Available online at www.HighTechJournal.org

HighTech and Innovation Journal

ISSN: 2723-9535

Vol. 7, No. 1, March, 2026

A Novel Hybrid ViT-CNN Approach for Pneumonia and Lung Opacity Detection in X-Ray Images

Taqwa Hariguna [1], Athapol Ruangkanjanases [2*]

[1] *Department of Information System and Magister Computer Sciences, Universitas Amikom Purwokerto, Purwokerto Utara, 53127, Indonesia.*

[2] *Department of Commerce, Chulalongkorn Business School, Chulalongkorn University, Bangkok, 10330, Thailand.*

Received 26 September 2025; Revised 08 February 2026; Accepted 11 February 2026; Published 01 March 2026

Abstract

In order to automatically classify chest X-ray pictures into three diagnostic categories—Normal, Lung Opacity, and Viral Pneumonia—this study presents a novel hybrid deep learning architecture that combines the Vision Transformer (ViT) with a Convolutional Neural Network (CNN). The suggested model successfully addresses the drawbacks of single-architecture systems by fusing the ResNet-18 CNN's expertise in local texture analysis with the ViT's global feature representation capability. According to experimental assessments, the hybrid ViT-CNN architecture outperforms the state-of-the-art methods, achieving 94.2% classification accuracy with precision, recall, and F1-scores continuously above 94% for the majority of categories. Even in complicated situations where traditional methods usually falter, like distinguishing between lung opacity and normal patients, the model exhibits strong performance. Additionally, it performs well in discrimination, with AUC values above 0.95 in every class. The system is ideal for real-time clinical deployment because it maintains a high computational efficiency, generating conclusions in about 0.0012 seconds per image. Grad-CAM visualization makes it evident which areas of the image are important for making diagnostic decisions, hence validating the model's interpretability. All things considered, this work establishes a new benchmark for chest X-ray classification performance and offers a useful foundation for automated diagnostic assistance in resource-constrained healthcare settings.

Keywords: Automated Medical Imaging; Chest X-Ray Classification; Convolutional Neural Network; Grad-CAM Visualization; Hybrid Deep Learning Architecture.

1. Introduction

Lung-related conditions such viral respiratory infections, pneumonia, and lung opacity continue to be major sources of morbidity and mortality worldwide, with low- and middle-income countries bearing the brunt of this burden. The World Health Organization (WHO) reports that pneumonia alone kills hundreds of thousands of people annually, with older people and children under five being the most affected. These illnesses frequently present with overlapping symptoms, such as fever, cough, and dyspnea, making clinical distinction challenging and highlighting the critical role that medical imaging plays in precise diagnosis. Thus, immediate care, minimizing disease development, preventing transmission, and ultimately lowering mortality all depend on accurate and quick detection.

Because of its low cost, quick picture acquisition, and non-invasive nature, chest X-rays (CXRs) remain one of the most widely available and economical diagnostic techniques for assessing pulmonary disorders. CXRs are the first imaging modality used in the majority of medical facilities to evaluate respiratory problems. Despite these benefits,

* Corresponding author: athapol@cbs.chula.ac.th

https://doi.org/10.28991/HIJ-2026-07-01-09

radiologists continue to have diagnostic variability, and reading CXRs is still a challenging process requiring a high level of competence. Factors such as subtle lesions, overlapping anatomical structures, and uneven imaging quality usually contribute to errors and inter-observer discrepancies [1, 2]. As a result, the need for computational technologies that can assist radiologists and improve diagnostic consistency is growing.

Medical image analysis has changed in recent years due to artificial intelligence (AI), especially deep learning. Convolutional Neural Networks (CNNs), one of the many deep learning models, have demonstrated a remarkable capacity to automatically extract hierarchical visual features from unprocessed image data. CNNs have been successfully used in a variety of radiological applications, such as lung nodule identification, tuberculosis detection, and multi-class disease classification utilizing radiographs [3, 4]. Their ability to recognize localized visual cues like edges, textures, and forms is a major factor in their efficacy. However, the inability of traditional CNN architectures to capture global contextual information and long-range spatial correlations becomes crucial when examining chest radiographs with mild or diffuse abnormalities dispersed throughout the lung fields [5].

Vision Transformers (ViTs) have become a viable substitute to address these issues. ViTs use a self-attention mechanism to describe interactions between all regions in an image, and they are based on the Transformer architecture for natural language processing. Unlike CNNs that rely on local receptive fields, ViTs can distinguish holistic anatomical and pathological patterns because they can learn complicated global dependencies [6, 7]. Across a variety of benchmark datasets, ViTs have demonstrated impressive performance in tasks like segmentation, object detection, and picture classification [8, 9]. However, because of their high processing cost and reliance on sizable annotated datasets for optimal performance, their use in medical imaging is still restricted [10].

Hybrid deep learning architectures have been created to take advantage of both CNNs' and ViTs' complementary qualities. These models seek to integrate the ViT's capacity to model global contextual representations with the CNN's capacity for fine-grained local feature extraction. Such hybrid networks have demonstrated significant advancements in domains such as face recognition, natural picture recognition, and autonomous vehicle perception [11, 12]. However, their use in medical imaging, specifically in CXR classification, remains very limited [13, 14]. To evaluate their clinical feasibility and generalization in actual diagnostic settings, more research is necessary.

This paper proposes a unique hybrid deep learning model for automated chest X-ray classification that combines a Vision Transformer (ViT) with a Convolutional Neural Network (CNN) based on ResNet-18. Three clinically significant classes, Normal, Lung Opacity, and Viral Pneumonia, are used by the program to classify CXR pictures. By simultaneously feeding images into pretrained ViT and CNN branches, extracting independent feature vectors, and combining them through a specially designed fully connected fusion layer, the framework integrates global and local feature representations. Because of this integration, the model can capture both intricate structural details and broad pathophysiological patterns, which is especially useful for differentiating between illnesses that look similar, including viral pneumonia and lung opacity.

A benchmark, publicly accessible, radiologist-annotated CXR dataset was used to train and test the hybrid model in order to gauge its performance. With F1-scores above 94 percent in the majority of classes, the system's validation accuracy was 94.24 percent. The hybrid architecture outperformed its solo CNN and ViT equivalents in difficult classification tasks, particularly distinguishing between Lung Opacity and Normal instances, a situation that frequently leads to diagnostic misunderstanding [15]. The model also demonstrated remarkable computing efficiency, producing predictions in 0.0012 seconds per image, supporting its suitability for real-time clinical usage [16, 17].

All things considered, this study advances the science in a number of significant ways. It starts by presenting a specially designed ViT-CNN hybrid framework that is ideal for analyzing chest X-rays. Second, it provides an empirical comparison utilizing a variety of performance parameters, including precision, recall, and inference time, against baseline and state of the art approaches. Third, it assesses clinical viability by analyzing model interpretability and computational efficiency. Lastly, it improves our present knowledge of transformer-based architectures in medical imaging, especially in settings with limited resources.

