Chest X-ray classification: Exploring with Vision Transformer

1st Luz Marina Pilco Pancca
Ciencia de la computación
Universidad Nacional de San Agustín
Arequipa, Perú
lpilco@unsa.edu.pe

2nd Junior Valentin Gomez contreras
Ciencia de la computación
Universidad Nacional de San Agustín
Arequipa, Perú
jgomezcon@unsa.edu.pe

3rd Erika Sejje Condori
Ciencia de la computación
Universidad Nacional de San Agustín
Arequipa, Perú
esejjec@unsa.edu.pe

***Abstract*—La clasificación de imágenes médicas es una tarea crucial en el diagnóstico y tratamiento de enfermedades. En este estudio, utilizamos un conjunto de datos compuesto por radiografías de tórax y tomografías computarizadas de pacientes que dieron positivo en la prueba de COVID-19, pacientes con neumonía y pacientes con resultados normales, abarcando así tres clases. Evaluamos y comparamos la eficacia de varios modelos de aprendizaje automático y aprendizaje profundo en la clasificación de estas imágenes. Los modelos considerados incluyen KNN, Random Forest, ResNet y Vision Transformer. Nuestro objetivo es identificar el modelo que proporciona la mayor precisión y eficiencia en la clasificación de las imágenes médicas, proporcionando así una herramienta útil para los profesionales de la salud en la toma de decisiones clínicas.**

***Index Terms*—Chest X-ray, Machine learning, deep learning, convolution neural network, Vision Transformer.**

I. Introducción

La pandemia de COVID-19 ha resaltado la importancia de nuevas tecnologías de diagnóstico rápido y preciso en su identificación y su respectivo tratamiento. Según la Organización Mundial de la Salud (OMS), La neumonía es responsable del 14% de todas las muertes en niños menores de 5 años, siendo una de las principales causas de mortalidad infantil [1]. Tanto el COVID-19, como la neumonía han cobrado vidas sin importar la edad o el genero. Es necesario detectar a tiempo y con precisión estas enfermedades respiratorias graves para su tratamiento.

En esta investigación, exploramos la efectividad de los modelos de aprendizaje profundo y visión transformer para el diagnóstico automatizado de neumonía, Covid o ninguno utilizando imágenes de rayos X de tórax. Nos basamos en el articulo [2] que presenta un análisis detallado e implementación de técnicas de aprendizaje profundo y visipon transformer para clasificar imágenes de rayos X de tórax. Nuestra investigación en aplicar metodologías similares a un nuevo conjunto de datos obtenido de Kaggle, que consta de 5,863 imágenes de rayos X de tórax etiquetadas, incluyendo 3,883 casos de neumonía y 1,349 casos normales.

II. Estado del arte

La identificación manual del diseño del implante a partir de imágenes radiográficas requiere mucho tiempo, de un profesional y es propenso a errores. No identificar correctamente el diseño del implante complica el proceso operatorio. Borja at al [3] utiliza una red neuronal convolucional profunda para detectar el diseño de prótesis de reemplazo total de cadera en radiografías. Fan et al [4] propone una red neuronal convolucional profunda (CNN) con un módulo de atención para detectar la osteoporosis en radiografías panorámicas obteniendo resultados prometedores.

Actualmente, los autores utilizan CNN modificadas con otros enfoques para clasificar fotografías de rayos X. Zhou et al. [2] compararon el rendimiento de varios modelos para la clasificación del fabricante de rayos X de implantes de hombro, los modelos incluyen técnicas de aprendizaje profundo y visión transformer especificamente CNN modificadas como ResNet, ViT, VGG-16, los resultados mostraron que los ViT lograron el mejor desempeño en estas tareas y que el aprendizaje por transferencia mejoró el ViT por un amplio margen. Sheeta et al [**?**] presentan un marco de aprendizaje profundo para la detección y clasificación del fabricante de implantes de hombros mediante radiografías, la extracción de características y el aumento de datos tienen un impacto significativo. Las investigaciones contrastan la eficacia de los modelos de aprendizaje profundo y visión transformer.

