NÖHÜ Müh. Bilim. Derg. / NOHU J. Eng. Sci., 2025; 14(4), 1385-1397
Niğde Ömer Halisdemir Üniversitesi Mühendislik Bilimleri Dergisi
Niğde Ömer Halisdemir University Journal of Engineering Sciences
Araştırma makalesi / Research article
www.dergipark.org.tr/tr/pub/ngumuh / www.dergipark.org.tr/en/pub/ngumuh

AI-powered diagnosis of respiratory diseases: Evaluating vision transformers and ResNet architectures for covid-19 and lung pathologies

Solunum hastalıklarının yapay zeka destekli teşhisi: Covid-19 ve akciğer patolojileri için görü dönüştürücüleri ve ResNet mimarilerinin değerlendirilmesi

Ahmet Solak[1,*]

[1] *Konya Teknical University, Department of Electrical and Electronics Engineering, 42250, Konya, Türkiye*

Abstract

This study systematically evaluates the efficacy of advanced deep learning architectures, namely Vision Transformers (ViT) and various ResNet models (ResNet50, ResNet101, ResNet152), in the classification of chest radiographs into four clinically significant diagnostic categories: Normal, Lung Opacity, Viral Pneumonia, and COVID-19. A meticulously curated dataset comprising 21,165 chest X-ray images was utilized to benchmark the models' performance across key evaluation metrics, including precision, recall, F1-score and accuracy. The experimental evaluation reveals that ViT model achieved 90.25% accuracy, 91.56% precision, 89.22% recall, and a 90.25% F1-score. These findings highlight the potential of AI-driven approaches in augmenting medical diagnostics, improving diagnostic accuracy, and enhancing healthcare delivery, particularly in resource-limited settings. The study underscores the applicability of Vision Transformers in complex medical imaging tasks and contributes to the growing body of research supporting AI-based solutions for respiratory diseases and other healthcare challenges.

Keywords: Chest radiographs, COVID-19 diagnosis, Deep learning, ResNet, Vision transformer

Öz

Bu çalışma, gelişmiş derin öğrenme mimarilerinin – özellikle Vision Transformers (ViT) ve çeşitli ResNet modellerinin (ResNet50, ResNet101, ResNet152) – göğüs röntgenlerini Normal, Akciğer Opasitesi, Viral Pnömoni ve COVID-19 olmak üzere dört klinik açıdan önemli tanısal kategoriye sınıflandırmadaki etkinliğini sistematik olarak değerlendirmektedir. Modellerin performansını, hassasiyet, geri çağırma, F1-skora ve doğruluk gibi temel değerlendirme metrikleri üzerinden ölçmek amacıyla özenle hazırlanmış 21.165 göğüs X-ışını görüntüsünden oluşan bir veri seti kullanılmıştır. Deneysel değerlendirmeler, ViT modelinin %90.25 doğruluk, %91.56 hassasiyet, %89.22 geri çağırma ve %90.25 F1-skora elde ettiğini ortaya koymaktadır. Bu bulgular, yapay zeka temelli yaklaşımların tıbbi tanı süreçlerini güçlendirme, tanı doğruluğunu artırma ve özellikle kaynak kısıtlı ortamlarda sağlık hizmetlerinin sunumunu iyileştirme potansiyeline işaret etmektedir. Çalışma, karmaşık tıbbi görüntüleme görevlerinde Vision Transformers'ın uygulanabilirliğini vurgulamakta ve solunum yolu hastalıkları ile diğer sağlık sorunlarına yönelik yapay zeka temelli çözümleri destekleyen artan araştırma literatürüne katkıda bulunmaktadır.

Anahtar kelimeler: Göğüs röntgenleri, COVID-19 tanısı, Derin öğrenme, ResNet, Vision transformer

1 Introduction

The onset of Coronavirus Disease 2019 (COVID-19), instigated by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), precipitated a global public health emergency of unparalleled magnitude. First identified in December 2019 in Wuhan, China, the rapid global spread of this novel virus led the World Health Organization (WHO) to declare a pandemic on March 11, 2020 [1]. By July 30, 2024, COVID-19 had resulted in 775686716 confirmed cases and 7054093 deaths worldwide [2, 3], profoundly burdening healthcare systems, disrupting economies, and reshaping societal norms.

In addition to COVID-19, other respiratory diseases—including pneumonia, influenza, and various bacterial and viral pulmonary infections—continue to represent substantial threats to global health. These conditions often present with overlapping clinical and radiological features, further complicating diagnosis and management. Early and accurate detection of such diseases is crucial not only for initiating appropriate therapeutic interventions and improving patient outcomes, but also for mitigating transmission and optimizing the use of limited healthcare resources. Delays or inaccuracies in diagnosis can lead to worsened prognosis, increased morbidity and mortality, and the inefficient allocation of critical infrastructure, such as hospital beds, ventilators, and personal protective equipment [4-6]. Moreover, robust early diagnostic systems are essential for effective epidemiological surveillance and public health response strategies, allowing authorities to monitor disease dynamics and implement targeted interventions [7].

* Sorumlu yazar / Corresponding author, e-posta / e-mail: asolak@ktun.edu.tr (A. Solak)
Geliş / Received: 0.02.2025 Kabul / Accepted: 29.08.2025 Yayımlanma / Published: 15.10.2025
doi: 10.28948/ngumuh.1635030

Currently, reverse transcription-polymerase chain reaction (RT-PCR) remains the gold standard for COVID-19 diagnosis. However, its practical application is constrained by factors such as prolonged turnaround times, restricted accessibility, and a notable risk of false-negative results [8]. Similarly, the diagnosis of other respiratory diseases frequently relies on clinical evaluation, laboratory testing, and medical imaging. Among these, chest radiography (CXR) and computed tomography (CT) have proven indispensable for detecting pulmonary abnormalities—including ground-glass opacities, consolidations, and interstitial changes—across a range of infectious and non-infectious etiologies [9]. Nevertheless, accurate interpretation of these imaging modalities demands specialized radiological expertise, a resource that is often stretched thin in high-demand clinical environments, particularly during pandemics and seasonal outbreaks.

The integration of artificial intelligence (AI) and deep learning algorithms into medical imaging analysis represents a transformative advancement in diagnostic medicine. These technologies directly address longstanding challenges in accuracy, efficiency, and scalability within imaging-based diagnostics. AI-based platforms are now capable of analyzing CXRs and CT scans to assist clinicians in the rapid and accurate detection of a wide array of thoracic diseases, including but not limited to COVID-19, pneumonia, and other viral or bacterial infections [10]. State-of-the-art deep learning frameworks, such as convolutional neural networks (CNNs) and ResNet architectures, have demonstrated outstanding performance in extracting discriminative features from complex medical images. More recently, Vision Transformers (ViT) have introduced advanced self-attention mechanisms, enabling the precise capture of intricate visual patterns crucial for differentiating among diseases with similar radiological presentations. By enhancing diagnostic accuracy, reducing inter-observer variability, and expediting large-scale screening processes, AI-driven methodologies offer significant potential to strengthen healthcare systems globally. These advances promise not only to improve the clinical management of COVID-19 and related respiratory diseases but also to bolster public health preparedness for future epidemics and pandemics.

A growing body of literature has investigated the application of deep learning techniques to enhance the diagnostic precision of COVID-19 through the analysis of CXR and CT imaging. These studies underscore the efficacy of CNNs and other advanced AI-based architectures in improving classification precision, optimizing diagnostic workflows, and mitigating human error in medical imaging analysis.