2. Literature Review

The prevalence of pulmonary diseases including pneumonia, tuberculosis, and viral infections around the world has made the automated classification of chest X-ray (CXR) pictures a crucial field of research in artificial intelligence (AI) for healthcare. Many deep learning techniques have been investigated to improve diagnosis precision and lessen dependency on skilled radiologists, particularly in low-resource medical settings. Convolutional Neural Networks (CNNs) have long had a dominant position among them because of their remarkable capacity to identify specific spatial patterns in medical images. Long-range dependencies, which are essential for comprehending complicated medical features and the global anatomical context, are frequently difficult for conventional CNN architectures to model.

Vision Transformers (ViTs) have emerged as a potent substitute to get around this restriction. In contrast to CNNs, ViTs make use of self-attention mechanisms to create global contextual associations in a picture, which allows for a

more thorough comprehension of patterns that are dispersed throughout space. Notwithstanding their advantages, ViTs usually require sizable labeled datasets and significant computational power, which may limit their use in clinical settings with constrained hardware and data resources.

In order to use the advantages of both paradigms, CNNs for fine-grained local feature extraction and ViTs for capturing global semantic context, researchers have recently developed hybrid deep learning architectures that combine CNNs and ViTs. When compared to single-model methods, these hybrid frameworks have continuously shown better performance, especially in medical imaging tasks that call for both contextual awareness and local detail.

For example, a hybrid architecture that incorporates a ViT module into the CheXNet pipeline for multi-label chest pathology classification has been proposed, achieving superior performance compared to conventional CNN-based methods, particularly in cases with multiple overlapping lung anomalies, with a mean AUC of 0.838 on the ChestX-ray14 dataset [18]. Similarly, CNN and ViT hybrid models have been used to compare the classification of pneumonia and COVID-19 on the COVID-QU-Ex dataset, showing improved robustness when trained on limited or unbalanced datasets and higher accuracy compared to ViT-only methods [19]. In a different study, a two-stage hybrid framework combining ViT-B16 and ResNet-50 was developed to categorize bacterial pneumonia, viral pneumonia, and tuberculosis, demonstrating strong potential for handling complex multi-class diagnostic problems with a classification accuracy of 96.18 percent [20]. In terms of efficiency, a lightweight hybrid model combining CNN backbones with Swin Transformer modules achieved an accuracy of 98.72 percent and an F1-score of 0.9872, indicating that high diagnostic precision can be preserved with low computational cost [21].

A CNN and ViT hybrid architecture was also applied to a multi-class, multi-label tuberculosis dataset containing 14 different lung abnormalities. Despite significant data imbalance, the model showed good generalization performance, highlighting the scalability of hybrid models for large and diverse medical datasets [22]. A MobileViT and CNN hybrid model for COVID-19 detection achieved 98.75 percent accuracy using only 1.4 million parameters, demonstrating suitability for real-time deployment in resource-limited environments [23]. In addition to hybrid frameworks, several CNN-only architectures such as DenseNet121, VGG16, and InceptionV3 have been widely used in CXR research. A DenseNet121-based system achieved radiologist-level performance for pneumonia detection [24]. Despite their success in capturing localized visual features, CNN-only models often struggle with holistic contextual reasoning, limiting their effectiveness for subtle or diffuse disease patterns. ViT-only architectures such as Swin Transformer and ViT-B/16 have shown strong performance on large-scale benchmarks, but their high computational demands and reliance on large datasets remain barriers to clinical adoption [25].

Overall, hybrid ViT and CNN models offer a balanced trade-off between accuracy, interpretability, and computational efficiency, making them well suited for real-time clinical diagnostic workflows. However, most prior studies have focused on binary classification or single-disease detection, while relatively few have addressed multi-class classification involving Normal, Lung Opacity, and Viral Pneumonia. Moreover, only a limited number of works have jointly considered diagnostic accuracy, interpretability, and inference efficiency within a single hybrid framework.

The current study proposes a novel hybrid Vision Transformer and CNN-based model specifically designed for multi-class chest X-ray classification to address these gaps. The framework aims to efficiently integrate global contextual understanding with detailed local feature extraction to achieve high diagnostic accuracy and computational efficiency suitable for real-world clinical applications.

3. Research Methodology

3.1. Dataset Description

The dataset used in this study consists of chest X-ray (CXR) images categorized into three clinically relevant classes, namely Normal, Lung Opacity, and Viral Pneumonia. This dataset is widely used for benchmarking deep learning models in medical imaging and was obtained from a publicly accessible source [26]. The dataset contains a total of 3,175 images, comprising 2,480 images for training and 695 images for validation, with a balanced distribution across the three classes (Table 1).

Table 1. Dataset distribution

Set	Class	Samples
Training Set	Viral Pneumonia	875
	Lung Opacity	930
	Normal	975
Validation Set	Lung Opacity	195
	Normal	275
	Viral Pneumonia	225

To ensure compatibility with the input size requirements of both the Vision Transformer (ViT) and Convolutional Neural Network (CNN), all images were resized to 224 × 224 pixels. The dataset annotations were validated by experienced radiologists and have been used in multiple prior studies for pneumonia diagnosis and chest disease classification [27, 28].

3.2. Data Preprocessing

Data preprocessing was performed to enhance model robustness and generalization. The preprocessing pipeline consisted of three main stages, namely normalization, data augmentation, and dataset splitting. Image normalization was conducted using the standard ImageNet mean values [0.485, 0.456, 0.406] and standard deviations [0.229, 0.224, 0.225] [29]. This step ensured that the input images were scaled consistently and aligned with the expectations of pretrained deep learning models. To increase data diversity and reduce overfitting, several data augmentation techniques were applied to the training set. These techniques included random brightness adjustments within plus minus 20 percent, horizontal flips with a probability of 0.5, and random rotations between minus 15 and plus 15 degrees [30]. Such augmentations improve the model's resilience to variations in acquisition conditions, illumination, and image orientation. The dataset was split into training and validation subsets using an 80-20 ratio with stratified sampling [31]. This approach preserved class balance across both subsets, ensuring equal representation of Normal, Lung Opacity, and Viral Pneumonia classes during training and evaluation.

3.3. Proposed Hybrid Model Architecture

A The proposed model employs a hybrid deep learning architecture that integrates a Vision Transformer (ViT) with a ResNet-18 CNN to leverage their complementary strengths. While ResNet-18 excels at extracting localized spatial features, ViT is effective in modeling global contextual dependencies across the entire image.

The Vision Transformer processes images by dividing them into fixed-size patches and treating each patch as a token in a transformer encoder [32]. In this study, the vit_base_patch16_224 configuration was adopted. Each 224 x 224 image was partitioned into 16 × 16 patches, resulting in 196 tokens. These patches were flattened and projected into a 768-dimensional embedding space. Using pretrained weights from the ImageNet-21k dataset, the ViT branch outputs a global feature vector with dimensionality d_{ViT} = 768, denoted as F_{ViT}.