Recientemente, los rayos X desempeñan un papel crucial en el diagnóstico de COVID-19 o neumonía. Varias investigaciones se enfocan en desarrollar esta tarea con precisión [5]–[12]. Hariri et al [5] explora el uso de aprendizaje profundo, específicamente redes neuronales convolucionales (CNNs), para diagnosticar COVID-19 y neumonía a partir de imágenes de radiografías de tórax. Kavya et at [6] de igual forma uso CNN como ResNet50 e InceptionV3, también se experimentó con un modelo basado en Mask-RCNN para la identificación de opacidades pulmonares, además uso transferencia de Aprendizaje para mejorar la precisión del diagnóstico. Chaturvedi et al [7] del mismo modo uso varios modelos CNN, incluyendo ResNet50 e InceptionV3, usando transferencia de aprendizaje y ajuste, para los datos se considero aumento de Datos, se utilizaron técnicas como volteo, escalado y rotación de las imágenes. Rodolfo at al [8] emplearon descriptores de textura y modelos de redes neuronales convolucionales (CNN) preentrenados para extraer características relevantes de las imágenes de CXR.

A diferencia de los anteriores autores Wang [9] incor-

pora Vision Transformers (ViTs), superando a varios modelos basados en CNN, esto puede deberse a la capacidad de los ViTs para captar mejor las características complejas de las imágenes. Singh at al [12] realizan un análisis exhaustivo de un marco Vision Transformer (ViT) para la detección de neumonía en radiografías de tórax, los resultados de visión transformer muestran superioridad respecto a los modelos CNN, su limitacion es la escases de datos. Chen at al [10], [11] también incorporan Visión Transformers, evalúa varias arquitecturas, incluidas redes neuronales eficientes (EfficientNet), transformadores de visión multiescala (MViT), transformadores de visión eficientes (EfficientViT) y transformadores de visión (ViT). Los modelos multiescala como MViT y EfficientNet tienden a sobreajustarse. Por el contrario, los modelos de transformador de visión obtienen buenos resultados.

III. Marco Teórico

A. Modelos tradicionales de aprendizaje automático

1) Random Forest: Random forest es un algoritmo de aprendizaje automático que utiliza múltiples árboles de decisión para mejorar la precisión de predicciones y clasificaciones. Cada árbol se entrena con una muestra diferente del conjunto de datos y hace una predicción. Luego, las predicciones de todos los árboles se combinan (promediando para regresión o votando para clasificación) para obtener el resultado final. Este método es robusto frente al sobreajuste y maneja bien datos complejos y con muchas variables [13].

2) K-Nearest Neighbor: La clasificación KNN surge de la necesidad de realizar análisis discriminantes en situaciones donde los métodos paramétricos confiables para calcular las densidades de probabilidad no están disponibles o son difíciles de obtener [14].

El algoritmo 1, nos muestra el pseudocódigo genérico de K-Nearest Neighbor.

Algorithm 1: Algoritmo K-Nearest Neighbor

```
Seleccionar el número de los k vecinos.;
Elegir aleatoriamente puntos inicales.;
repeat
    Calcular las distancias de un punto con sus vecinos.;
    Obtener los k vecinos más cercanos según la distancia calculada.;
    Contar el número de puntos en cada categoría.;
    Asignar los nuevos puntos a la máxima categoría.;
until Recorrer todos los puntos del conjunto de datos.;
```

B. Modelos de aprendizaje profundo

1) VGG-16 Architecture: VGG16 es una red neuronal convolucional profunda de 16 capas, es muy popular usualmente se utiliza para tareas de clasificación de imágenes, se hizo famosa por su simplicidad y efectividad, obteniendo excelentes resultados en la clasificación de grandes bases de datos de imágenes [15]. Aunque VGG-16 es una arquitectura muy poderosa, requiere alrededor de 144 millones de parámetros y lleva mucho tiempo entrenar en la computadora portátil [2]. En la figura 1 se muestra la arquitectura de VGG16.