Oh et al. developed an advanced classification framework utilizing a diverse, multi-source dataset that included cases of normal lungs, tuberculosis (TB), bacterial pneumonia, viral pneumonia (including COVID-19), and other non-COVID viral pneumonias. To promote robust model performance and generalizability, the dataset was meticulously divided into three distinct subsets: 70% for training, 10% for validation, and 20% for testing. To facilitate a fair performance comparison, an additional benchmarking dataset was curated specifically for evaluating the model against the COVID-Net framework. The classification system was constructed using the ResNet18 architecture, with interpretability enhanced through the integration of a probabilistic gradient-weighted class activation map (Grad-CAM). This interpretability mechanism identified critical regions within chest X-rays that were most indicative of COVID-19-related patterns. The ResNet18 model exhibited strong classification performance, attaining an overall accuracy of 88.9%. The proposed method yielded a precision of 83.4%, a recall of 85.9%, an F1-score of 84.4%, and a specificity of 96.4%. These performance metrics underscore the method's efficacy in accurately differentiating diagnostic categories while sustaining a high rate of true negative identifications [11].

Butt et al. introduced a deep learning framework that leverages pulmonary computed tomography image analysis to enhance the early diagnosis of COVID-19 pneumonia. Their study introduced a location-attention classification model, integrated with a Noisy-OR Bayesian function, to enhance classification accuracy and effectively differentiate between healthy cases, influenza-A viral pneumonia (IAVP), and COVID-19. Based on a dataset of 618 CT samples, the proposed model achieved an overall accuracy of 86.7%. This performance is attributed to the integration of critical radiological features—namely, ground-glass opacities and pleural abnormalities—that are indicative of COVID-19 pathology. This approach underscores the significant potential of deep learning in augmenting clinical diagnostic workflows by improving detection efficiency and reliability, while serving as a complementary tool to traditional methods like RT-PCR [12].

Khan et al. introduced CoroNet, a convolutional neural network derived from the Xception architecture, specifically engineered to improve the identification of COVID-19 from CXR images. The network was both trained and validated on a dataset compiled from two publicly available repositories, which includes 284 cases of COVID-19, 310 normal cases, 330 cases of bacterial pneumonia, and 327 cases of viral pneumonia. To mitigate the issue of class imbalance, a random sub-sampling strategy was implemented, ensuring a more balanced distribution of training samples. Prior to model training, all images were resized to 224×224 pixels to standardize input dimensions. CoroNet employed a transfer learning framework by initializing the network with weights pre-trained on the ImageNet dataset. Subsequently, the model underwent fine-tuning using a COVID-19-specific dataset to refine its feature extraction capabilities for precise disease classification. This approach resulted in an overall classification accuracy of 89.6%, thereby demonstrating the substantial efficacy of transfer learning in the context of medical image analysis. [13].

Xu et al. conducted a study aimed at enhancing early COVID-19 screening by leveraging CT imaging to differentiate COVID-19 cases from IAVP and healthy lung conditions. The dataset utilized in their research comprised 618 CT scans, obtained from three hospitals in Zhejiang Province, China. This dataset included scans from 110

patients diagnosed with COVID-19, 224 individuals with IAVP, and 175 healthy subjects. To facilitate accurate classification, the researchers developed a three-dimensional (3D) deep learning framework designed to segment regions of potential infection within the lung images. The segmented regions were subsequently categorized into three distinct classes—COVID-19, IAVP, and infection-unrelated regions (ITI)—based on confidence scores generated through a location-attention classification model. This approach enabled more precise identification of COVID-19-related abnormalities, thereby contributing to the advancement of automated diagnostic tools for respiratory diseases. The final classification step employed the Noisy-OR Bayesian function, yielding an overall accuracy of 86.7%. This study highlights the effectiveness of combining 3D segmentation with Bayesian classification in screening for COVID-19 through CT imaging [14].

Shadin et al. investigated the efficacy of deep learning models for COVID-19 detection using CXR images. Their study compared two methodological approaches: a bespoke CNN and a transfer learning strategy employing InceptionV3 architecture. The analysis was performed on a dataset comprising 1,553 CXR images that represent a range of respiratory conditions. The custom CNN achieved a training accuracy of 79.74% and a validation accuracy of 84.92%, whereas the InceptionV3-based model attained a higher performance with a training accuracy of 85.41% and a validation accuracy of 85.94%. These results emphasize the advantages of transfer learning, particularly in contexts characterized by limited datasets, as pre-trained models can effectively utilize features extracted from extensive datasets such as ImageNet [15].

Park et al. developed a robust ViT framework to advance the automated diagnosis of COVID-19 and other pulmonary infections using CXR images. Their research utilized a comprehensive, multi-institutional dataset comprising 17,548 CXR images, categorized into normal, other infections (including bacterial pneumonia and tuberculosis), and COVID-19 cases. The ViT model was rigorously evaluated across three independent external institutional test sets—CNUH, YNU, and KNUH—to assess its performance and generalizability in diverse clinical settings. The model demonstrated consistent and strong classification results, achieving average accuracy scores of 86.4% on CNUH, 85.9% on YNU, and 85.2% on KNUH test sets. Notably, these outcomes surpassed those of conventional models such as ResNet-50 and standard ViT architectures, highlighting the effectiveness of leveraging low-level CXR feature embeddings in combination with transformer-based learning. The study's findings underscore the potential of the proposed ViT approach to deliver accurate, stable, and generalizable diagnostic support for COVID-19 and related diseases in real-world healthcare environments [16].

Cannata et al. conducted a comprehensive study to develop an automated COVID-19 infection screening tool using chest X-ray (CXR) images, aiming to provide a rapid, cost-effective alternative to RT-PCR. The researchers utilized a large, publicly available CXR dataset containing four diagnostic classes: COVID-19, viral pneumonia, lung opacity (non-COVID lung infection), and normal cases. The methodology employed advanced artificial intelligence techniques, leveraging transfer learning with pre-trained networks to address data scarcity and computational efficiency. Four deep learning architectures were evaluated: InceptionV3, Xception, ResNet50, and Vision Transformer (ViT). All models were trained and tested using a consistent data split (70% training, 10% validation, 20% test), and the same dataset version (3,616 COVID-19, 10,192 normal, 6,012 lung opacity, and 1,345 viral pneumonia images). Experimental results showed that ViT significantly outperformed all convolutional neural network (CNN) architectures, achieving a test accuracy of 99.3%, compared to 85.58% for ResNet50 (the best CNN baseline). ViT also demonstrated high precision, recall, and F1-scores across all four classes, successfully distinguishing COVID-19 from other respiratory diseases and healthy cases. The authors highlighted the clinical potential of ViT-based computer-aided diagnostic tools to assist, accelerate, and automate the COVID-19 diagnosis process using CXR images [17].

Despite considerable progress in the application of deep learning to chest X-ray and CT imaging for the diagnosis of COVID-19 and other pulmonary diseases, several limitations persist in the existing literature. Many previous studies have relied on limited, single-center datasets or lacked external validation, thereby restricting the generalizability and clinical applicability of their findings. Furthermore, the majority of works have focused on conventional CNN architectures, with relatively few investigations evaluating the effectiveness of emerging transformer-based models, such as the Vision Transformer (ViT), on large, diverse, and multi-class chest X-ray datasets. In addition, the practical integration of these AI models into clinical workflows, particularly as real-time decision support tools in emergency and high-volume healthcare environments, remains underexplored. To address these gaps, the present study systematically compares the performance of ViT and ResNet architectures using a comprehensive, multi-institutional CXR dataset encompassing four clinically relevant diagnostic categories. The aim is to provide robust, generalizable evidence on the efficacy of transformer-based approaches and to discuss their potential clinical integration in real-world settings.

Building upon the foundations of prior research, this study investigates the effectiveness of advanced deep learning architectures, specifically ViT and various ResNet models (ResNet50, ResNet101, and ResNet152), in the classification of CXRs across four clinically relevant categories: Normal, Lung Opacity, Viral Pneumonia, and COVID-19. In contrast to traditional CNNs, which primarily focus on local feature extraction, ViTs employ self-attention mechanisms to capture long-range dependencies within images. This innovative approach not only diverges from conventional methods but also holds a significant promise for advancing medical image analysis by potentially enhancing diagnostic accuracy and enabling more comprehensive feature integration. This study systematically compares the performance of the ViT with that of the well-established ResNet architecture. A rigorous evaluation

framework is employed, leveraging key performance metrics—including accuracy, precision, recall, F1-score, and specificity—to provide a comprehensive assessment of each model's capabilities.