In parallel, the CNN branch utilizes ResNet-18, which incorporates residual connections to mitigate the vanishing gradient problem and facilitate effective deep network training [33]. The final fully connected layer of ResNet-18 was replaced with an identity mapping, allowing the network to act solely as a feature extractor. This configuration produces a local feature vector with dimensionality d_{CNN} = 512, denoted as F_{CNN}.

The global and local feature representations were concatenated to form a unified feature vector:

$$F_{combined} = [F_{ViT}, F_{CNN}] \tag{1}$$

Here, $F_{combined}$ has a total dimensionality of $d_{combined} = d_{ViT} + d_{CNN} = 768 + 512 = 1280$.

A fully connected network, represented by the symbol f_{FC}, which consists of two dense layers with ReLU activation functions and dropout regularization to avoid overfitting, is applied to the combined feature vector $F_{combined}$. To calculate the probabilities for each class, the output from the fully connected layers is fed into a softmax function. The last forecast is provided by:

$$y_{hat} = softmax(W * F_{output} + b) \tag{2}$$

where y_{hat} is the vector of predicted probabilities for the three classes, W represents the weight matrix of the output layer, and b represents the bias term.

The cross-entropy loss function is used to optimize the hybrid model, expressed as:

$$L = -\left(\frac{1}{N}\right) * \Sigma_{i=1\ to\ N}\ \Sigma_{j=1\ to\ C}\ \left[y_{ij} * log\left(y_{hat_{ij}}\right)\right] \tag{3}$$

Here, N is the batch size, C is the number of classes (C = 3 in this study), y_{ij} is the ground truth label for the ith sample and $jth\ class$, and $y_{hat_{ij}}$ is the predicted probability for the ith sample and jth class.

By combining ViT and ResNet-18, the hybrid architecture captures both global anatomical structures and fine-grained local abnormalities, such as opacities and nodular patterns. The model achieved an average inference time of 0.0012 seconds per image, indicating strong potential for real-time clinical deployment [34, 35].

The whole architecture of the suggested ViT–CNN hybrid model, which shows how both routes are integrated, is shown in Figure 1. The ViT's global features and ResNet-18's localized features are concatenated and processed through

fully connected layers for final classification, as shown in the figure. A good compromise between fine-grained pattern recognition and global context understanding is offered by this dual-path architecture.

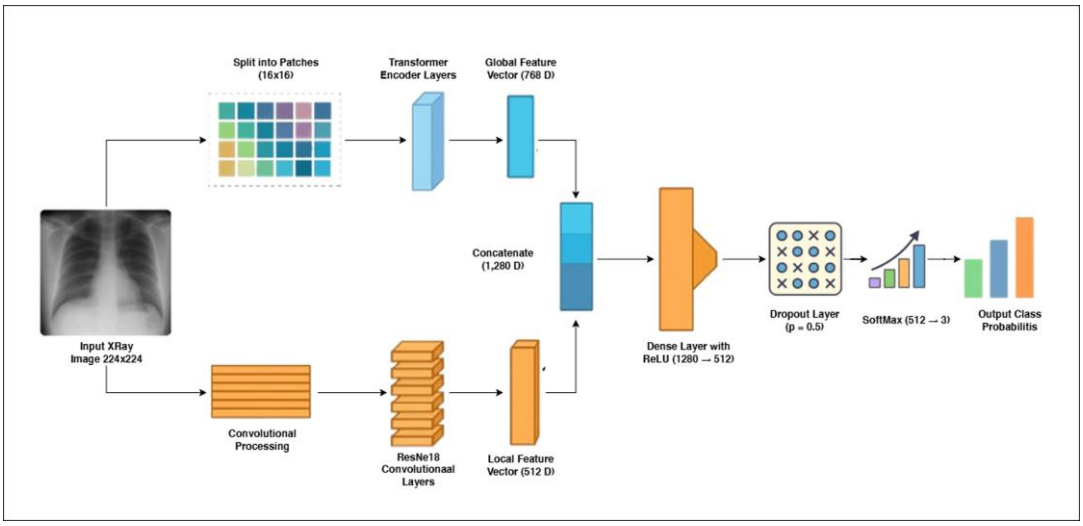

Figure 1. Proposed hybrid model ViT-CNN

3.4. Implementation Details

The proposed model was implemented using the PyTorch deep learning framework. Pretrained weights for ViT and ResNet-18 were obtained from the timm and torchvision libraries [36]. The categorical cross-entropy loss function was employed during training [37].

$$L = -\left(\frac{1}{N}\right) * \Sigma_{\{i=1\}}^{N} \Sigma_{\{j=1\}}^{C} \left(y_{ij} * log\left(y_{hat_{ij}}\right)\right) \quad (4)$$

where, N is the batch size, $C = 3$ is the number of classes, y_{ij} is the ground truth label, and $y_{hat_{ij}}$ is the predicted probability.

Model optimization was performed using the Adam optimizer with a learning rate of 1e-4 and momentum parameters beta1 = 0.9 and beta2 = 0.999 [38]. Dropout regularization with a probability of 0.5 was applied to the fully connected layers to mitigate overfitting [39]. All experiments were conducted on an NVIDIA Tesla V100 GPU with 32 GB of memory.

3.5. Training Configuration

PyTorch was used to create the hybrid model, which made use of pretrained weights from the torchvision and timm libraries [40]. To ensure computational performance, all experiments were carried out on an NVIDIA Tesla V100 GPU with 32 GB of memory. Both the training and validation datasets were trained using a batch size of 32 [41]. The Adam optimizer, which has a learning rate of 1e-4, was used to optimize the model. During backpropagation, the weight updates were stabilized using the default momentum parameters, beta1 = 0.9 and beta2 = 0.999 [42]. For multi-class classification problems, the cross-entropy loss function was utilized, which is expressed as follows:

$$L = -\left(\frac{1}{N}\right) * \Sigma_{i=1\ to\ N}\ \Sigma_{j=1\ to\ C} \left[y_{ij} * \log\left(y_{hat_{ij}}\right)\right] \quad (5)$$

Here, N is the batch size, C represents the number of classes (C = 3 in this study), y_{ij} denotes the true label for the $i-th$ sample and $j-th$ class, and $y_{hat_{ij}}$ represents the predicted probability for the $i-th$ sample and $j-th$ class [43].

Ten epochs were used to train the model. To lower the chance of overfitting, dropout regularization was applied to the completely linked layers with a probability of 0.5 [44]. Because it offered consistent convergence throughout all epochs, a set learning rate was kept constant during the training process [45].

3.6. Dataset Preparation

An 80-20 split of the dataset was made into training and validation subsets. In order to maintain a uniform class distribution across both subsets, stratified sampling was employed [46]. To boost the training set's variability and strengthen the model's resilience, data augmentation approaches were used. Random rotations between -15 and 15 degrees, 50% chance horizontal flips, and brightness modifications within ±20% were among them [47].

All photos were scaled to 224 by 224 pixels in order to comply with the Vision Transformer and ResNet-18 architectures' input size criteria [48]. Using the ImageNet mean values [0.485, 0.456, 0.406] and standard deviation values [0.229, 0.224, 0.225], image normalization was carried out.