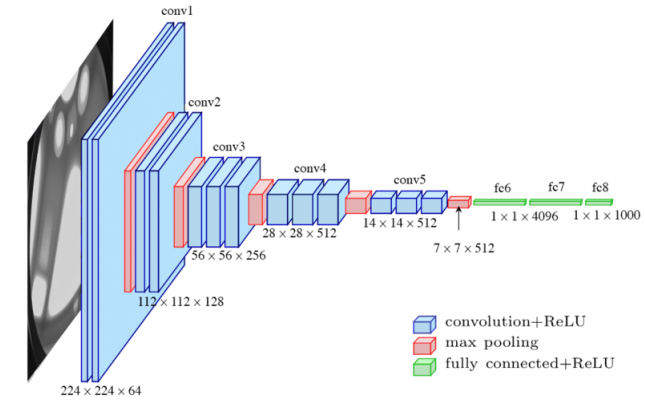

Fig. 1: Arquitectura VGG16 Referencia

2) ResNet-50: ResNet-50 (Residual Network with 50 layers) es una arquitectura de red neuronal profunda desarrollada por Kaiming He at al [16]. ResNet-50 presento el concepto de bloques residuales que permiten entrenar redes mucho más profundas sin los problemas de degradación comunes en redes tradicionales. Existen diversas variaciones de redes residuales, pero ResNet-50 ofrece una profundidad adecuada para abordar problemas de clasificación. ResNet-50 es un modelo poderoso para realizar aprendizaje por transferencia, en la figura 3 se muestra la arquitectura de ResNet-50.

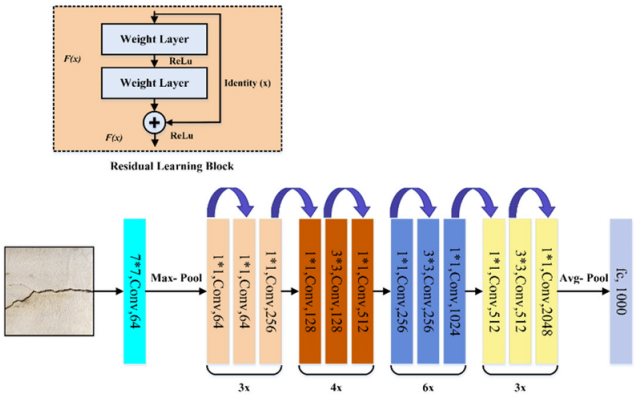

Fig. 2: Arquitectura ResNet-50 [17].

3) InceptionV3: InceptionV3 es una red neuronal convolucional profunda para clasificacion y reconocimiento de imágenes, desarrollada por Google Research [18]. InceptionV3 mejora la eficiencia y precisión en tareas de clasificación de imágenes mediante el uso de módulos de Inception, que permiten el procesamiento simultáneo de convoluciones de diferentes tamaños [19]. Se introduce un nuevo concepto llamado Factorización en convoluciones pequeñas, que consiste simplemente en reemplazar un filtro $nn \times nn$ por un filtro $nn \times 1$ y un filtro $1 \times nn$, esta técnica reduce la cantidad de

parámetros, ayudando a prevenir el sobreajuste.. En la figura 3 se muestra la arquitectura de InceptionV3.

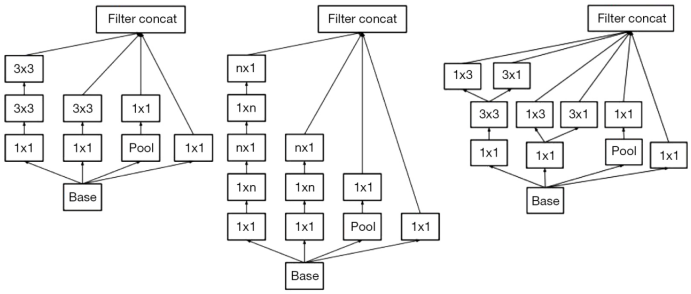

Fig. 3: Arquitectura InceptionV3 [20].

C. Transformers

Los transformadores (*Transformer*) han dominado el campo del procesamiento del lenguaje natural y recientemente han tenido un impacto en el área de la visión por computadora [21]. Los transformadores son un tipo de red neuronal profunda basada principalmente en el mecanismo de autoatención. Los modelos basados en transformadores funcionan de manera similar o mejor que otros tipos de redes, como las redes neuronales convolucionales y recurrentes [22].

D. Visión transformers

Dado su alto rendimiento, *transformer* está recibiendo cada vez más atención por parte de la comunidad de visión por computadora, los modelos más representativos son el transformador de detección (DETR) [23], ViT [24], el transformador de imagen con eficiencia de datos (DeiT) [25] y Swin-Transformer [26].

