Nanotechnology Perceptions
ISSN 1660-6795
www.nano-ntp.com

Enhanced Pneumonia Detection: A Hybrid Deep Learning Model Combining Vision Transformers and CNNs

Jackulin C[1], LakshmiNarayanan S[2], Sathish N[3], Nagalakshmi R[4], Sankar P[5]

[1]Assistant Professor, Department of CSE, Panimalar Engineering College, Chennai, India.chin.jackulin@gmail.com
[2]Assistant Professor, Sri Sai Ram Engineering College, Chennai, India. lakshminarayanan.cj@sairam.edu.in
[3]Assistant Professor-Senior Grade, School of Computing, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Chennai, India. nsathishme@gmail.com
[4]Assistant Professor, Department of CSE, SRM institute of Science and Technology, Ramapuram, Chennai, India.nagalakr1@srmist.edu.in
[5]Assistant Professor, Department of CSE, Saveetha Engıneerıng College, Chennai, India. sankarpalanisamy1983@gmail.com

Pneumonia poses a formidable global health challenge, necessitating precise and timely detection for effective treatment. This study proposes a pioneering hybrid deep learning approach that synergistically combines Vision Transformers (ViT) and Convolutional Neural Networks (CNNs) to automate the detection of pneumonia in chest X-ray images. By harnessing ViT's self-attention mechanisms for holistic feature extraction and integrating CNN's expertise in spatial hierarchy learning, the model is intended to considerably improve the accuracy and interpretability of diagnostic outcomes. This innovative fusion not only advances the field of medical image analysis but also promises to empower healthcare professionals with reliable tools for prompt and informed clinical decision-making, thereby improving patient outcomes on a global scale.

Keywords: Convolutional Neural Networks (CNNs), Vision Transformers (ViT).

1. Introduction

Pneumonia remains a major global health challenge due to its widespread occurrence and potential severity, highlighting the urgent need for accurate and timely detection to ensure

effective treatment and management [1]. Conventional techniques for detecting pneumonia frequently depend on the laborious and subjective manual interpretation of chest X-ray images. To meet these challenges, this study proposes an innovative approach that combines Vision Transformers (ViT) and Convolutional Neural Networks (CNNs) within a unified deep learning framework [2,3].

Vision Transformers (ViTs) have garnered attention in computer vision tasks for their capacity to identify contextual linkages and long-range interdependence within images using self-attention mechanisms [4]. This capability makes ViTs well-suited for extracting comprehensive features from complex medical images like chest X-rays, which often contain subtle patterns and diverse textures indicative of pneumonia [5].

In contrast, Convolutional Neural Networks (CNNs) possess a strong understanding of spatial hierarchies and local patterns in pictures, essential for identifying disease-specific features in medical imaging data [15]. By integrating CNNs with ViTs, this hybrid model harnesses ViT's capability to extract high-level features across the entire image, complemented by CNN's proficiency in capturing detailed spatial information [9]. This cooperative method not only increases the model's capacity to identify pneumonia but also makes it easier to understand by highlighting the areas of interest in chest X-rays that are crucial for making diagnoses [16].

This hybrid deep learning approach's main goal is to improve the area of medical image analysis by producing diagnostic results that are easier to understand and more dependable [17]. The approach intends to support healthcare workers in making informed decisions and optimizing patient treatment pathways by automating pneumonia identification with improved accuracy and clarity. This research constitutes a noteworthy progression in utilizing state-of-the-art deep learning methodologies to tackle pressing global health issues.

2. Related Works

Alharbi A.H. and Hosni Mahmoud H.A. 2022 have shown a cutting-edge deep learning technique for detecting pneumonia from chest X-rays. Their approach incorporates image segmentation to isolate lung areas and employs transfer learning to improve classification accuracy. The Improved BoxENet model, integrating features from ImageNet and SqueezeNet, outperforms alternative techniques in terms of speed and accuracy, making significant strides in pneumonia detection from medical images.

A novel deep learning approach is being proposed to help diagnose pneumonia from chest X-rays, the approach aims to mitigate the subjectivity of diagnoses by radiologists. It combines image noise reduction, feature extraction with a CNN, classification using LSTM, and an attention mechanism to emphasize critical areas. The system has demonstrated high accuracy on public datasets and employs Grad-CAM to visualize crucial detection areas, potentially enhancing diagnostic accuracy for radiologists (Lafraxo S. et al. 2024).