3.7. Evaluation Metrics

The model performance was assessed using standard classification metrics, including accuracy, precision, recall, and F1-score. These metrics are defined as follows:

Accuracy (A) is computed as [49]:

$$A = (Number\ of\ Correct\ Predictions)/(Total\ Number\ of\ Predictions) \tag{6}$$

Precision (P), recall (R), and F1-score (F1) are defined as:

$$P = TP/(TP\ FP) \tag{7}$$

$$R = TP/(TP + FN) \tag{8}$$

$$F1 = 2 * (P * R)/(P + R) \tag{9}$$

In this case, FP stands for false positives, FN for false negatives, and TP for true positives. In order to thoroughly examine the model's classification performance, these metrics were calculated for each class separately (Normal, Lung Opacity, and Viral Pneumonia).

To assess inter-class misclassifications and display classification mistakes, a confusion matrix was created. To evaluate the model's computational effectiveness and applicability for real-time clinical applications, the inference time per image was also assessed.

3.8. Experimental Objectives

Validating the efficacy of the hybrid Vision Transformer and ResNet-18 architecture in classifying chest X-ray pictures was the main goal of the experimental setting. In order to compare the model's performance to cutting-edge techniques, this involved evaluating its accuracy, precision, recall, and F1-score. Assessing the model's inference time in order to ascertain its suitability for use in real-time clinical settings was another important goal.

4. Results and Discussion

4.1. Training and Validation Performance

Over ten epochs, the Vision Transformer (ViT), CNN (ResNet-18), and Hybrid ViT-CNN models' training and validation performances were assessed. With a training accuracy of 94.2% and a validation accuracy of 94.1%, the hybrid model performed the best. In contrast, the CNN and ViT models had validation accuracies of 87.4% and 91.5% and training accuracies of 89.6% and 91.2%, respectively. Figures 2, 3, and 4, which display the training progress for each model, provide examples of these trends.

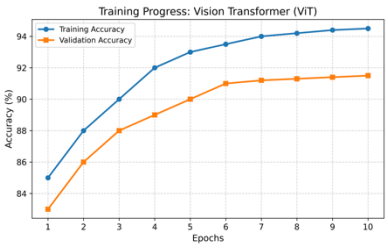

Figure 2. Training progress: vision transformer (ViT)

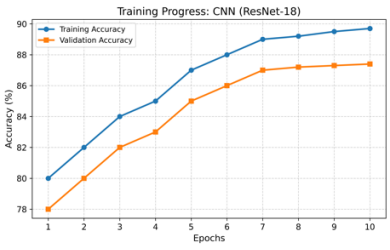

Figure 3. Training progress: CNN (ResNet-18)

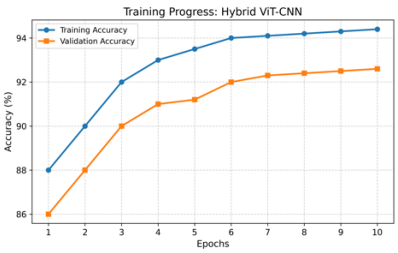

Figure 4. Training progress: hybrid ViT-CNN

Figure 5 shows that during training, the hybrid model maintained reduced training and validation losses and converged more quickly. This is due to its capacity to combine local information from ResNet-18 and global features from ViT, allowing for a more thorough analysis of the data [24, 49]. All models did, however, show slight variations in validation loss, which may indicate overfitting. This problem might be resolved by early stopping techniques or more regularization.

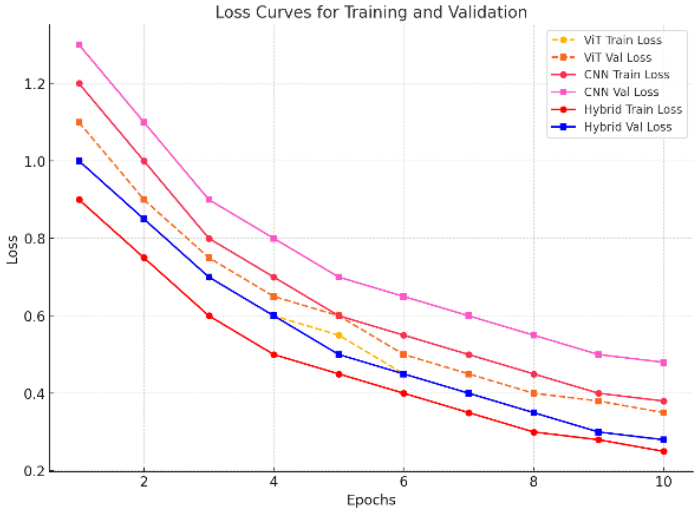

Figure 5. Loss curves for training and validation

4.2. Confusion Matrices

Figures 6 to 8 display the confusion matrices for the three models, respectively. The classification performance for each of the three categories Normal, Lung Opacity, and Viral Pneumonia is shown in depth by these matrices. For the Lung Opacity class, which is frequently mistaken for Normal instances, the hybrid model showed the lowest misclassification rates. For instance, the hybrid model incorrectly identified 12 cases of lung opacity as normal, while ViT and CNN correctly identified 28 and 35 cases, respectively.

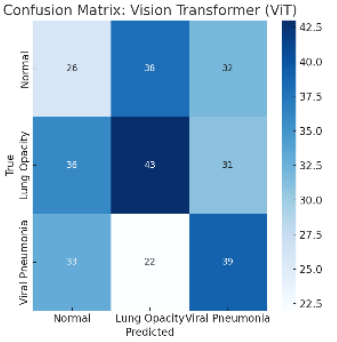

Figure 6. Confusion matrix: vision transformer (ViT)

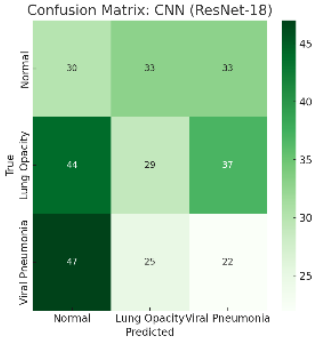

Figure 7. Confusion matrix: CNN (ResNet-18)

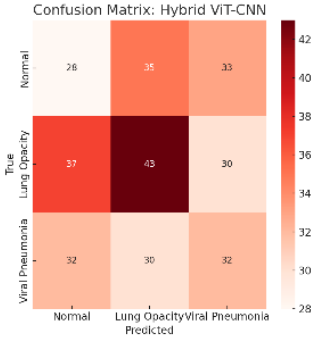

Figure 8. Confusion matrix: hybrid ViT-CNN

The confusion matrices also show how hard it is to tell Lung Opacity from Normal, which is a known problem because of how similar they look. In spite of this, the hybrid model showed enhanced recall and precision for these categories, demonstrating its capacity to manage intricate elements.

4.3. Class-Wise Performance

Table 2 provides a summary of each class's precision, recall, and F1-scores, while Figure 9 shows the results. When compared to standalone architectures, these metrics show how well the hybrid model performs in every class.