1) DETR: Propuesto por Carion et al. [23], fue la primera aplicación de transformadores, se enfoco en la deteccion de objetos, DETR es un modelo de detección de extremo a extremo que utiliza un codificador transformador para modelar la relación entre las características de la imagen extraídas por una CNN, un decodificador transformador para generar consultas de objetos y una red *feed-forward* para asignar etiquetas y vincular los cuadros alrededor de los objetos.

2) ViT: Propuesto por Dosovitskiy et al. [24], es un modelo de clasificación de imágenes, la imagen de entrada se convierte en una serie de parches, cada uno de ellos junto con un método de codificación posicional que codifica las posiciones espaciales de cada parche para proporcionar información espacial. Luego, los parches, junto con un token de clase, se introducen en el transformador para calcular el MHSA y generar las incorporaciones aprendidas de los parches. El estado del token de clase de la salida de ViT sirve como representación de la imagen. Por último, se utiliza un perceptrón multicapa (MLP) para clasificar la representación de la imagen aprendida. Además de las imágenes sin procesar, los mapas de características de las CNN se pueden introducir en un ViT para realizar mapas relacionales.

3) Deit: Propuesto por Touvron et al. [25], para resolver el problema de los datos de entrenamiento a gran escala que requiere ViT y garantizar el rendimiento en datos a pequeña escala.

4) Swin: Propuesto por Liu et al. [26], para reducir el costo de calcular la atención de imágenes de alta resolución y abordar los distintos tamaños de parche.

E. Transfer Learning

El aprendizaje por transferencia (transfer learning) es una técnica de aprendizaje profundo, se toma como punto de partida un modelo previamente entrenado, en lugar de entrenar un modelo de cero, esto permite aprovechar el conocimiento del modelo preentrenado, acelerando el proceso de entrenamiento, obteniendo un mejor rendimiento sin la necesidad de tener conjuntos de datos grandes [27]. El aprendizaje por transferencia es una buena opción para realizar tareas en las que los datos son limitados, para realizar la clasificación los datos son insuficientes, por tanto se opta por aplicar la técnica del aprendizaje por transferencia.

IV. Base de Datos

A. Base de Datos Orginal

El conjunto de datos utilizado se recopiló de varios fuentes de acceso público los datos se recopilarán indirectamente de médicos e instituciones [28]–[30], disponibles en https://github.com/ieee8023/covid-chestxray-dataset, https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia, https://github.com/agchung, https://data.mendeley.com/datasets/dvntn9yhd2/1. La base de datos contiene radiografías de tórax y tomografías computarizadas de pacientes que dieron positivo en la prueba de COVID-19 o que se sospecha que tienen otras neumonías virales y bacterianas (MERS, SARS y SDRA). El conjunto de datos comprende un total de 5,863 imágenes de rayos X de tórax etiquetadas, incluyendo 3,883 casos de neumonía y 1,349 casos normales.

V. Propuesta

Para realizar la clasificación los datos son insuficientes, por tanto se opta por aplicar *data augmentation* y la técnica del aprendizaje por transferencia. Todos los modelos KNN, Random Forest, VGG-16, ResNet-50 y Visión Transformer se utilizan para el aprendizaje por transferencia con los pesos de ImageNet. Las capas convolucionales en la red servirán para extraer la característica local de cualquier imagen de entrada, por lo que los pesos de esas capas se congelan para que no se puedan entrenar. El modelo solo entrenará en las últimas dos capas personalizadas completamente conectadas para la tarea de implantes de hombro.

VI. Resultados

KNN se utilizó la distancia euclidiana como métrica de medición y con k = 30. Se obtuvo un accuracy de 0.891 La Figura 4 muestra la matriz de confusión de KNN.