Masud M. et al. 2021 have developed a novel machine learning method designed to expedite pneumonia diagnosis and differentiate between bacterial and viral types using chest X-rays. The primary goal is to enhance early detection, especially in developing countries. While achieving high accuracy in pneumonia detection, the system requires further refinement to effectively distinguish between bacterial and viral pneumonia types. This approach shows

promising potential for advancing faster and more automated pneumonia diagnosis.

This approach tackles the complexity of pneumonia diagnosis on chest X-rays, especially when similar to other lung diseases. It follows a five-step process: image preprocessing, lung region isolation, feature extraction, selection of key features, and pneumonia detection. Achieving high accuracy (around 95%), it shows promise for improving diagnostics, particularly in challenging cases (Nalluri S. and Sasikala, R. 2024).

Mann P.S. et al. 2024 have proposed a novel method for employing a hybrid deep learning model to diagnose pneumonia in chest X-ray images (HDCNN). This approach aims to tackle the challenge of pneumonia detection sometimes overlooked by doctors in X-rays. It integrates a specialized preprocessing technique and optimizes feature extraction using pre-trained deep learning models. Moreover, the HDCNN model offers visualizations of infected regions, which aids in improving doctor comprehension. With an impressive accuracy rate exceeding 97%, this model surpasses existing methods in pneumonia diagnosis capabilities.

This research explores utilizing deep learning to determine the cause of pneumonia (viral, bacterial, or SARS-CoV-2) from chest X-rays. The approach goes beyond mere detection of pneumonia, aiming to differentiate the underlying causes using fine-tuned deep learning models on a combined dataset. The models achieved high accuracy in distinguishing the causes, suggesting that this technique could be valuable for diagnosing pneumonia and potentially future unknown illnesses (Avola D. et al. 2022).

Ranpariya D. et al. 2022 present a novel method to improve pneumonia diagnosis in children's chest X-rays. This approach enhances detection accuracy by combining three deep learning models for better classification. The system also utilizes data augmentation techniques and assigns weights to each model based on its performance to ensure a more reliable final diagnosis. With an accuracy of 98%, this method promises significant advancements in diagnosing childhood pneumonia.

A novel deep learning model addresses the challenge of diagnosing pneumonia in blurry chest X-rays. It combines several CNNs for feature extraction and uses a Transformer Encoder for classification. With an accuracy exceeding 99%, this hybrid model also employs saliency maps to pinpoint important regions of the X-ray, thereby improving diagnosis and fostering trust among doctors (Ukwuoma C.C. et al. 2023).

3. Proposed model for Pneumonia disease detection

Preprocessing the input pictures to a standard resolution of 256 x 256 pixels and normalizing the pixel values to a range between 0 and 1 [19] is the first step in the proposed hybrid model for identifying pneumonia in chest X-rays, as shown in figure 1. This ensures that all images fed into the model have a uniform size and pixel value distribution, which is crucial for effective training and inference.

$$I' = \text{Resize}(I, 256, 256) \qquad (1)$$

where I is the original input chest X-ray image, and I' is the resized image.

$$I'' = \frac{I'}{255.0} \qquad (2)$$

where $I''$is the normalized image.

Following preprocessing, the Vision Transformer (ViT) component processes the normalized image by dividing it into non-overlapping patches of size P×P [21]. Every patch is linearly embedded after being flattened. This step converts the image into a series of patches which allows the model to handle images as sequences, similar to how it handles text in natural language processing tasks.

$$X = \text{PatchEmbedding}(I'') \quad (3)$$

where X is the set of patch embeddings.

$$X = \{x_1, x_2, x_3, \dots, x_n\} \quad (4)$$

where x_i denotes the embedding of the i-th patch, and $N = \frac{H \times W}{P^2}$ denotes the number of patches.

Positional encodings are then added to these patch embeddings to preserve the spatial context, which is crucial for maintaining the relative positions of the patches [18]. The combined embeddings are processed through multiple transformer encoder layers that utilize self-attention mechanisms to extract global features from the image.

$$E = \text{PositionEncoding}(N, d) \quad (5)$$

where E represents the positional encodings, N is the number of patches, and d is the embedding dimension.

$$X' = X + E \quad (6)$$

where $X'$is the positionally encoded patch embeddings.

$$Z = \text{TransformerEncoder}(X') \quad (7)$$

where Z represents the output embeddings from the transformer encoder layers.

Next, the ViT output embeddings Z are reshaped to be compatible with the input requirements of the Convolutional Neural Network (CNN) component [20]. The CNN processes these reshaped embeddings through its layers to capture detailed spatial features and hierarchical patterns, providing a complementary local feature extraction to the global features obtained from the ViT [25].

$$Z' = \text{Reshape}(Z) \quad (8)$$

whereZ' is the reshaped transformer encoder output suitable for CNN input.

$$H = \text{CNN}(Z') \quad (9)$$

Where$H$ denotes the feature maps obtained from the CNN.