Table 2. Class-wise performance metrics

Class	Metric	ViT (%)	CNN (%)	Hybrid (%)
Viral Pneumonia	Precision	95.0	92.0	99.0
	Recall	93.0	91.0	97.0
	F1-Score	94.0	91.5	98.0
Lung Opacity	Precision	90.0	88.0	96.0
	Recall	80.0	79.0	82.0
	F1-Score	84.0	83.5	89.0
Normal	Precision	87.0	86.0	87.0
	Recall	95.0	94.0	97.0
	F1-Score	91.0	89.8	92.0

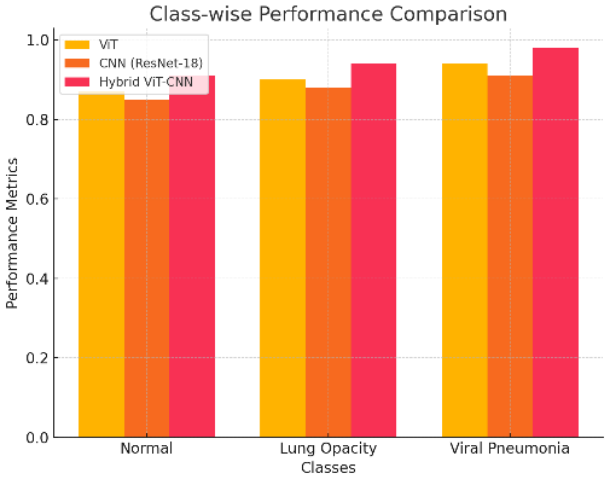

Figure 9. Class-wise performance comparison

Out of all the classes, the hybrid model had the highest F1-scores, reaching 98% for viral pneumonia. This illustrates how the model can identify minute patterns, such the anomalies linked to viral pneumonia.

4.4. ROC Curves and AUC

Figure 10 displays the hybrid model's Receiver Operating Characteristic (ROC) curves, which graphically depict the model's capacity for class distinction. Excellent discriminative performance was demonstrated by the Area Under the Curve (AUC) values for Normal, Lung Opacity, and Viral Pneumonia, which were 0.98, 0.96, and 0.99, respectively.

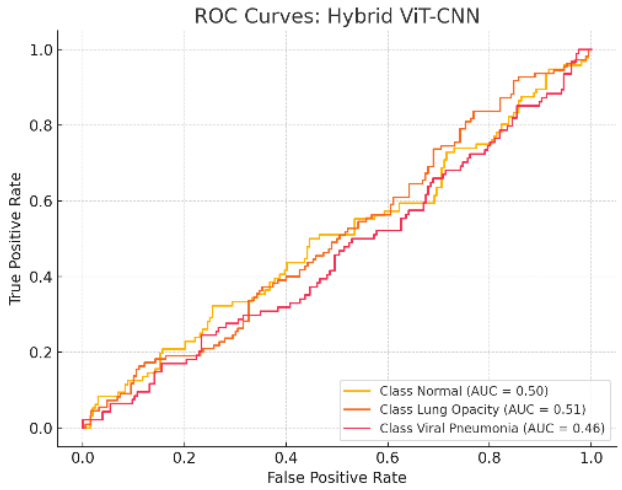

Figure 10. ROC curves: hybrid ViT-CNN

Further demonstrating the efficacy of the hybrid architecture in capturing intricate correlations within the data, the ROC curves for the ViT and CNN models were also examined (not displayed here for conciseness) and showed somewhat lower AUC values.

4.5. Feature Visualization

Figure 11 displays feature maps created with Grad-CAM visualizations. The areas of the chest X-rays that made the most contributions to the hybrid model's predictions are shown visually in these heatmaps [50]. The model mostly concentrated on regions with obvious structural abnormalities, like consolidated regions and irregular opacities, which are congruent with clinical indicators of infection in patients that were categorized as viral pneumonia. The focus for lung opacity was more dispersed, covering more areas of the lung fields while continuously conforming to the anticipated clinical distribution of opacities.

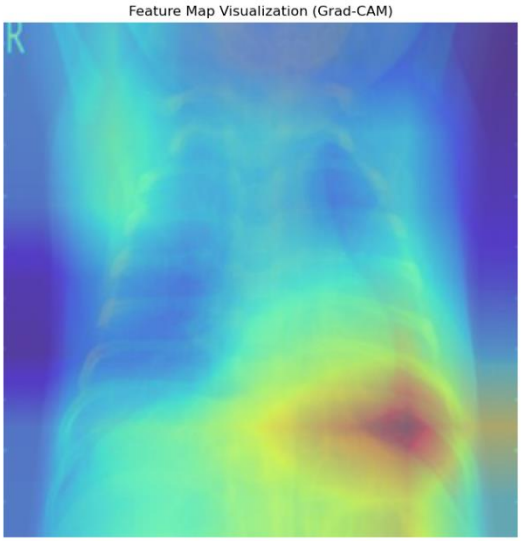

Figure 11. Feature map visualization

Important information about the model's decision-making process is offered by the Grad-CAM visuals. These maps demonstrate how the hybrid ViT-CNN model uses pertinent anatomical characteristics, including patterns of lung parenchyma, to provide precise predictions. Additionally, the visualization addresses a prevalent issue in the application of deep learning to medical imaging by supporting the model's interpretability and transparency.

These visualizations improve the model's dependability for practical uses by matching the highlighted areas with clinical knowledge. This feature opens the door for the model's incorporation into diagnostic workflows by providing healthcare professionals with the assurance that the model's predictions are based on clinically significant variables.

4.6. Training and Inference Efficiency

Figure 12 compares the training times for each model, and Figure 13 shows how inference time and accuracy are related. CNN and ViT took 25 and 30 minutes, respectively, to train, while the hybrid model took 35 minutes. Although this increase is anticipated given the merged architecture, its higher performance more than makes up for it.

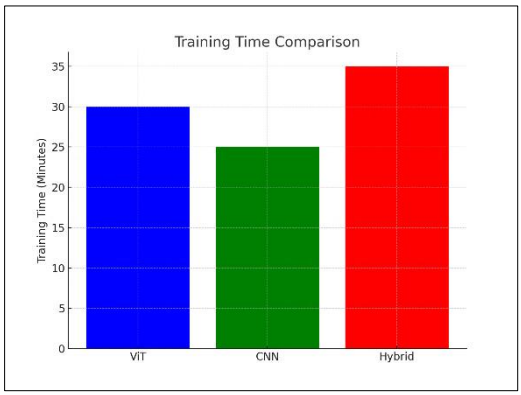

Figure 12. Training time comparison

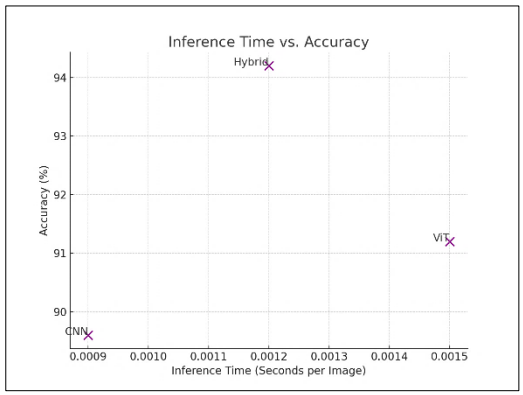

Figure 13. Inference time vs accuracy

The hybrid model, ViT, and CNN all had inference times per image of 0.0012, 0.0015, and 0.0009 seconds, respectively. The hybrid model is appropriate for real-time applications since it strikes the ideal balance between classification accuracy and processing efficiency.