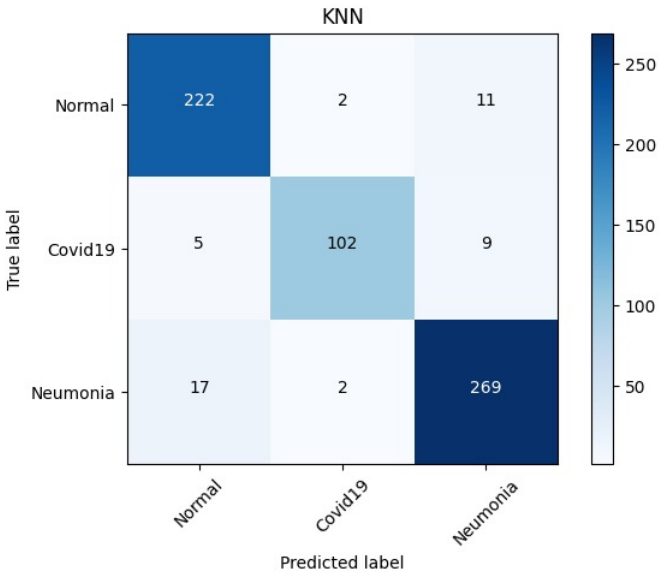

Fig. 4: Matriz de confusión KNN

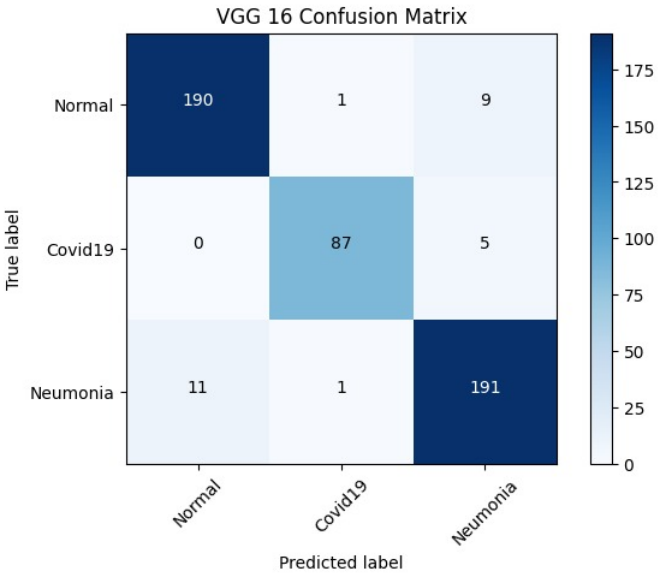

Fig. 6: Matriz de confusión VGG-16

La Figura 5 muestra la matriz de confusión de ResNet-50 con aprendizaje por transferencia y validación cruzada. InceptionV3 también se implementó mediante Transfer Learning. Solo Random Forest, ResNet-50 e InceptionV3 utilizaron técnicas de validación cruzada. Los resultados de ResNet-50, acc 57% no son los esperamos, se tendria que revisar el modelo. La Figura 6 muestra la matriz de confusión de VGG-16, obtuvo un acc de 94%, un resultado óptimo.

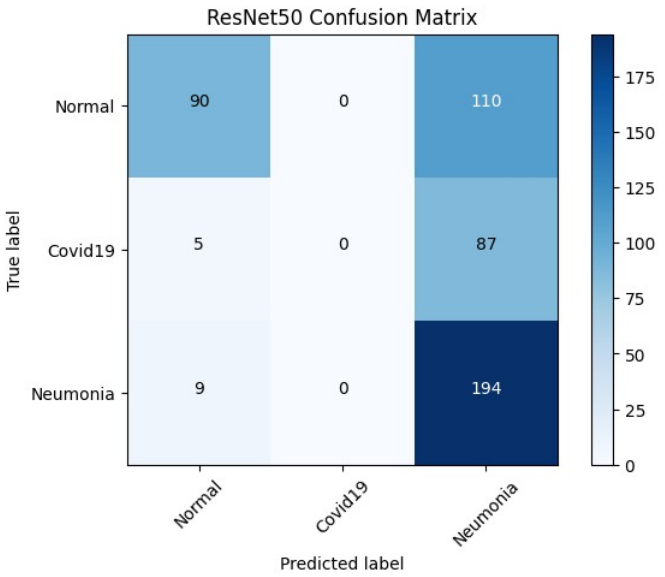

Fig. 5: Matriz de confusión Restnet-50

La Figura 7 muestra la matriz de confusión de Vision Transformer con el aprendizaje por transferencia. La puntuación de precisión de Vision Transformer es mejor que la de todos los demás modelos, un acc del 96%.