This study aims to advance the development of robust, AI-driven diagnostic tools for COVID-19 and similar respiratory ailments. By leveraging the capabilities of ViTs in medical imaging, the research seeks to enhance both the speed and precision of COVID-19 detection. Furthermore, it proposes a scalable framework designed to address current challenges while also accommodating future respiratory disease outbreaks. The core objectives of this investigation are outlined as follows:

1) Performance Evaluation: To assess and compare the classification performance of ViTs and ResNet models (ResNet50, ResNet101, ResNet152) in categorizing chest radiographs into four diagnostic groups: Normal, Lung Opacity, Viral Pneumonia, and COVID-19.

2) Clinical Relevance: To evaluate the potential of these AI models to assist healthcare professionals by providing rapid, accurate, and reliable diagnostic insights, enabling the differentiation of COVID-19 from other pulmonary pathologies.

This investigation endeavors to reconcile advanced artificial intelligence techniques with their clinical implementations, thereby promoting the incorporation of AI into medical diagnostics to enhance accuracy, efficiency, and accessibility.

2 Materials and methods

2.1 Dataset

The dataset utilized in this study comprises a total of 21,165 CXR images, systematically categorized into four diagnostic groups: 3,616 images of COVID-19-positive cases, 6,012 images with lung opacity (non-COVID lung infections), 10,192 normal images, and 1,345 images of viral pneumonia [18, 19]. This comprehensive collection was developed through a collaborative effort led by researchers from Qatar University and the University of Dhaka, together with international partners and medical professionals from Pakistan and Malaysia. The dataset has undergone multiple updates, with incremental additions to each class to support ongoing research in automated thoracic disease detection [20]. Figure 1 illustrates representative samples from each category. Prior to neural network ingestion, all images were uniformly resized to 128 × 128 pixels. Additionally, data augmentation methods—including horizontal flipping, rotation, and zooming—were implemented to increase variability and enhance the robustness of the model.

2.2 Model architecture

2.2.1 Residual networks (ResNet)

Residual Networks (ResNet) constitute a groundbreaking innovation in the design of deep CNNs, effectively addressing the degradation problem—an issue wherein increasing network depth unexpectedly results in higher training error. This phenomenon arises primarily from the challenge of training deeper networks, as gradients tend to diminish during backpropagation. Introduced by He et al., the ResNet architecture resolves this issue through residual learning, a paradigm that has since become foundational in computer vision tasks [21]. The hallmark of ResNet lies in its residual learning framework, wherein layers are designed to learn residual mappings relative to their inputs rather than complete transformations. This is accomplished through shortcut (or skip) connections, which bypass one or more layers, facilitating direct gradient propagation during backpropagation and thereby enhancing network training stability and efficiency. Such a mechanism not only facilitates the training of substantially deeper networks but also alleviates the vanishing gradient problem.

The residual block constitutes the fundamental building unit of the ResNet architecture, playing a critical role in preserving efficient gradient propagation and alleviating the vanishing gradient issue in deep neural networks. Let x denote the input to the residual block and define $F(x, \{W_i\})$ as the residual mapping that the network is designed to learn, where W_i represents the set of weights associated with the convolutional layers within the block. The output y of the residual block is consequently expressed as shown in Equation (1).

$$y = F(x, \{W_i\}) + x \tag{1}$$

In this formulation, let y denote the output of the residual block and x its corresponding input. Notably, the residual function $F(x, \{W_i\})$ is structured to incorporate x directly into its output, thereby establishing an identity mapping.

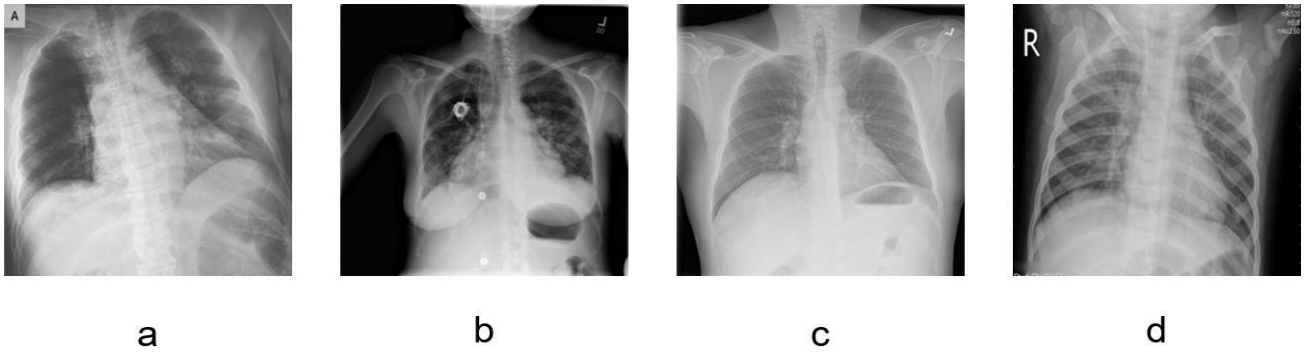

Figure 1. Illustrating four diagnostic categories. **(a)** COVID-19**, (b)** Lung opacity**, (c)** Normal, **(d)** Viral pneumonia

This approach simplifies the learning process by allowing the network to concentrate on learning residuals rather than the complete transformation. Figure 2 illustrates a schematic overview of the residual block architecture, highlighting its essential structural components.

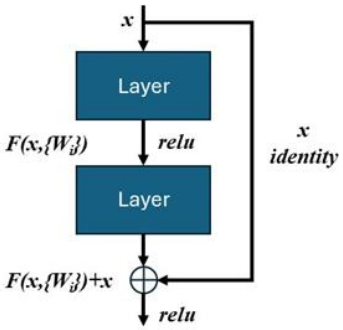

Figure 2. Residual block

Basic and Bottleneck Blocks:

ResNet employs two primary types of residual blocks, with the choice of blocks depending on the depth of the network architecture:

Basic Block: The basic block architecture is employed in shallow variants of the ResNet framework, specifically ResNet-18 and ResNet-34. The proposed architecture comprises two convolutional layers, each immediately followed by a batch normalization layer and a rectified linear unit (ReLU) activation function. This configuration is designed to enhance training stability and alleviate the vanishing gradient problem. Additionally, an identity mapping mechanism is incorporated, facilitating direct propagation of the input to the output, thereby preserving essential feature representations and improving gradient flow during backpropagation. Equation (2) provides a formal mathematical representation of the output y from the basic block.

$$y = ReLU(F(x, \{W_1, W_2\}) + x) \quad (2)$$

In this framework, W_1 and W_2 denote the weight matrices corresponding to the two convolutional layers that comprise the residual block. This design is intentionally simple, ensuring efficient gradient flow and stable training in relatively shallow networks. Its simplicity helps mitigate the vanishing gradient problem, which is critical for effective network optimization.

Bottleneck Block: For deeper neural network architectures, including ResNet-50, ResNet-101, and ResNet-152, the bottleneck block is utilized to enhance computational efficiency and facilitate the training of significantly deeper models. The bottleneck block follows a three-layer structure:

- A 1x1 convolution to reduce dimensionality,
- A 3x3 convolution to perform spatial feature extraction, and
- Another 1x1 convolution to restore dimensionality.

Equation (3). delineates the mathematical formulation of the bottleneck block's output, denoted by y.

$$y = ReLU(F(x, \{W_1, W_2, W_3\}) + x) \quad (3)$$

Here, W_1, W_2, and W_3 represent the weights of the three convolutional layers. By incorporating dimensionality reduction using 1x1 convolutions, the bottleneck block significantly reduces computational complexity. This efficiency facilitates the training of deeper networks without excessive computational cost, making it a cornerstone of deeper ResNet architectures. Figure 3 shows basic and bottleneck blocks.

