiento de los Datos

Se aplicaron técnicas de aumento de datos (*data augmentation*) para mejorar la capacidad de generalización de los modelos y reducir el riesgo de sobreajuste. Las transformaciones incluyeron rotaciones aleatorias, desplazamientos horizontales y verticales, zoom y volteo horizontal. Estas operaciones introducen variaciones en las imágenes de entrenamiento, permitiendo que los modelos aprendan características más robustas. Además, los valores de los píxeles se normalizaron al rango $[0, 1]$ mediante un escalado por $1/255$.

II-D. Modelos Utilizados

Se implementaron y compararon tres modelos diferentes, cada uno configurado con una capa de salida *softmax* para clasificar entre las cuatro categorías y un *dropout* de 0.5 para regularización. La figura 1 muestra un resúmen de los 3 modelos entrenados. Los detalles específicos de cada modelo son los siguientes:

1. **CNN Personalizada**: Este modelo fue diseñado específicamente para este problema. Su arquitectura consta de tres capas convolucionales, seguidas por dos capas de agrupamiento máximo de tamaño 2×2 y una capa de agrupamiento promedio (*average pooling*).

Característica	Modelo 1 (CNN)	Modelo 2 (VGG16)	Modelo 3 (ResNet50)
Filtros Convolucionales	32, 64, 128	Basados en VGG16	Basados en ResNet50
Capas Densas	256, 4	256, 4	256, 4
Activaciones	ReLU, Softmax	ReLU, Softmax	ReLU, Softmax
GlobalAveragePooling2D	Sí	Sí	Sí
Transfer Learning	No	Sí	Sí
Congelación de capas	No	15 capas	15 capas

Figura 1. Características de los modelos entrenados

2. **VGG16**: Un modelo preentrenado ampliamente utilizado en tareas de clasificación de imágenes. Se utilizó *Transfer Learning* congelando los últimos 15 pesos para aprovechar características previamente aprendidas, mientras que las capas superiores fueron ajustadas al problema específico.
3. **ResNet50**: Otro modelo preentrenado que emplea *skip connections* para mitigar problemas de degradación en redes profundas. También se configuró mediante *Transfer Learning*, congelando los últimos 15 pesos para ajustar sus capas superiores al conjunto de datos actual.

II-E. Entrenamiento y Métricas

El entrenamiento de los modelos se llevó a cabo utilizando el optimizador Adam con una tasa de aprendizaje inicial de $0{,}0001$, y la función de pérdida utilizada fue *categorical crossentropy*. Las métricas empleadas para la evaluación incluyeron la exactitud (*accuracy*), precisión (*precision*), exhaustividad (*recall*) y puntuación F1 (*F1-score*), proporcionando una visión integral del desempeño de cada modelo. Los modelos fueron entrenados durante 10 épocas con un tamaño de batch de 32 imágenes.

II-F. Configuraciones Adicionales

Para optimizar el proceso de entrenamiento y evitar sobreajuste, se implementaron dos *callbacks*:

1. ***Early Stopping***: Detiene el entrenamiento si el desempeño del modelo en el conjunto de validación comienza a deteriorarse.
2. ***Model Checkpoint***: Guarda automáticamente el modelo con el mejor desempeño en el conjunto de validación, asegurando que los análisis posteriores se realicen con la versión óptima del modelo.

III. Resultados

En esta sección se presentan los resultados obtenidos tras entrenar y evaluar los tres modelos propuestos para la clasificación de imágenes de radiografías de tórax en las cuatro categorías definidas. Se evaluaron métricas clave de desempeño: exactitud (*accuracy*), precisión (*precision*), exhaustividad (*recall*) y puntuación F1 (*F1-score*). Además, se generaron matrices de confusión para cada modelo, que se presentan como figuras asociadas.

III-A. Desempeño de los Modelos

Los resultados de las métricas para cada modelo se resumen en la Tabla I. El modelo **ResNet50** obtuvo el mejor desempeño general, con una exactitud de $91\,\%$, mientras que el modelo **VGG16** logró una exactitud cercana, del $88\,\%$. Por otro lado, la **CNN Personalizada** alcanzó un desempeño más limitado, con una exactitud de $76\,\%$.