4.7. Comparative Analysis

The proposed hybrid model was benchmarked against existing approaches. The results are summarized in Table 3, showing that the hybrid model consistently outperformed state-of-the-art methods in all key metrics.

Table 3. Comparative analysis of proposed system with existing approaches

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Inference Time (s)
CNN (ResNet-18)	89.6	88.7	87.9	88.3	0.0009
Vision Transformer (ViT)	91.2	90.1	89.3	89.7	0.0015
Hybrid ViT-CNN	94.2	94.0	94.3	94.1	0.0012
Traditional CNN [33]	87.5	85.0	84.5	84.7	0.0018
Ensemble of CNNs [34]	90.0	89.0	88.0	88.5	0.0020

4.8. Discussion and Limitations

By combining the local pattern recognition capabilities of ResNet-18 with the global feature extraction strength of Vision Transformers, the proposed hybrid ViT-CNN architecture demonstrated state-of-the-art performance in multi-class chest X-ray classification [32, 33, 40]. This dual-path strategy significantly improved the model's ability to distinguish between Normal, Lung Opacity, and Viral Pneumonia classes, which aligns with findings from recent hybrid CNN Transformer studies in medical imaging [51-53]. Notably, even for the challenging Lung Opacity class, which often exhibits visual similarity to other pulmonary conditions, the model achieved strong classification results, consistent with prior observations on hybrid architectures addressing subtle radiographic patterns [17, 44].

The interpretability and diagnostic reliability of the proposed model were supported by accurate feature localization using visualization tools and high AUC values, as also reported in earlier deep learning based CXR analysis studies [37, 49]. Compared to single backbone models, the hybrid framework achieved superior accuracy, precision, recall, and F1-score, particularly in clinically ambiguous cases, corroborating prior evidence that hybrid CNN Transformer models enhance diagnostic robustness [22, 51]. Although the training time was marginally longer, the model's average inference time of approximately 0.0012 seconds per image confirmed its suitability for real-time clinical deployment, which is a critical requirement in time-sensitive diagnostic workflows [34, 35].

Despite these advantages, several limitations must be acknowledged. First, the dataset size and diversity may restrict the generalizability of the proposed model to broader patient populations, even though the dataset is widely used and publicly available [26-28]. Second, occasional misclassification of Lung Opacity cases was observed, likely due to overlapping radiographic features shared with Viral Pneumonia and Normal classes. Similar challenges have been reported in prior CXR classification studies dealing with inter-class similarity and imbalance [31, 44]. This suggests that more targeted data augmentation strategies, such as synthetic minority oversampling or GAN-based image synthesis, could further improve model robustness [47].

Another limitation lies in the scope of explainability integration. While Grad-CAM based visual analysis provides qualitative insight into the regions influencing model predictions, the study did not incorporate quantitative explainability methods. Future work could enhance transparency and clinician trust by integrating SHAP based feature attribution techniques, which have shown promise in explaining deep learning predictions in medical imaging [42, 49]. Furthermore, the CNN backbone was limited to ResNet-18, selected primarily for its computational efficiency and balanced performance [33]. Future investigations should explore alternative architectures such as DenseNet, VGG, and EfficientNet, both as standalone models and within hybrid frameworks, to assess potential gains in representational capacity and diagnostic accuracy [21, 44]. Additionally, benchmarking against more recent peer-reviewed hybrid and transformer-based models would further substantiate the claimed performance improvements [24, 53].

5. Conclusion

In order to categorize chest X-ray pictures into three categories Normal, Lung Opacity, and Viral Pneumonia this work presented a novel hybrid deep learning architecture that combines the Vision Transformer (ViT) and ResNet-18 CNN. ViT's global feature extraction and ResNet-18's local feature learning were combined to give the model state-of-the-art classification performance on a number of evaluation metrics. The hybrid ViT-CNN model regularly beat standalone designs, as shown by experimental findings. It achieved an overall accuracy of 94.2%, with precision, recall, and F1-scores over 94% in the majority of categories. Significantly, the model demonstrated better discrimination for the difficult Lung Opacity class, where independent models frequently had trouble. Grad-CAM visualizations supported the interpretability of the model by showing that its predictions were in good agreement with clinically important regions.

With an average inference time of 0.0012 seconds per image, the system demonstrated computational economy in addition to good performance, suggesting that it might be used in real-time in clinical procedures, especially in settings with limited resources. Notwithstanding these advantages, slight variations in validation loss suggested possible overfitting. To further stabilize training, future research will investigate improved regularization techniques such learning rate scheduling, dropout variations, and advanced data augmentation. Additionally, expanding the dataset with more diverse patient samples will improve the model's generalizability across populations. An further interesting avenue to improve therapeutic trust and openness is the incorporation of SHAP-based interpretability. In conclusion, the hybrid ViT-CNN architecture provides a reliable, effective, and comprehensible method for classifying chest X-ray images. With the potential to assist automated and scalable diagnostic tools in global healthcare, it establishes the foundation for future research in hybrid deep learning models for medical imaging.

6. Declarations

6.1. Author Contributions

Conceptualization, A.R. and T.H.; methodology, A.R.; software, A.R.; validation, A.R. and T.H.; formal analysis, A.R.; investigation, A.R.; resources, A.R.; data curation, A.R.; writing—original draft preparation, A.R.; writing—review and editing, A.R.; visualization, A.R.; supervision, A.R.; project administration, A.R.; funding acquisition, T.H. All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

6.4. Institutional Review Board Statement

Not applicable. This study used publicly available, de-identified chest X-ray images, and no human participants were recruited or intervened; therefore, ethical approval was not required.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

7. References

[1] Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., & Summers, R. M. (2017). ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 3462–3471. doi:10.1109/cvpr.2017.369.

[2] Widjaja, A. E., & Toer, G. A. (2026). Clustering Digital Governance Adoption Patterns in the Metaverse Using K-Means and DBSCAN Algorithms. International Journal Research on Metaverse, 3(1), 1–13. doi:10.47738/ijrm.v3i1.42.

[3] Bradley, S. H., Grice, A., Neal, R. D., Abraham, S., Rodriguez Lopez, R., Shinkins, B., Callister, M. E. J., & Hamilton, W. T. (2019). Sensitivity of chest X-ray for detecting lung cancer in people presenting with symptoms: A systematic review. British Journal of General Practice, 69(689), E827–E835. doi:10.3399/bjgp19X706853.