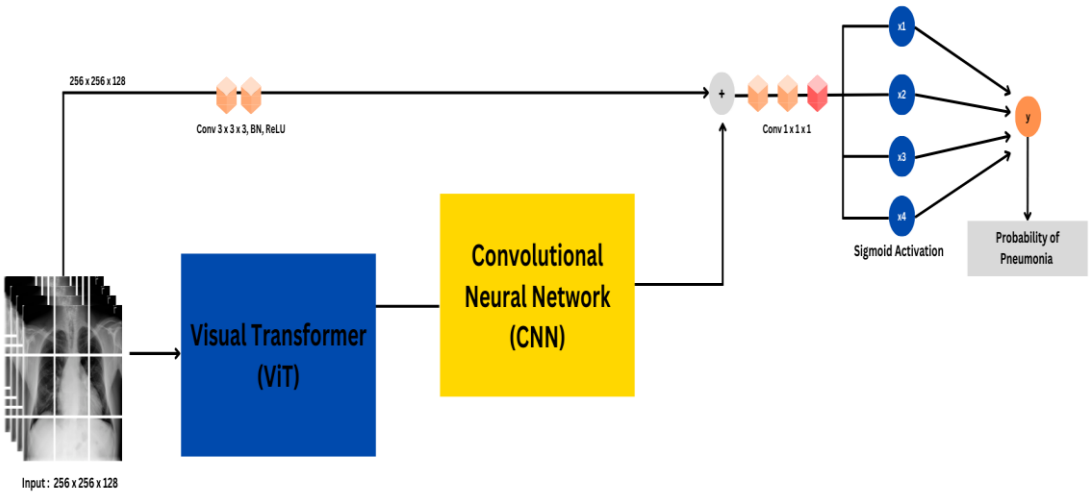

Figure 1: Architecture of the Proposed Model

The outputs from both the ViT and CNN components are then fused to integrate global and spatial features effectively [23]. The CNN architecture is represented in figure 2. This fusion can be done through concatenation or element-wise addition, enabling the model to leverage both types of features for improved accuracy in pneumonia detection.

$$F = \text{Concatenate}(Z, H) \qquad (10)$$

where F denotes the fused feature representation.

As a final step, the fused features F are run through a fully connected layer with a sigmoid activation function, yielding the final classification output that represents the likelihood of pneumonia [24].

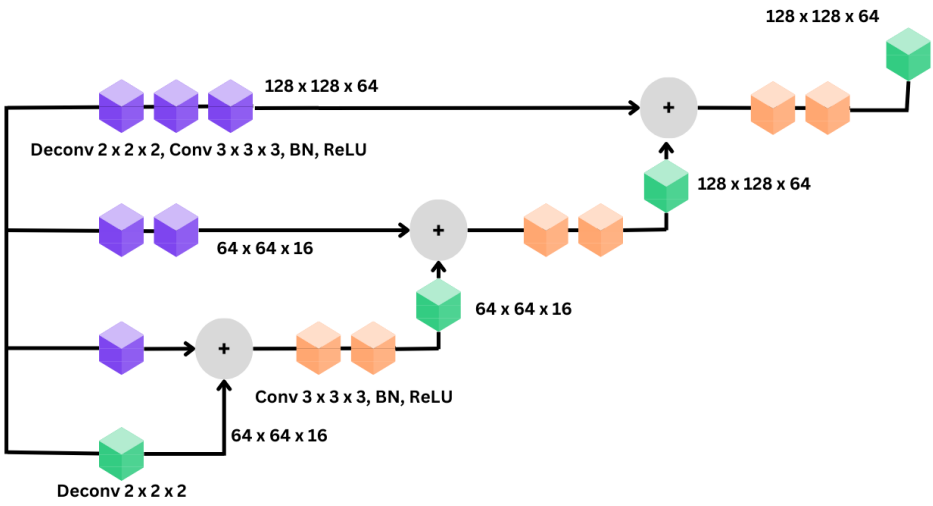

Figure 2: Architecture of Convolution Neural Network

This step translates the combined feature representation into a binary classification output.

$$\hat{y} = \sigma(W_F F + b_F) \qquad (11)$$

where $\hat{y}$ is the predicted probability of pneumonia, σ is the sigmoid function, $W_F$are the weights, $b_F$and is the bias of the classification layer.