En términos de precisión, el modelo **ResNet50** destacó con un valor de $90\,\%$, superando tanto a VGG16 ($89\,\%$) como a la CNN personalizada ($78\,\%$). En cuanto a la exhaustividad, ResNet50 alcanzó un $91\,\%$, lo que indica una alta capacidad para identificar correctamente las imágenes de cada categoría. Finalmente, en la puntuación F1, que balancea precisión y exhaustividad, el modelo ResNet50 mantuvo su liderazgo con $91\,\%$, mientras que VGG16 quedó ligeramente por detrás con $88\,\%$, seguido de la CNN personalizada con $74\,\%$.

Cuadro I
Resumen de Métricas de Desempeño por Modelo

Modelo	Exactitud	Precisión	Exhaustividad	F1-Score
CNN	0.76	0.78	0.73	0.74
VGG16	0.88	0.89	0.86	0.88
ResNet50	0.91	0.90	0.91	0.91

III-B. Matrices de Confusión

Las matrices de confusión correspondientes a cada modelo se presentan en las Figuras 2, 3 y 4. Estas matrices muestran la distribución de verdaderos positivos, falsos positivos, verdaderos negativos y falsos negativos para las cuatro categorías.

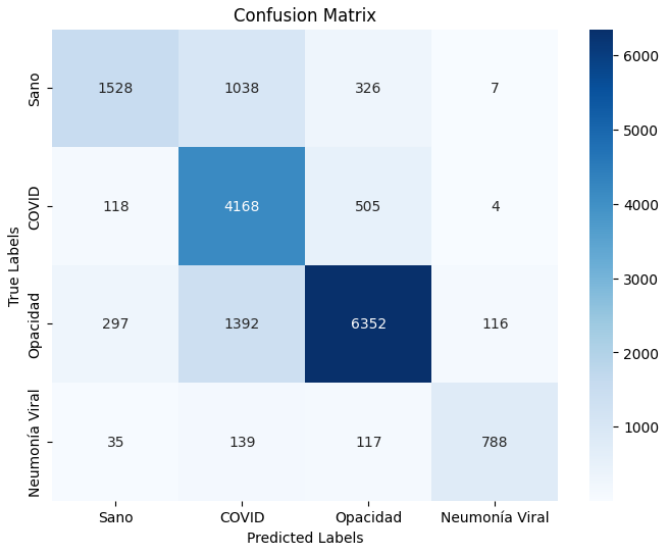

Figura 2. Matriz de confusión para el modelo CNN Personalizada.

III-C. Análisis Comparativo

Al analizar los resultados, se observa que el modelo ResNet50 presenta un equilibrio destacado entre las métricas clave, logrando los valores más altos en todas las categorías.

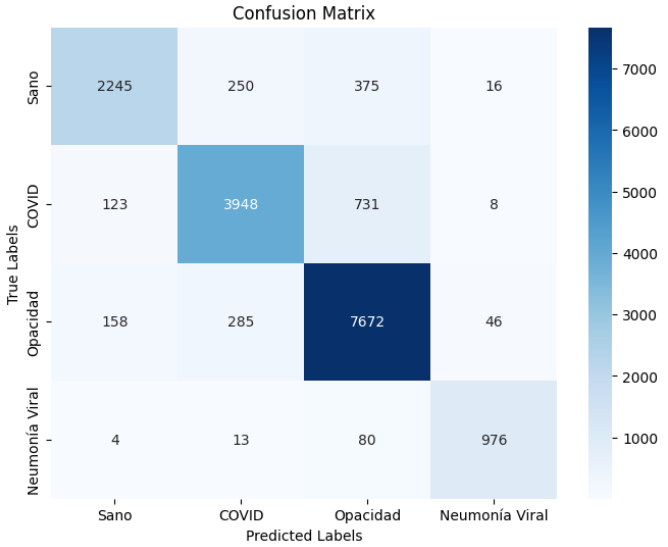

Figura 3. Matriz de confusión para el modelo VGG16.