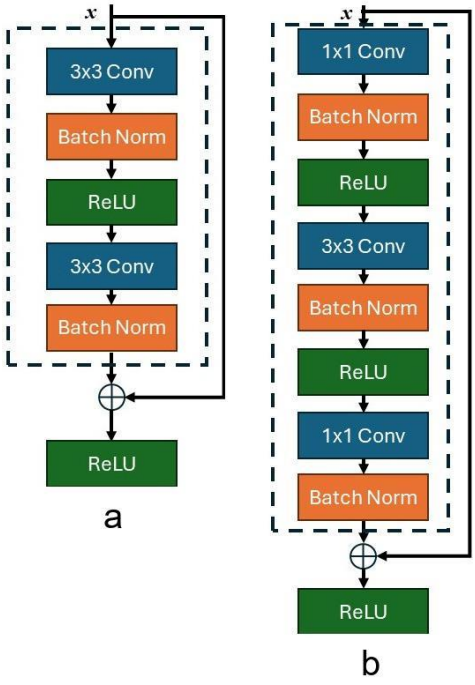

Figure 3. a) basic block **b)** bottleneck block

ResNet architectures are available in multiple variants, differentiated primarily by their depth and computational complexity. While deeper models possess a greater capacity to capture intricate patterns in data, they also demand significantly more computational resources. In ResNet architectures, the model's depth is quantified by the aggregate number of convolutional layers incorporated into its design.

ResNet-18 and ResNet-34: These relatively shallow variants utilize the basic block design, consisting of 18 and 34 convolutional layers, respectively. Their reduced computational requirements make them well-suited for applications with constrained resources or tasks involving less complex feature extraction.

ResNet-50, ResNet-101, and ResNet-152: Deep neural network architectures often incorporate bottleneck blocks, an effective design strategy that maximizes parameter efficiency while enabling significantly greater network

depth. For example, the ResNet family includes models such as ResNet-50, ResNet-101, and ResNet-152, which consist of 50, 101, and 152 convolutional layers, respectively. The proposed architectural models exhibit robust performance in addressing a wide range of advanced computer vision tasks, including image classification, object detection, and semantic segmentation. Their robustness and scalability make them particularly well-suited for deployment in large-scale datasets, such as ImageNet, where they have demonstrated superior performance in feature extraction and pattern recognition.

A defining characteristic of ResNet architectures is the use of an identity shortcut connection, which directly adds the input to the output of a residual block. When the input and output dimensions differ—such as during downsampling—a 1x1 convolution is employed to align dimensions, ensuring the addition operation remains mathematically valid. This innovative design supports the scalability of ResNet models, enabling their adaptation to varying levels of depth and complexity, thereby accommodating diverse applications in computer vision.

2.2.2 Vision transformer (ViT)

The Vision Transformer (ViT) represents a seminal advancement in computer vision by repurposing the Transformer architecture—originally developed for natural language processing (NLP)—to address image classification challenges [22]. In contrast to traditional CNNs, which rely on localized receptive fields and hierarchical feature extraction through convolutional operations, ViT decomposes an image into a sequence of patches. It then employs self-attention mechanisms to capture long-range dependencies and global contextual information, thereby offering a fundamentally different approach to visual representation.

Patch Embedding: In the ViT framework, an input image $\boldsymbol{I} \in \mathbb{R}^{HxWxC}$—with H, W, and C representing the image's height, width, and number of channels, respectively, initially segmented into a set of non-overlapping patches, each of dimensions P x P. This partitioning yields a total of $N = \frac{HW}{P^2}$ patches. Each patch is subsequently flattened into a one-dimensional vector and mapped into a lower-dimensional embedding space through a linear projection, thereby generating a sequence of patch embeddings $\boldsymbol{E} \in \mathbb{R}^{NxD}$ where D denotes the embedding dimension.

Positional Encoding: To maintain the spatial coherence of image patches, fixed positional encodings are incorporated into the patch embeddings, ensuring the preservation of spatial relationships within the input data. These encodings, represented as $\boldsymbol{P} \in \mathbb{R}^{NxD}$, are of the same dimensionality as the embeddings. The resulting positional encoding is expressed as Equation (4).

$$E_{pos} = E + P \tag{4}$$

Transformer Encoder: At the core of the ViT lies its Transformer encoder, an adaptation of the seminal architecture proposed by Vaswani et al. [23]. This encoder is structured as a stack of L identical layers, with each layer integrating two essential components: a multi-head self-attention mechanism (MHSA) as formulated in Equation (5) and a position-wise feed-forward network (FFN) as defined in Equation (6).

$$Z_l = LayerNorm(Z_{l-1} + MHSA(Z_{l-1})) \tag{5}$$

$$Z_l = LayerNorm(Z_l + FFN(Z_l)) \tag{6}$$

where Z(l-1) is the input to the l-th layer.

Classification Head: After processing through the Transformer layers, the embedding corresponding to the designated [CLS] token is extracted. This embedding serves as the definitive feature representation for downstream classification tasks. This classification-specific token is processed through a fully connected layer, where it undergoes transformation to generate the final class probability distribution.

Multi-Head Self-Attention (MHSA): The multi-head self-attention (MHSA) mechanism is a fundamental element of transformer architecture. It enables the model to concurrently attend to multiple segments of the input, thereby facilitating the capture of intricate dependencies and contextual relationships. Specifically, for any given input sequence, the MHSA mechanism computes attention weights that are subsequently used to form multiple attention heads. This parallel processing approach allows the model to extract a diverse range of contextual dependencies across various feature representations. These heads are subsequently concatenated and linearly transformed. Formally, for an input sequence $\boldsymbol{Z} \in \mathbb{R}^{NxD}$, the output of a single attention head is defined as Equation (7).

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{7}$$

In this framework, the matrices Q, K, and V correspond to the query, key, and value representations, respectively, while d_k denotes the dimensionality of the key vectors.

Feed-Forward Network (FFN): In Transformer architecture, each layer incorporates a position-wise feed-forward network (FFN) that is critical for modeling intricate dependencies within the input data. This network is structured as two successive fully connected layers, with a ReLU activation function interposed between them. The introduction of non-linearity via the ReLU function enhances the model's capacity to learn complex and high-level feature representations. This operation can be formally represented as Equation (8).

$$FFN(\boldsymbol{Z}) = \max(0, \boldsymbol{Z}W_1 + b_1)\,W_2 + b_2 \tag{8}$$

Here W_1, W_2, b_1, and b_2 are learnable parameters.

ViT architecture is configurable through several hyperparameters, notably the number of transformer layers (L), the number of attention heads, and the dimensionality of the embedding space (D). The resulting variants include, for example, the following configurations:

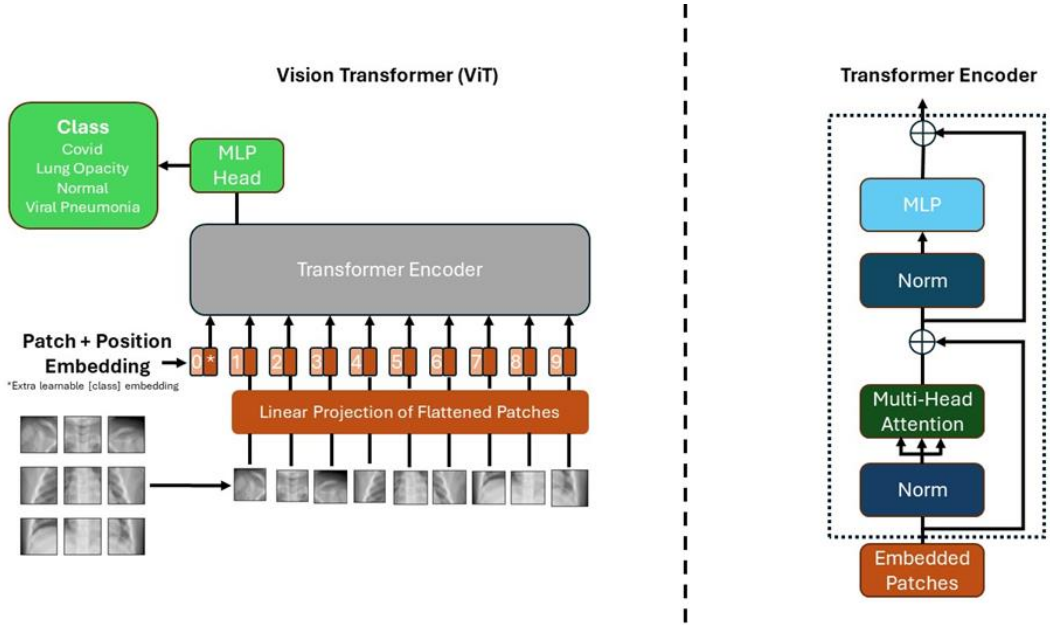

Figure 4. ViT architecture

ViT-Base (B/16): This configuration comprises 12 transformer layers, each equipped with 12 attention heads, and an embedding dimension of 768. It processes input images using patch sizes of 16×16.