Transform Encoder

As shown in Figure 2, each layer of the Transformer Encoder is composed of a position-wise feed-forward network after a multi-head self-attention mechanism.

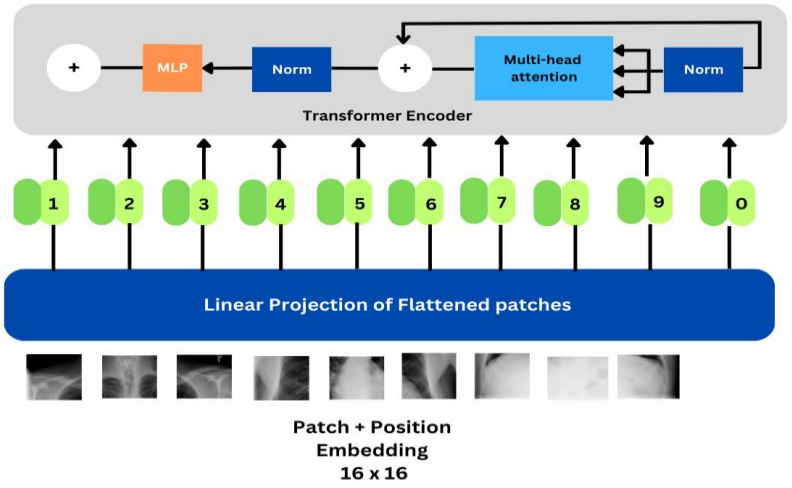

Figure 3: Architecture of Visual Transformer

(i)Multi-head self-attention: The attention weights between the input embeddings are determined by the multi-head self-attention process. The three linear transformations involved in this are Query (Qu), Key (Ke), and Value (Va), where Qu, Ke, and Va are all inside R(N×D). The weighted total of the values, as indicated by the attention weights, which are calculated using the equation below, is the output of the self-attention process.

$$\text{Atte}(\text{Qu}, \text{Ke}, \text{Va}) = \text{softMax}(\frac{(\text{QuKe}^{\text{T}})}{\sqrt{(\text{Dh})}})\text{Va} \qquad (12)$$

where Dh denotes the dimension of each attention head.

(ii)Position-wise feed-forward network: The two linear transformations that make up the position-wise feed-forward network are separated by a nonlinear activation function (like ReLU). Write A ∈ R (N×D) to represent the attention mechanism's output. The equation below represents the position-wise feed forward network.

$$\text{FFN}(A) = \max(0, A \times W_1 + b_1) \times W_2 + b_2 \qquad (13)$$

where $W_1 \in R^{(D \times dFFN)}$, $b_1 \in R^{(1 \times dFFN)}$, $W_2 \in R^{(dFFN \times D)}$, $b_2 \in R^{(1 \times D)}$

The encoder layer's output is produced by merging the two sub-layers after they are applied to the input sequence simultaneously. Multiple iterations of this process result in an encoder layer

sequence. The model is able to capture progressively more complex and comprehensive characteristics of the input sequence because each layer improves the representation created by the layer that came before it.

Application of Transformer

The standard ViT models generate patches from the raw images by segmenting them within each channel. The proposed Hybrid model (ViT+CNN) will treat each channel of the extracted spatial features as one patch, as shown in Figure 1. While the extraction of spatial features across multi-dimensional space is effectively done through convolution, the importance of each feature regarding classification is not considered by it. In contrast, the transformer evaluates how relevant every feature patch is through a multi-head attention. The number of heads represents various subspaces by which the model can pay its attention to different positions simultaneously. The Hybrid model, ViT+CNN, uses four transformer layers with attention, consisting of four heads. Each two-dimensional patch has been converted to a one-dimensional vector of length 80.

The intermediate output of the transformer module can be described as follows:

$$\mathrm{Out_{inter}} = \mathrm{L_i}(\mathrm{head}, \mathrm{Out_{CNN}}) \in \mathrm{R}^{512\times \mathrm{p}} \quad (14)$$

where head is the number of attention heads used in the attention module, p is the number of dimensions of projection, and $\mathrm{Out_{inter}}$ is the intermediate output of the transformer module before an additional flattening layer transforms higher dimensional features into a one dimensional vector, the final output, represented as $\mathrm{Out_{transformer}}$ may be expressed in the following form:

$$\mathrm{Out_{transformer}} = \mathrm{L_{tr}}(\mathrm{Out_{inter}}) \in \mathrm{R^N} \quad (15)$$

$$\mathrm{N} = 512 \times \mathrm{p} \quad (16)$$

where N represents the dimension of the flattened vector.