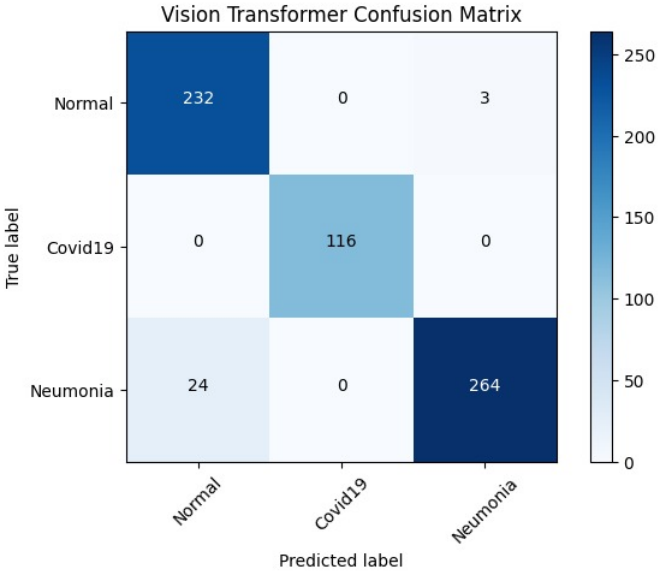

Fig. 7: Matriz de confusión Visión Transformer

Al analizar los resultados obtenidos, se demuestra que la extracción de características y el aumento de datos tienen un efecto significativo en el rendimiento en los modelos. El modelo Vision Transformer obtiene el mejor resultado, el mecanismo de autoatención es extremadamente eficaz no solo en el ámbito de la PNL, sino también en la visión por

computadora. Dividir la imagen en múltiples *patches* (parches) permite al modelo comprender mejor la imagen. Estos parches se envían al codificador del transformador, donde se aplica el mecanismo de autoatención, se identifica las características más significativas para cada clase y predice una nueva imagen de entrada basándose en las partes más relevantes.

VII. Conclusiones

La clasificación de imagenes de rayos X es fundamental para el diagnóstico y tratamiento del COVID-19 o neumonía. En este estudio, hemos utilizado un conjunto de datos compuesto por radiografías de tórax de pacientes con COVID-19, neumonía y resultados normales. Hemos evaluado varios modelos de aprendizaje automático y profundo, incluyendo KNN, Random Forest, VGG-16, ResNet y Visión Transformer. Los resultados muestran que los modelos basados en aprendizaje profundo, especialmente ResNet y Vision Transformer, proporcionan mayor precisión y eficiencia en la clasificación, ofreciendo un acc del 96%, siendo un resultado óptimo para la toma de decisiones clínicas.

VIII. Trabajos Futuros

En trabajos futuros, planeamos expandir este estudio incorporando conjuntos de datos más amplios y diversos que incluyan imágenes de pacientes con otras afecciones respiratorias para mejorar la robustez y generalización de los modelos. También se prevé el desarrollo de enfoques híbridos que combinen múltiples modelos de aprendizaje profundo y métodos tradicionales de aprendizaje automático para mejorar aún más la precisión de la clasificación.
Finalmente, se evaluará la implementación de estos modelos en aplicaciones reales, analizando su impacto en la práctica médica y su aceptación por parte de los profesionales de la salud.

IX. Anexos

El código de este artículo se puede encontrar en https://github.com/zulmarina1687/MCC-VA-TrabajoFinal
El conjunto de datos original disponible en https://www.kaggle.com/datasets/alsaniipe/chest-x-ray-image/code

References

[1] I. Rudan, C. Boschi-Pinto, Z. Biloglav, K. Mulholland, and H. Campbell, "Epidemiology and etiology of childhood pneumonia," *Bulletin of the world health organization*, vol. 86, pp. 408–416B, 2008.

[2] M. Zhou and S. Mo, "Shoulder implant x-ray manufacturer classification: Exploring with vision transformer," *arXiv preprint arXiv:2104.07667*, 2021.

[3] A. Borjali, A. F. Chen, O. K. Muratoglu, M. A. Morid, and K. M. Varadarajan, "Detecting total hip replacement prosthesis design on plain radiographs using deep convolutional neural network," *Journal of Orthopaedic Research®*, vol. 38, no. 7, pp. 1465–1471, 2020.

[4] H. Fan, J. Ren, J. Yang, Y.-X. Qin, and H. Ling, "Osteoporosis prescreening using panoramic radiographs through a deep convolutional neural network with attention mechanism," *arXiv preprint arXiv:2110.09662*, 2021.

[5] M. Hariri and E. Avşar, "Covid-19 and pneumonia diagnosis from chest x-ray images using convolutional neural networks," *Network Modeling Analysis in Health Informatics and Bioinformatics*, vol. 12, no. 1, p. 17, 2023.