ViT-Large (L/16): This model variant comprises 24 layers and 16 attention heads, featuring an embedding dimension of 1024. Additionally, it employs patch sizes of 16×16, facilitating enhanced spatial feature extraction and representation.

ViT-Huge (H/14): The most advanced configuration comprises 32 layers, 16 attention heads, and an embedding dimension of 1280, utilizing smaller patch sizes of 14×14 to enhance feature extraction and model performance.

The choice of ViT variant depends on the computational resources available and the complexity of the task at hand. Notably, ViTs typically require pretraining on large-scale datasets (e.g., ImageNet-21k) to achieve optimal performance. This necessity stems from the absence of inductive biases in ViTs that are inherently present in CNNs, including locality and translation equivariance.

2.3 Experimental setup

The experimental framework was implemented on the Google Colab platform, leveraging the computational capabilities of a Tesla T4 GPU equipped with 320 Turing Tensor Cores and 16 GB of GDDR6 VRAM. This setup was selected to ensure efficient training and inference of deep learning models. TensorFlow 2.15 was utilized as the primary deep learning framework, offering comprehensive support for the development, training, and optimization of advanced neural network architectures. Its robust computational capabilities facilitated efficient model implementation, enabling precise and scalable deep learning applications. The cloud-based environment facilitated high computational efficiency, reproducibility, and seamless collaboration among researchers.

The dataset was partitioned into training, validation, and testing sets using an 80:10:10 ratio to enable a rigorous and impartial assessment of model performance. To improve generalizability and counteract overfitting, various data augmentation strategies—such as image flipping, rotation, and zooming—were employed. Prior to model training, all input images were standardized to a resolution of 128×128×3 pixels. Model optimization was performed using the AdamW optimizer [24] with a learning rate of 0.0003 and a weight decay coefficient of 0.00003. Training was conducted with a batch size of 32 over 100 epochs, and a dropout rate of 0.1 was integrated into the network architecture to further mitigate the risk of overfitting. The learning process was guided by the sparse categorical cross-entropy loss function, which ensured stable convergence of the model. All hyperparameters were selected manually based on preliminary experiments.

3 Experimental Results

3.1 Evaluation metrics

A rigorous evaluation was conducted to benchmark the diagnostic performance of the ViT and the ResNet models (ResNet50, ResNet101, and ResNet152) in categorizing chest X-ray images into four classes— Normal, Lung Opacity, Viral Pneumonia, and COVID-19—using an extensive suite of evaluation metrics. These metrics provide a detailed and objective analysis of each model's classification efficacy, particularly within the field of medical image processing, where significant challenges—most notably, the prevalent issue of class imbalance—are routinely encountered.

Within the suite of evaluation metrics, accuracy is a principal measure, defined as the ratio of correctly classified observations to the total number of observations. This metric

offers a general performance indicator and is formally defined by Equation (9).

$$\text{Accuracy} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})} \quad (9)$$

True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) collectively define the classification outcomes in a predictive model. While accuracy serves as a fundamental performance metric, it can be misleading in scenarios involving imbalanced datasets, as it does not account for class distribution disparities.

Precision is a fundamental performance metric in classification tasks, defined as the ratio of true positive predictions to the total number of instances predicted as positive. This measure assesses the classifier's ability to accurately identify positive cases and is crucial for evaluating model performance, particularly in contexts where minimizing false positives is of paramount importance. It is formally defined in Equation (10).

$$\text{Precision} = \frac{\text{TP}}{(\text{TP} + \text{FP})} \quad (10)$$

Recall quantifies a model's effectiveness in identifying all actual positive instances within a dataset. More formally, it is defined as the ratio of true positive predictions to the total number of genuine positive cases, as delineated in Equation (11).

$$\text{Recall} = \frac{\text{TP}}{(\text{TP} + \text{FN})} \quad (11)$$

The F1-score, defined as the harmonic mean of precision and recall, offers a balanced evaluation of these two performance metrics. This balance renders it especially useful in applications where an optimal trade-off between precision and recall is imperative. Equation (12) formally delineates the definition of this metric.

$$\text{F1} - \text{Score} = \frac{2 * (\text{Precision} * \text{Recall})}{(\text{Precision} + \text{Recall})} \quad (12)$$

A confusion matrix serves as a robust tool for evaluating a model's predictive performance by systematically aligning actual class labels with those predicted by the model. In a four-class classification scenario, a 4×4 confusion matrix is employed, where each row denotes an actual class, and each column corresponds to a predicted class. This structured approach enables a thorough examination of class-specific accuracies, error distributions, and misclassification trends.

In multi-class classification scenarios, individual metrics—namely precision, recall, and F1-score—are computed for each class to evaluate the model's performance across distinct categories. To achieve a comprehensive performance assessment, these per-class metrics are commonly aggregated using two well-established methods:

Macro-Averaging: This method calculates the unweighted mean of the performance metrics across all classes, treating each class equally irrespective of its size or prevalence.

Weighted-Averaging: Here, the metrics for each class are weighted according to the proportion of true instances belonging to that class, ensuring a more representative measure for datasets with class imbalances.

Collectively, these evaluation metrics provide an integrated assessment of the model's performance across all four diagnostic categories, thereby enabling a more nuanced and precise analysis of its classification accuracy and overall efficacy.

Reporting both macro and weighted averages is particularly important in the presence of class imbalance, as it ensures that the evaluation reflects both per-class performance and the real-world prevalence of each diagnostic group.

3.2 Results

The performance of the models—ViT, ResNet50, ResNet101, and ResNet152—was meticulously evaluated on the test dataset using a diverse set of metrics, including accuracy, precision, recall, F1-score, and confusion matrices. This multifaceted evaluation framework enabled a thorough examination of each model's classification capabilities across four chest X-ray categories thereby providing a comprehensive assessment of their diagnostic performance.

To ensure clarity and systematic comparisons, the performance of each model is analyzed in dedicated subsections. The use of confusion matrices enables a detailed examination of misclassifications, thereby identifying areas where predictive accuracy can be improved. Such an approach ensures a rigorous and impartial assessment of the models' effectiveness, facilitating the identification of their respective strengths and limitations. This comprehensive analysis is crucial for refining model architectures and optimizing training methodologies to enhance performance in future applications.

3.2.1 ResNet-50

ResNet-50, a 50-layer deep residual network introduced by He et al. [21], mitigates the vanishing gradient issue by employing skip connections, thereby enhancing the training efficiency of deep neural architectures. Table 1 presents the model's performance on the test set by reporting essential evaluation metrics—precision, recall, and F1-score—across four chest X-ray diagnostic categories: COVID-19, Lung Opacity, Normal, and Viral Pneumonia. In addition, both macro-averaged and weighted-average metrics are provided to offer a comprehensive evaluation of the classifier's overall efficacy.