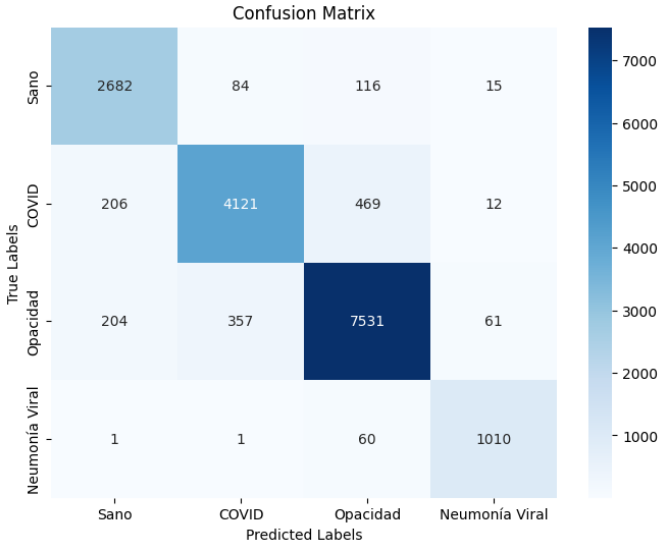

Figura 4. Matriz de confusión para el modelo ResNet50.

Esto sugiere que su arquitectura preentrenada, combinada con el uso de *skip connections*, le permitió capturar patrones más complejos en las imágenes. Por otro lado, VGG16 también mostró un desempeño competitivo, siendo más eficiente que la CNN personalizada en todos los aspectos evaluados.

El desempeño más modesto de la CNN personalizada puede atribuirse a la menor profundidad de su arquitectura y la falta de características preentrenadas, lo que limita su capacidad para generalizar en un conjunto de datos tan complejo.

En la Figura 5 se muestra una visualización comparativa de las métricas, resaltando las diferencias entre los tres modelos.

IV. Discusión

Los resultados obtenidos en este estudio muestran que, en general, el modelo **ResNet50** logró el mejor desempeño

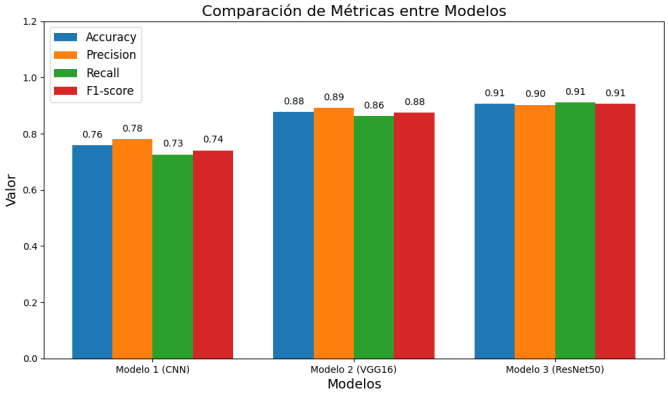

Figura 5. Comparación gráfica de métricas entre los tres modelos.

en comparación con los otros dos modelos evaluados. Sin embargo, es importante señalar algunos factores que podrían haber influido en el rendimiento de los modelos y que podrían mejorar en futuras iteraciones del trabajo.

Uno de los principales desafíos fue el tamaño de las imágenes de entrada. Los modelos **VGG16** y **ResNet50** fueron originalmente diseñados para trabajar con imágenes de mayor resolución, típicamente de 244×244 píxeles. Sin embargo, debido a las limitaciones de tiempo y recursos computacionales, se optó por redimensionar las imágenes a 64×64 píxeles. Esta reducción en la resolución de las imágenes pudo tener un impacto negativo en el desempeño de los modelos, ya que se pierde información visual importante que podría haber mejorado la precisión de las predicciones. Si se hubiera mantenido la resolución original, es probable que los modelos hubieran logrado un mejor desempeño, aunque este ajuste hubiera requerido mayores tiempos de entrenamiento y más recursos computacionales. Este es un aspecto a considerar en futuros trabajos, donde se podría balancear entre resolución de imágenes y tiempos de entrenamiento.