[6] N. S. Kavya, N. Veeranjaneyulu, D. D. Priya, *et al.*, "Detecting covid19 and pneumonia from chest x-ray images using deep convolutional neural networks," *Materials Today: Proceedings*, vol. 64, pp. 737–743, 2022.

[7] A. Chaturvedi, A. Shrivastava, D. Verma, I. Kumar, M. Gupta, and A. K. Rao, "Covid-19 pneumonia identification from chest x-ray image using deep learning," in *2023 4th International Conference on Intelligent Engineering and Management (ICIEM)*, pp. 1–5, IEEE, 2023.

[8] R. M. Pereira, D. Bertolini, L. O. Teixeira, C. N. Silla, and Y. M. Costa, "Covid-19 identification in chest x-ray images on flat and hierarchical classification scenarios," *Computer Methods and Programs in Biomedicine*, vol. 194, p. 105532, 2020.

[9] T. Wang, Z. Nie, R. Wang, Q. Xu, H. Huang, H. Xu, F. Xie, and X.-J. Liu, "Pneunet: deep learning for covid-19 pneumonia diagnosis on chest x-ray image analysis using vision transformer," *Medical & Biological Engineering & Computing*, vol. 61, no. 6, pp. 1395–1408, 2023.

[10] T. Chen, I. Philippi, Q. B. Phan, L. Nguyen, N. T. Bui, C. daCunha, and T. T. Nguyen, "A vision transformer machine learning model for covid-19 diagnosis using chest x-ray images," *Healthcare Analytics*, vol. 5, p. 100332, 2024.

[11] T. Chen, I. Philippi, Q. B. Phan, L. Phan, T. Nguyen, *et al.*, "High-accuracy fine-tuned vision transformer model for diagnosing covid-19 from chest x-ray images," *Authorea Preprints*, 2024.

[12] S. Singh, M. Kumar, A. Kumar, B. K. Verma, K. Abhishek, and S. Selvarajan, "Efficient pneumonia detection using vision transformers on chest x-rays," *Scientific Reports*, vol. 14, no. 1, p. 2487, 2024.

[13] S. J. Rigatti, "Random forest," *Journal of Insurance Medicine*, vol. 47, no. 1, pp. 31–39, 2017.

[14] D. Rajoot, P. Singh, and M. Bhattacharya, "A new technique for feature selection and cluster center initialization.," pp. 119–125, 01 2010.

[15] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[17] L. Ali, F. Alnajjar, H. Jassmi, M. Gochoo, W. Khan, and M. Serhani, "Performance evaluation of deep cnn-based crack detection and localization techniques for concrete structures," *Sensors*, vol. 21, p. 1688, 03 2021.

[18] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

[19] V. Ashish, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, p. I, 2017.

[20] Q. Guan, X. Wan, H. Lu, B. Ping, D. Li, L. Wang, Y. Zhu, Y. Wang, and J. Xiang, "Deep convolutional neural network inception-v3 model for differential diagnosing of lymph node in cytological images: a pilot study," *Annals of Translational Medicine*, vol. 7, no. 14, 2019.

[21] K. He, C. Gan, Z. Li, I. Rekik, Z. Yin, W. Ji, Y. Gao, Q. Wang, J. Zhang, and D. Shen, "Transformers in medical image analysis," *Intelligent Medicine*, vol. 3, no. 1, pp. 59–78, 2023.

[22] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, *et al.*, "A survey on vision transformer," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 1, pp. 87–110, 2022.

[23] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*, pp. 213–229, Springer, 2020.

[24] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[25] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *International conference on machine learning*, pp. 10347–10357, PMLR, 2021.

[26] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

[27] L. Torrey and J. Shavlik, "Transfer learning," in *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, pp. 242–264, IGI global, 2010.

[28] J. P. Cohen, P. Morrison, L. Dao, K. Roth, T. Q. Duong, and M. Ghassemi, "Covid-19 image data collection: Prospective predictions are the future," *arXiv 2006.11988*, 2020.
[29] J. P. Cohen, P. Morrison, and L. Dao, "Covid-19 image data collection," *arXiv 2003.11597*, 2020.
[30] Kumar and Sachin, "Covid19-pneumonia-normal chest x-ray images," Mendeley Data, 2022.