The model exhibited a notable performance across multiple evaluation metrics. Specifically, the model achieved a maximum precision of 96.7% for the Normal class and recorded its highest recall of 98.62% for the COVID-19 category, thereby demonstrating its effectiveness in identifying critical cases. Additionally, the Viral Pneumonia class attained the highest F1-score at 97.06%, reflecting an optimal balance between precision and recall. Overall, the model reached an accuracy of 89.69% on the test set, with macro-average and weighted-average F1-scores of

90.82% and 89.77%, respectively. These results underscore the robust classification capabilities of the ResNet-50 model in a multi-class diagnostic setting.

Table 1. ResNet-50 test results

	Test Data	Precision	Recall	F1-Score
COVID-19	363	79.56%	98.62%	88.07%
Lung Opacity	602	85.37%	89.20%	87.25%
Normal	1024	96.70%	85.74%	90.89%
Viral Pneumonia	135	96.35%	97.78%	97.06%
Macro avg	2124	89.49%	92.84%	90.82%
Weighted avg	2124	90.54%	89.69%	89.77%
Overall Accuracy	2124		89.69%	

To facilitate a more detailed evaluation of the model's performance, a confusion matrix was generated (see Figure 5). Among the 363 test images classified as COVID-19, only 5 instances were misclassified, reflecting a high degree of accuracy for this category. Similarly, the Viral Pneumonia class exhibited robust performance, with only 3 misclassifications out of 135 test images. In contrast, the Lung Opacity class demonstrated a higher error rate, with 65 out of 602 images misclassified. The Normal class also faced significant challenges, as 146 of the 1024 test images were incorrectly predicted. The experimental findings indicate that the model demonstrates robust performance in differentiating COVID-19 cases from viral pneumonia. However, the analysis also reveals that its classification accuracy for Normal and lung opacity images is suboptimal, suggesting the need for further refinement in these areas.

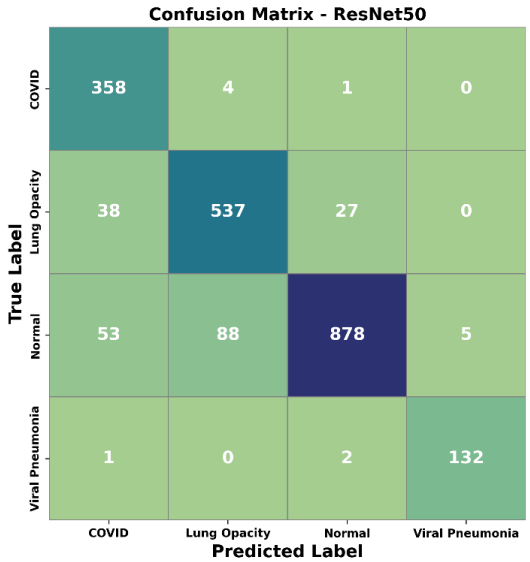

Figure 5. Confusion matrix for ResNet-50 test results

3.2.2 ResNet-101

Introduced by He et al. [21], ResNet-101 is an advanced deep convolutional neural network that extends the ResNet-50 architecture by incorporating 101 layers, thereby enhancing both depth and representational capacity. Table 2 presents a comprehensive evaluation of the model's performance on the test dataset, detailing precision, recall, and F1-score for four CXR categories. Moreover, the table summarizes the overall classification effectiveness through macro-averaged and weighted-average metrics.

In the evaluation, the model achieved a perfect precision of 100% for the Viral Pneumonia category, and a recall of 100% for the COVID-19 category, indicating its efficacy in accurately identifying all true instances of COVID-19 within the test dataset. Notably, the highest F1-score (70.94%) was observed for the Normal class. Despite these class-specific performances, the overall accuracy of ResNet-101 on the test set was 56.87%, significantly lower than the 89.69% accuracy obtained by ResNet-50.

Table 2. ResNet-101 test results

	Test Data	Precision	Recall	F1-Score
COVID-19	363	30.02%	100%	46.18%
Lung Opacity	602	88.98%	34.88%	50.12%
Normal	1024	93.03%	57.32%	70.94%
Viral Pneumonia	135	100%	35.56%	52.46%
Macro avg	2124	78.01%	56.94%	54.92%
Weighted avg	2124	81.56%	56.87%	59.63%
Overall Accuracy	2124		56.87%	

Figure 6 illustrates the confusion matrix corresponding to the ResNet-101 model. Notably, the model achieved a perfect recall (100%) for the COVID-19 class, accurately classifying all 363 test images in that category. However, a substantial number of misclassifications were observed among the other classes. Specifically, within the Lung Opacity class, 359 of the 602 images were incorrectly classified as COVID-19, while an additional 33 were misclassified as Normal. Similarly, for the Normal class, 411 out of 1024 images were erroneously predicted as COVID-19. In the case of the Viral Pneumonia class, 76 of the 135 images were misclassified as COVID-19.

This pattern of misclassification indicates a strong bias in the model towards overpredicting the COVID-19 class, which is reflected in the high recall (100%) but low precision (30.02%) for this category. The notably low overall accuracy (56.87%) compared to ResNet-50 suggests suboptimal generalization and points to possible issues such as overfitting, class imbalance, or inappropriate model hyperparameters (e.g., learning rate, batch size). These factors, along with potential configuration errors during model training, may have contributed to the observed performance drop and misclassification trends.

Conversely, the model demonstrated perfect precision for the Viral Pneumonia class (100%), attributable to the absence of any misclassification of non-Viral Pneumonia images as Viral Pneumonia.

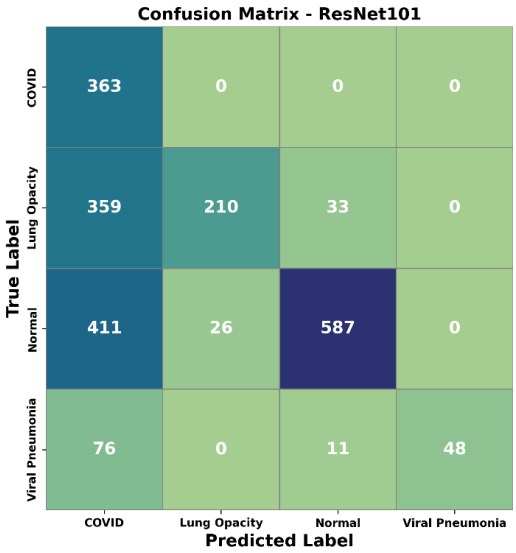

Figure 6. Confusion matrix for ResNet-101 test results

3.2.3 *ResNet-152*

ResNet-152, a 152-layer Residual Network introduced by He et al., represents the deepest architecture examined in this study [21]. Table 3 presents a comprehensive evaluation of the model's performance on the test set. Specifically, it reports precision, recall, and F1-scores for each of the four chest X-ray categories. Moreover, the table includes both macro-averaged and weighted-average metrics, offering a detailed overview of the model's diagnostic efficacy across these classes.

The model achieved the highest precision (100%) and F1-score (93.70%) for the Viral Pneumonia category, while the highest recall (95.61%) was observed for the Normal category. Overall, ResNet-152 achieved an accuracy of 87.66%, outperforming ResNet-101 (56.87%) but slightly trailing ResNet-50 (89.69%).

Table 3. ResNet-152 test results

	Test Data	Precision	Recall	F1-Score
COVID-19	363	99.17%	65.56%	78.94%
Lung Opacity	602	80.92%	87.38%	84.03%
Normal	1024	87.80%	95.61%	91.54%
Viral Pneumonia	135	100%	88.15%	93.70%
Macro avg	2124	91.97%	84.17%	87.05%
Weighted avg	2124	88.57%	87.66%	87.39%
Overall Accuracy	2124		87.66%	

Figure 7 presents the confusion matrix for the ResNet-152 model, summarizing its classification performance across four diagnostic categories. The model was evaluated on a test set comprising 363 COVID-19 images, correctly identifying 238 of them. In contrast, misclassifications included 74 images erroneously labeled as lung opacity and 51 images incorrectly categorized as normal. For the Lung Opacity category, 526 out of 602 images were accurately identified, with 75 misclassified as Normal and 1 as COVID-19. Regarding the Normal class, 979 of the 1024 test images were correctly classified, whereas 44 were erroneously assigned to Lung Opacity and 1 to COVID-19. Finally, for the Viral Pneumonia category, the model achieved correct classification for 119 of 135 images, while 6 were misclassified as Lung Opacity and 10 as Normal.