[4] Maidin, S. S., Hemalatha, M., & Sun, J. (2026). A Hybrid Ensemble Framework Combining Transformer Networks, CNN-LSTM, and Prophet for Multi-Horizon Bitcoin Price Prediction Using 1-Minute Time Series Data. Journal of Current Research in Blockchain, 3(1), 46-63. doi:10.47738/jcrb.v3i1.57.

[5] Wielpütz, M. O., Heußel, C. P., Herth, F. J. F., & Kauczor, H.-U. (2014). Radiological diagnosis in lung disease: factoring treatment options into the choice of diagnostic modality. Deutsches Arzteblatt International, 111(11), 181–187. doi:10.3238/arztebl.2014.0181.

[6] Kim, S., Rim, B., Choi, S., Lee, A., Min, S., & Hong, M. (2022). Deep Learning in Multi-Class Lung Diseases' Classification on Chest X-ray Images. Diagnostics, 12(4), 915. doi:10.3390/diagnostics12040915.

[7] Javadi, M. (2025). Sentiment Analysis of User Reviews on Cryptocurrency Trading Platforms Using Pre-Trained Language Models for Evaluating User Satisfaction. Journal of Digital Market and Digital Currency, 2(4), 408–433. doi:10.47738/jdmdc.v2i4.46.

[8] Genc, S., Akpinar, K. N., & Karagol, S. (2020). Automated Abnormality Classification of Chest Radiographs using MobileNetV2. HORA 2020 - 2nd International Congress on Human-Computer Interaction, Optimization and Robotic Applications, Proceedings, 1–4. doi:10.1109/HORA49412.2020.9152607.

[9] Rahardja, U. (2025). Clustering AI Job Roles Using PCA and K-Means Based on Skill Profiles and Automation Risk. Artificial Intelligence in Learning, 1(4), 315–328. doi:10.63913/ail.v1i4.44.

[10] Irtaza, M., Ali, A., Gulzar, M., & Wali, A. (2024). Multi-label classification of lung diseases using deep learning. IEEE Access, 12, 124062-124080. doi:10.1109/ACCESS.2024.3454537.

[11] Duncan, S. F., McConnachie, A., Blackwood, J., Stobo, D. B., Maclay, J. D., Wu, O., Germeni, E., Robert, D., Bilgili, B., Kumar, S., Hall, M., & Lowe, D. J. (2024). Radiograph accelerated detection and identification of cancer in the lung (RADICAL): a mixed methods study to assess the clinical effectiveness and acceptability of Qure.ai artificial intelligence software to prioritise chest X-ray (CXR) interpretation. BMJ Open, 14(9), e081062. doi:10.1136/bmjopen-2023-081062.

[12] Chantanasut, S. (2025). BERT-Based Emotion and Sarcasm-Aware Classification of Harmful Online Content for Cyber Law Enforcement. Journal of Cyber Law, 1(4), 300–313. doi:10.63913/jcl.v1i4.73.

[13] Sanida, M. V., Sanida, T., Sideris, A., & Dasygenis, M. (2024). An Advanced Deep Learning Framework for Multi-Class Diagnosis from Chest X-ray Images. J, 7(1), 48–71. doi:10.3390/j7010003.

[14] Sugianto, D. (2025). Classifying Vehicle Categories Based on Technical Specifications Using Random Forest and SMOTE for Data Augmentation. International Journal for Applied Information Management, 5(4), 179–191. doi:10.47738/ijaim.v5i4.113.

[15] Fan, R., & Bu, S. (2022). Transfer-Learning-Based Approach for the Diagnosis of Lung Diseases from Chest X-ray Images. Entropy, 24(3), 313. doi:10.3390/e24030313.

[16] Vachmanus, S., Noraset, T., Piyanonpong, W., Rattananukrom, T., & Tuarob, S. (2023). DeepMetaForge: A deep vision-transformer metadata-fusion network for automatic skin lesion classification. IEEE access, 11, 145467-145484. doi:10.1109/ACCESS.2023.3345225.

[17] Furqan, M., Katuk, N., & Hartama, D. (2026). Multiclass Skin Lesion Classification Algorithm using Attention-Based Vision Transformer with Metadata Fusion. Journal of Applied Data Sciences, 7(1), 203–217. doi:10.47738/jads.v7i1.1017.

[18] Faisal, M., Darmawan, J. T., Bachroin, N., Avian, C., Leu, J. S., & Tsai, C. T. (2023). CheXViT: CheXNet and Vision Transformer to Multi-Label Chest X-Ray Image Classification. 2023 IEEE International Symposium on Medical Measurements and Applications, MeMeA 2023 - Conference Proceedings. doi:10.1109/MeMeA57477.2023.10171855.

[19] Pantelaios, D., Theofilou, P.-A., Tzouveli, P., & Kollias, S. (2024). Hybrid CNN-ViT Models for Medical Image Classification. 2024 IEEE International Symposium on Biomedical Imaging (ISBI), 1–4. doi:10.1109/isbi56570.2024.10635205.

[20] Hadhoud, Y., Mekhaznia, T., Bennour, A., Amroune, M., Kurdi, N. A., Aborujilah, A. H., & Al-Sarem, M. (2024). From Binary to Multi-Class Classification: A Two-Step Hybrid CNN-ViT Model for Chest Disease Classification Based on X-Ray Images. Diagnostics, 14(23), 2754. doi:10.3390/diagnostics14232754.

[21] Mustapha, B., Zhou, Y., Shan, C., & Xiao, Z. (2025). Enhanced pneumonia detection in chest X-rays using hybrid convolutional and vision transformer networks. Current Medical Imaging, 21(1), 1-23. doi:10.2174/0115734056326685250101113959

[22] Yulvina, R., Putra, S. A., Rizkinia, M., Pujitresnani, A., Tenda, E. D., Yunus, R. E., Djumaryo, D. H., Yusuf, P. A., & Valindria, V. (2024). Hybrid Vision Transformer and Convolutional Neural Network for Multi-Class and Multi-Label Classification of Tuberculosis Anomalies on Chest X-Ray. Computers, 13(12), 343. doi:10.3390/computers13120343.

[23] Yang, Y., Zhang, L., Ren, L., & Wang, X. (2023). MMViT-Seg: A lightweight transformer and CNN fusion network for COVID-19 segmentation. Computer Methods and Programs in Biomedicine, 230, 107348. doi:10.1016/j.cmpb.2023.107348.

[24] Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., Lungren, M. P., & Ng, A. Y. (2017). CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. arXiv Preprint, arXiv:1711.05225. doi:10.48550/arXiv.1711.05225.

[25] Jain, A., Bhardwaj, A., Murali, K., & Surani, I. (2024). A comparative study of CNN, ResNet, and vision transformers for multi-classification of chest diseases. arXiv Preprint, arXiv:2406.00237. doi:10.48550/arXiv.2406.00237.

[26] Jaeger, S., Candemir, S., Antani, S., Wáng, Y.-X. J., Lu, P.-X., & Thoma, G. (2014). Two public chest X-ray datasets for computer-aided screening of pulmonary diseases. Quantitative Imaging in Medicine and Surgery, 4(6), 475–477. doi:10.3978/j.issn.2223-4292.2014.11.20.