Otro aspecto relevante fue la distribución de las clases en el conjunto de datos. Aunque se intentó realizar un balanceo de las clases, este proceso resultó ser muy costoso en términos de tiempo de ejecución. Debido a restricciones de tiempo, se decidió no continuar con esta técnica de balanceo, lo que podría haber afectado ligeramente el desempeño, especialmente en clases menos representadas como en este caso fue la de neumonía viral que, como se evidencia en las matrices de confusión, fue en todos los modelos la que menos verdaderos positivos tuvo y era la que menos datos de entrenamiento tenía (1076) comparado con la clase que mas datos tenía (8153) para los pulmones sanos. Para modelos futuros, aplicar una estrategia de balanceo de clases, como el aumento de datos específico para las clases minoritarias o el uso de técnicas de ponderación durante el entrenamiento, podría mejorar la capacidad del modelo para clasificar correctamente las clases menos representadas y, en consecuencia, mejorar las métricas de precisión y recall.

Adicionalmente, cabe mencionar que el proceso de preprocesamiento y aumento de datos, aunque útil para mitigar el so-

breajuste, también podría mejorarse en futuras investigaciones. Técnicas adicionales de aumento de datos, como la rotación y la distorsión de las imágenes, podrían contribuir a una mayor robustez del modelo y a un mejor desempeño general.

En resumen, los resultados obtenidos son prometedores, pero existen varias áreas de mejora, como la resolución de las imágenes, el balanceo de las clases y la ampliación de las técnicas de aumento de datos. Estos aspectos deberían ser considerados para futuros trabajos con el objetivo de mejorar el desempeño y la generalización de los modelos en tareas de clasificación médica utilizando imágenes de radiografías de tórax.

V. Conclusiones

En este trabajo se evaluaron tres modelos para la clasificación de imágenes de radiografías de tórax en cuatro categorías: pulmones sanos, COVID-19, opacidad pulmonar y neumonía viral. Los resultados obtenidos muestran que el modelo **ResNet50** fue el de mejor desempeño, alcanzando un *accuracy* de 0.91, un *precision* de 0.90, un *recall* de 0.91 y un *F1-score* de 0.91. Este rendimiento superior puede atribuirse a la arquitectura profunda de ResNet50, que es capaz de aprender representaciones más complejas de las imágenes.

El modelo **VGG16** también mostró un desempeño notable, con un **accuracy** de 0.88, *precision* de 0.89, *recall* de 0.86 y *F1-score* de 0.88. A pesar de ser más ligero que ResNet50, su desempeño sigue siendo robusto, especialmente considerando que fue configurado mediante *transfer learning*.

El modelo **CNN personalizada** mostró un desempeño inferior, con un **accuracy** de 0.76, *precision* de 0.78, *recall* de 0.73 y *F1-score* de 0.74. Aunque es un modelo más sencillo, los resultados obtenidos sugieren que se podría mejorar su desempeño mediante una mayor complejidad en su arquitectura o la aplicación de técnicas de preprocesamiento más avanzadas.

En general, se concluye que los modelos preentrenados como **ResNet50** y **VGG16** son altamente efectivos para tareas de clasificación médica con imágenes de radiografías de tórax. Sin embargo, la reducción en el tamaño de las imágenes, debido a limitaciones computacionales, pudo afectar negativamente el desempeño, por lo que futuros trabajos podrían explorar resoluciones más altas para mejorar los resultados. También se sugiere que el balanceo de clases podría beneficiar la precisión de las predicciones, especialmente en clases menos representadas como lo fue la de neumonía viral.

Referencias

[1] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical image analysis*, vol. 42, pp. 60–88, 2017.

[2] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[3] L. Wang, Z. Q. Lin, and A. Wong, "Deep learning for covid-19 detection in medical imaging," *Nature Machine Intelligence*, vol. 2, no. 6, pp. 342–348, 2020.

[4] T. Rahman, M. E. H. Chowdhury, A. Khandakar, R. Mazhar, M. A. Kadir, Z. B. Mahbub, K. R. Islam, M. S. Khan, A. Iqbal, N. Al-Emadi, M. B. I. Reaz, and M. T. Islam, "Covid-19 radiography database," https://www.kaggle.com/datasets/tawsifurrahman/covid19-radiography-database, 2020, winner of the COVID-19 Dataset Award by Kaggle Community.