The model achieved a precision of 100% for the Viral Pneumonia class, which can be attributed to the absence of false positive predictions for this category, mirroring the performance observed with ResNet-101. Furthermore, the model demonstrated its most robust classification capability with the Normal class, as reflected in its recall rate of 95.61% and the highest count of correctly classified images.

3.2.4 *ViT*

ViT adopts a fundamentally distinct architectural approach compared to traditional CNN-based models. In this study, the ViT model was trained from scratch without the use of pre-trained weights, allowing the network to learn all relevant features directly from the chest X-ray dataset. Table 4 presents a detailed evaluation of the ViT model's performance on the test dataset. The table reports essential metrics for each of the four CXR categories, thereby providing a granular assessment of the model's diagnostic capabilities. Additionally, macro-averaged and weighted-average values for these metrics are reported to offer a comprehensive assessment.

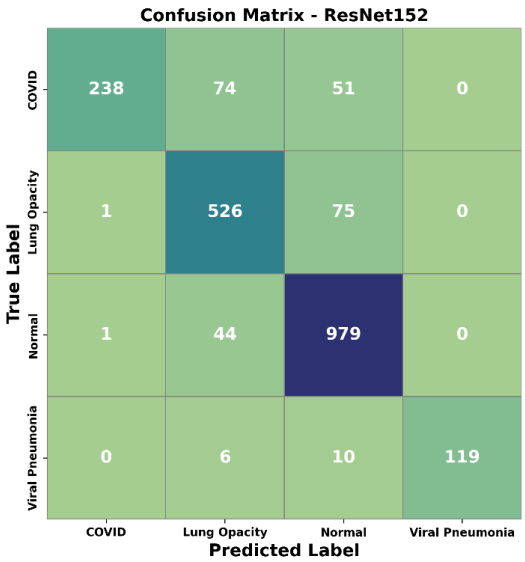

Figure 7. Confusion matrix for ResNet-152 test results

Among the evaluated classes, the COVID-19 category achieved the highest precision at 96.79%, while the Normal category demonstrated the highest recall at 93.95%. Notably, the Viral Pneumonia class attained the highest F1-score

(92.54%), underscoring the model's strong performance in accurately classifying this category. Among the models evaluated, ViT achieved the highest performance, recording a test set accuracy of 90.25%.

ViT employs a fundamentally different architecture from CNN-based models. Table 4 provides a comprehensive evaluation of the model's performance on the test set. Specifically, it details the precision, recall, and F1-score for each of the four chest X-ray categories while also reporting the corresponding macro and weighted averages of these metrics.

Table 4. ViT test results

	Test Data	Precision	Recall	F1-Score
COVID-19	363	96.79%	83.20%	89.48%
Lung Opacity	602	85.46%	87.87%	86.65%
Normal	1024	90.75%	93.95%	92.32%
Viral Pneumonia	135	93.23%	91.85%	92.54%
Macro avg	2124	91.56%	89.22%	90.25%
Weighted avg	2124	90.44%	90.25%	90.24%
Overall Accuracy	2124		90.25%	

Figure 8 presents a detailed confusion matrix for the ViT model. For the COVID-19 class, 302 of 363 test images were correctly classified, while misclassifications included 39 images labeled as Lung Opacity, 17 as Normal, and 5 as Viral Pneumonia. In the Lung Opacity category, 529 out of 602 images were correctly identified; however, 70 images were incorrectly classified as Normal, 2 as COVID-19, and 1 as Viral Pneumonia. Similarly, for the Normal class, 962 of 1024 images were correctly classified, with 51 images misidentified as Lung Opacity, 8 as COVID-19, and 3 as Viral Pneumonia. Lastly, for the Viral Pneumonia class, 124 of 135 images were correctly classified, with the remaining 11 images misclassified as Normal.

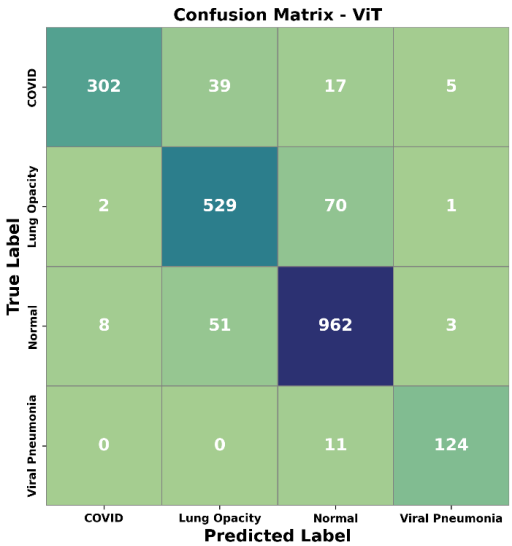

Figure 8. Confusion matrix for ViT test results

ViT model demonstrated superior efficacy in classifying viral pneumonia, achieving an F1-score of 92.54%. The confusion matrix underscores the model's overall robust classification ability, effectively differentiating between the four CXR categories.

4 Discussion

An evaluation on the test image dataset indicates that ViT model achieved the highest classification accuracy, reaching 90.25%. In contrast, among the ResNet-based architectures, ResNet-50 demonstrated the best performance with an accuracy of 89.69%, while ResNet-101 and ResNet-152 attained accuracies of 56.87% and 87.66%, respectively. These findings suggest that increasing the depth of the network does not inherently lead to improved accuracy, thereby underscoring the potential efficacy of shallower architectures in certain scenarios.

To validate the reliability of these results, a comparative analysis with existing studies is presented in Table 5. The ViT model consistently outperformed other models utilized in multi-class and binary classification tasks across prior research. Oh et al. [11] documented a classification accuracy of 88.9% when distinguishing among five categories—normal, bacterial, tuberculosis, viral, and COVID-19 images. In a related study, Butt et al. [12] achieved an accuracy of 86.7% in differentiating COVID-19, Influenza A, and normal images. Similarly, Khan et al. [13] reported an accuracy of 89.6% for categorizing images into COVID-19, normal, viral pneumonia, and bacterial pneumonia classes. Furthermore, Xu et al. [14] attained an accuracy of 86.7% in discriminating between COVID-19, IAVP, and normal images, while Shadin et al. [15] observed an accuracy of 85.94% in the binary classification of COVID-19 versus normal images. More recently, Park et al. [16] developed a robust ViT framework and achieved average accuracy scores of 86.4%, 85.9%, and 85.2% across three independent institutional test sets (CNUH, YNU, and KNUH, respectively) for classifying normal, other infections, and COVID-19 cases, surpassing the performance of conventional models such as ResNet-50 and standard ViT architectures.

The collective findings underscore the superior performance of the ViT model, which consistently achieves higher classification accuracy than both the evaluated ResNet architectures and those documented in previous studies. The model's reliability is quantitatively supported by high accuracy, precision, recall, and F1-score values across all classes in both internal and comparative evaluations. While direct inference speed was not measured, the ViT architecture is recognized in the literature for its computational efficiency and suitability for real-time clinical applications, further supporting its potential as a rapid and reliable diagnostic tool for CXR image analysis. This reinforces the effectiveness of transformer-based architectures for CXR image classification, further validating the robustness of the study's findings within the broader context of the literature.