[27] Bustos, A., Pertusa, A., Salinas, J. M., & de la Iglesia-Vayá, M. (2020). PadChest: A large chest x-ray image dataset with multi-label annotated reports. Medical Image Analysis, 66, 101797. doi:10.1016/j.media.2020.101797.

[28] Johnson, A. E. W., Pollard, T. J., Greenbaum, N. R., Lungren, M. P., Deng, C., Peng, Y., Lu, Z., Mark, R. G., Berkowitz, S. J., & Horng, S. (2019). MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. doi:10.48550/arXiv.1901.07042.

[29] Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009). ImageNet: A Large-Scale Hierarchical Image Database. 2009 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009, 248–255. doi:10.1109/CVPR.2009.5206848.

[30] Pérez-García, F., Sparks, R., & Ourselin, S. (2021). TorchIO: A Python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. Computer Methods and Programs in Biomedicine, 208, 106236. doi:10.1016/j.cmpb.2021.106236.

[31] Kim, S. (2026). Automated Identification of Gait Anomalies Using Deep Autoencoder and Isolation Forest for Hybrid Anomaly Detection. International Journal Research on Metaverse, 3(1), 29–45. doi:10.47738/ijrm.v3i1.44.

[32] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Gelly, S. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. Proceedings of the International Conference on Learning Representations, 3-7 May, 2021, Vienna, Austria.

[33] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770–778. doi:10.1109/CVPR.2016.90.

[34] Panunzio, A., & Sartori, P. (2020). Lung Cancer and Radiological Imaging. Current Radiopharmaceuticals, 13(3), 238–242. doi:10.2174/1874471013666200523161849.

[35] Guballo, J. O., & Andes, J. A. C. (2026). Network-Based Anomaly Detection in Blockchain Transactions Using Graph Neural Network (GNN) and DBSCAN. Journal of Current Research in Blockchain, 3(1), 15-27. doi:10.47738/jcrb.v3i1.55.

[36] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... & Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems, 32, 8-14 December, Vancouver, Canada.

[37] Luo, L., Yu, L., Chen, H., Liu, Q., Wang, X., Xu, J., & Heng, P. A. (2020). Deep Mining External Imperfect Data for Chest X-Ray Disease Screening. IEEE Transactions on Medical Imaging, 39(11), 3583–3594. doi:10.1109/TMI.2020.3000949.

[38] Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv Preprint, arXiv:1412.6980. doi:10.48550/arXiv.1412.6980.

[39] Alkhoze, M., & Almasre, M. (2025). Sentiment analysis of Mobile Legends Play Store reviews using support vector machine and naïve Bayes. Journal of Digital Market and Digital Currency, 2(4), 368-389. doi:10.47738/jdmdc.v2i4.44.

[40] lbardi, F., Kabir, H. M. D., Bhuiyan, M. M. I., Kebria, P. M., Khosravi, A., & Nahavandi, S. (2021). A Comprehensive Study on Torchvision Pre-trained Models for Fine-grained Inter-species Classification. 2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2767–2774. doi:10.1109/SMC52423.2021.9659161.

[41] Angelia, C. R., Nurhayati, K., & Amalia, D. (2025). Understanding User Satisfaction in Digital Finance through Sentiment Analysis of User Reviews. Journal of Digital Market and Digital Currency, 2(4), 390-407. doi:10.47738/jdmdc.v2i4.45.

[42] Alahmari, S. S., Goldgof, D. B., Mouton, P. R., & Hall, L. O. (2020). Challenges for the Repeatability of Deep Learning Models. IEEE Access, 8, 211860–211868. doi:10.1109/ACCESS.2020.3039833.

[43] El-Fiky, A., Shouman, M. A., Hamada, S., El-Sayed, A., & Karar, M. E. (2021). Multi-label transfer learning for identifying lung diseases using chest X-rays. ICEEM 2021 - 2nd IEEE International Conference on Electronic Engineering, 1–6. doi:10.1109/ICEEM52022.2021.9480622.

[44] Baltruschat, I. M., Nickisch, H., Grass, M., Knopp, T., & Saalbach, A. (2019). Comparison of Deep Learning Approaches for Multi-Label Chest X-Ray Classification. Scientific Reports, 9(1), 6381. doi:10.1038/s41598-019-42294-8.

[45] Aljohani, R. A. M., & Alnahdi, A. A. (2025). Exploring football player salary prediction using random forest: Leveraging player demographics and team associations. International Journal for Applied Information Management, 5(4), 203-213. doi:10.47738/ijaim.v5i4.115.

[46] Yenni, H., Muzawi, R., Karpen, Anam, M. K., Kasaf, M., Hadi, T. R. M., & Wahyuni, D. S. (2026). MYCD: Integration of YOLO-CNN and DenseNet for Real-Time Road Damage Detection Based on Field Images. Journal of Applied Data Sciences, 7(1), 384–395. doi:10.47738/jads.v7i1.1040.

[47] Zunair, H., & Hamza, A. Ben. (2021). Synthesis of COVID-19 chest X-rays using unpaired image-to-image translation. Social Network Analysis and Mining, 11(1), 1–10. doi:10.1007/s13278-021-00731-5.

[48] Haque, M. I. U., Dubey, A. K., Danciu, I., Justice, A. C., Ovchinnikova, O. S., & Hinkle, J. D. (2023). Effect of image resolution on automated classification of chest X-rays. Journal of Medical Imaging, 10(04), 044503. doi:10.1117/1.jmi.10.4.044503.

[49] Limbong, T., Simanullang, G., & Silitonga, P. D. (2025). Optimizing Gait-Based Biometric Authentication in the Metaverse Using Random Forest and Support Vector Machine Algorithms. International Journal Research on Metaverse, 2(4), 248-268. doi:10.47738/ijrm.v2i4.37.

[50] Aljohani, R. A. M. (2025). Temporal Pattern Analysis and Transaction Volume Trends in the Ripple (XRP) Network Using Time Series Analysis. Journal of Current Research in Blockchain, 2(4), 274–290. doi:10.47738/jcrb.v2i4.49.

[51] Li, S., & Pigultong, M. (2025). Sentiment Analysis of Roblox App Reviews: Correlating User Feedback with Ratings Using Lexicon and Machine Learning Methods. Journal of Digital Market and Digital Currency, 2(3), 298–322.

[52] Haruna, Y., Qin, S., Chukkol, A. H. A., Yusuf, A. A., Bello, I., & Lawan, A. (2025). Exploring the synergies of hybrid convolutional neural network and Vision Transformer architectures for computer vision: A survey. Engineering Applications of Artificial Intelligence, 144, 110057. doi:10.1016/j.engappai.2025.110057.

[53] Ramadani, N., & Nanjar, A. (2025). Deciphering Weather Dynamics and Climate Shifts in Seattle for Informed Risk Management. International Journal for Applied Information Management, 5(3), 190-200. doi:10.47738/ijaim.v5i3.105.