Table 5. Comparison of studies

Studies	Models	Accuracy	Precision	Recall	F1-Score
[11]	ResNet-18	88.9%	83.4%	85.9%	84.4%
[12]	ResNet-18	86.7%	81.3%	86.7%	83.9%
[13]	CoroNet	89.6%	89.84%	89.93%	89.82%
[14]	ResNet-18	86.7%	86.86%	86.66%	86.7%
[15]	InceptionV3	85.94%	87.5%	76.24%	81.48%
[16]	ViT	86.4%	-----	87.0%	-----
This Study	ResNet-50	89.69%	89.49%	92.84%	90.82%
	ResNet-101	56.87%	78.01%	56.94%	54.92%
	ResNet-152	87.66%	91.97%	84.17%	87.05%
	ViT	**90.25%**	**91.56%**	**89.22%**	**90.25%**

5 Conclusion and suggestions

Despite a marked decline in COVID-19 incidence, the disease continues to pose a significant global health challenge, underscoring the need for rapid and reliable diagnostic strategies to improve patient outcomes. In this context, our study investigates the utility of CXR imaging as an efficient, cost-effective, and widely accessible alternative to traditional diagnostic techniques such as RT-PCR and CT. We assembled a comprehensive dataset comprising 21,165 images, which included 3,616 confirmed COVID-19 cases, 6,012 instances exhibiting lung opacity, 10,192 normal cases, and 1,345 cases of viral pneumonia. The diagnostic performance of four state-of-the-art machine learning models—ResNet-50, ResNet-101, ResNet-152, and ViT—was systematically evaluated using this dataset.

Among the models evaluated, ViT demonstrated the most robust performance. On the test set, ViT achieved an accuracy of 90.25%, a precision of 91.56%, a recall of 89.22%, and an F1-score of 90.25%. These results underscore the efficacy of the ViT architecture in accurately capturing the underlying patterns of the data compared to its counterparts. Comparative analysis with prior research confirmed that the ViT model outperformed other approaches across binary, three-class, four-class, and five-class classification tasks, thereby underscoring the robustness and reliability of the proposed method.

In addition to its strong quantitative performance, the proposed AI-based framework can be readily integrated into clinical workflows as a decision support system. For instance, in emergency departments or triage settings, the model can rapidly analyze incoming chest X-rays to assist healthcare professionals in the early detection and differentiation of COVID-19, viral pneumonia, lung opacity, and normal cases. Such real-time integration can facilitate prompt isolation and treatment decisions, alleviate the diagnostic workload for radiologists, and ultimately improve patient outcomes, particularly in high-demand or resource-constrained environments.

Future investigations will concentrate on bolstering classification accuracy through the integration of state-of-the-art data preprocessing methodologies and the meticulous optimization of model hyperparameters. This multifaceted approach is expected to significantly enhance the diagnostic performance of AI-driven systems in chest X-ray image analysis.

Conflict of Interest

The authors declare that they have no conflict of interest.

Similarity Rate (iThenticate): %15

References

[1] WHO, WHO Director-General's opening remarks at the media briefing on COVID-19. March 2020. Available: https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020

[2] WHO, WHO COVID-19 dashboard-cases. 30 July 2024. Available: https://data.who.int/dashboards/covid19/cases?n=o

[3] WHO, WHO COVID-19 dashboard-deaths. 30 July 2024. Available: https://data.who.int/dashboards/covid19/deaths?n=o

[4] R. T. Gandhi, J. B. Lynch, and C. Del Rio, Mild or moderate Covid-19. New England Journal of Medicine, 383 (18), 1757-1766, 2020.

[5] J. Hellewell, S. Abbott, A. Gimma, N. I. Bosse, C. I. Jarvis, T. W. Russell, et al., Feasibility of controlling COVID-19 outbreaks by isolation of cases and contacts. The Lancet Global Health, 8 (4), e488-e496, 2020. https://doi.org/10.1016/S2214-109X(20)30074-7

[6] E. J. Emanuel, G. Persad, R. Upshur, B. Thome, M. Parker, A. Glickman, et al., Fair allocation of scarce medical resources in the time of Covid-19. New England Journal of Medicine, 382, 2049-2055, 2020.

[7] C. Dong, S. Cao, and H. Li, Young children's online learning during COVID-19 pandemic: Chinese parents' beliefs and attitudes. Children and Youth Services Review, 118, 105440, 2020. https://doi.org/10.1016/j.childyouth.2020.105440

[8] G. Liu and J. F. Rusling, COVID-19 antibody tests and their limitations. ACS Sensors, 6 (3), 593-612, 2021. https://doi.org/10.1021/acssensors.0c02621

[9] W.-C. Dai, P. Zhang, H. Wang, X. Cheng, L. Xu, and Y. Yin, CT imaging and differential diagnosis of COVID-19. Canadian Association of Radiologists Journal, 71 (2), 195-200, 2020. https://doi.org/10.1177/0846537120913033

[10] F. Shi, L. Xia, F. Shan, B. Song, D. Wu, Y. Wei, et al., Large-scale screening to distinguish between COVID-19 and community-acquired pneumonia using infection size-aware classification. Physics in Medicine & Biology, 66 (6), 065031, 2021. https://doi.org/10.1088/1361-6560/abe838

[11] Y. Oh, S. Park, and J. C. Ye, Deep learning COVID-19 features on CXR using limited training data sets. IEEE Transactions on Medical Imaging, 39 (8), 2688-2700, 2020. https://doi.org/10.1109/TMI.2020.2993291

[12] C. Butt, J. Gill, D. Chun, and B. A. Babu, Deep learning system to screen coronavirus disease 2019 pneumonia. Applied Intelligence, 50, 3622-3634, 2020. https://doi.org/10.1007/s10489-020-01714-3

[13] A. I. Khan, J. L. Shah, and M. M. Bhat, CoroNet: A deep neural network for detection and diagnosis of

COVID-19 from chest x-ray images. Computer Methods and Programs in Biomedicine, 196, 105581, 2020. https://doi.org/10.1016/j.cmpb.2020.105581

[14] X. Xu, X. Jiang, C. Ma, P. Du, X. Li, S. Lv, et al., A deep learning system to screen novel coronavirus disease 2019 pneumonia. Engineering, 6 (10), 1122-1129, 2020. https://doi.org/10.1016/j.eng.2020.04.010

[15] N. S. Shadin, S. Sanjana, and N. J. Lisa, COVID-19 diagnosis from chest X-ray images using convolutional neural network (CNN) and InceptionV3. International Conference on Information Technology (ICIT), pp. 799-804, Amman, Jordan, 14-15 July 2021.

[16] S. Park, G. Kim, Y. Oh, J. B. Seo, S. M. Lee, J. H. Kim, et al., Vision transformer for covid-19 cxr diagnosis using chest x-ray feature corpus. arXiv preprint arXiv:2103.07055, 2021.

[17] S. Cannata, C. Paschero, M. Enescu, F. A. Fiorini, and M. Panella, Deep learning algorithms for automatic COVID-19 detection on chest X-ray images. IEEE Access, 10, 119905-119913, 2022. https://doi.org/10.1109/ACCESS.2022.3221531

[18] M. E. Chowdhury, T. Rahman, A. Khandakar, R. Mazhar, M. A. Kadir, Z. B. Mahbub, et al., Can AI help in screening viral and COVID-19 pneumonia? IEEE Access, 8, 132665-132676, 2020. https://doi.org/10.1109/ACCESS.2020.3010287

[19] T. Rahman, A. Khandakar, Y. Qiblawey, A. Tahir, S. Kiranyaz, S. B. Abul Kashem, et al., Exploring the effect of image enhancement techniques on COVID-19 detection using chest X-ray images. Computers in Biology and Medicine, 132, 104319, 2021. https://doi.org/10.1016/j.compbiomed.2021.104319

[20] Kaggle, COVID-19 Radiography Database. https://www.kaggle.com/datasets/tawsifurrahman/covid19-radiography-database/data, Accessed 25 June 2025.

[21] K. He, X. Zhang, S. Ren, and J. Sun, Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770-778, Las Vegas, NV, USA, 27-30 June 2016.

[22] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, et al., An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020. https:/doi.org/10.48550/arXiv.2010.11929

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, et al., Attention is all you need. arXiv preprint arXiv:1706.03762, 2017. https:/doi.org/10.48550/arXiv.1706.03762

[24] I. Loshchilov and F. Hutter, Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101, 2017. https:/doi.org/10.48550/arXiv.1711.05101