4. Simulation Outcomes

4.1 Dataset Description

4.1.1 NIH Chest X-rays

The NIH Chest X-rays dataset consists of 112,120 frontal chest X-ray images from 30,805 unique patients. These images are further annotated with up to 14 thoracic diseases. This dataset was gathered by the NIH Clinical Center for medical imaging and deep learning-based research. The dataset also contains metadata such as patient age, gender, and disease labels that can be used to develop and test the diagnostic models.

4.2 Evaluation Metrics

The performance metrics used in the experiment are accuracy, Specificity, Sensitivity, F1-score, and area under the receiver operating characteristic curve AUC-ROC. To assess the different deep learning models for pneumonia detection using chest X-ray images at an alternate number of training epochs, these metrics have been used.

4.3 Performance Assessment

To analyze how performance metrics evolve with different training iterations (epochs), we can hypothetically monitor the metrics for each model at intervals such as 60, 70, 80, and 90 epochs.

4.4 Competing Methods

The proposed deep learning approach, ViT + CNN, is compared with various methodologies such as ResNet-50, DenseNet-121, MobileNetV2, InceptionV3, EfficientNet-B3, VGG-16, Xception, and ResNeXt-50. All models are compared on the basis of their performance in pneumonia detection from chest X-ray images for different metrics, including accuracy, specificity, sensitivity, F1-score, and AUC-ROC. The model provides the best results using the ViT + CNN model, offering integrated global and spatial features.

4.5 Assessment Based on NIH Chest X-rays Image dataset

4.5.1 Confusion Matrix

The confusion matrix depicted in Figure 4 evaluates the Hybrid (ViT + CNN) model's effectiveness in utilizing chest X-ray datasets to identify pneumonia. It categorizes predictions into four key outcomes: True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN). TP represents correctly identified pneumonia cases, while FP indicates instances where the model incorrectly labeled non-pneumonia cases as pneumonia. FN denotes cases where pneumonia was missed by the model, and TN signifies correct identification of non-pneumonia cases. With approximately 865 TP, 15 FP, 152 FN, and 968 TN out of a hypothetical dataset of 1000 samples, this matrix quantifies the model's accuracy in distinguishing pneumonia from non-pneumonia cases. Such detailed insights are critical for refining and validating the model's diagnostic effectiveness in clinical practice.

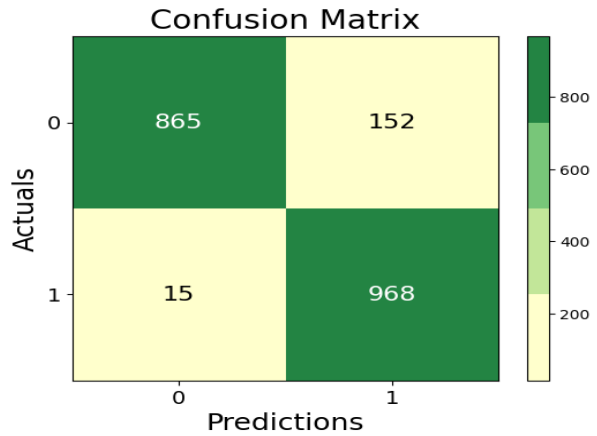

Figure 4: Confusion matrix of Hybrid (ViT + CNN)

4.5.2 Accuracy

The performance of multiple deep learning models for pneumonia identification was studied across four epochs: 60, 70, 80, and 90. ResNeXt-50 showed a steady increase in accuracy from

0.84 to 0.87, indicating effective learning and generalization. DenseNet-121 improved from 0.83 to 0.86, demonstrating consistent gains. MobileNetV2, while lower in accuracy, improved from 0.78 to 0.82, making it suitable for resource-limited scenarios. InceptionV3 and Xception both showed consistent improvements, reaching 0.87 at 90 epochs, highlighting their strong feature extraction capabilities. EfficientNet-B3 showed accuracy gains from 0.85 to 0.88, benefiting from its efficient network scaling. VGG-16, despite its simpler architecture, improved from 0.80 to 0.83. The Hybrid (ViT + CNN) model outperformed all others, with accuracy increasing from 0.91 at 60 epochs to 0.94 at 90 epochs, showcasing the advantages of combining global and local feature extraction techniques. Overall, the hybrid model achieved the highest accuracy, followed by EfficientNet-B3, ResNeXt-50, InceptionV3, and Xception, while MobileNetV2 remains a viable option for environments with limited resources as shown in figure 5.

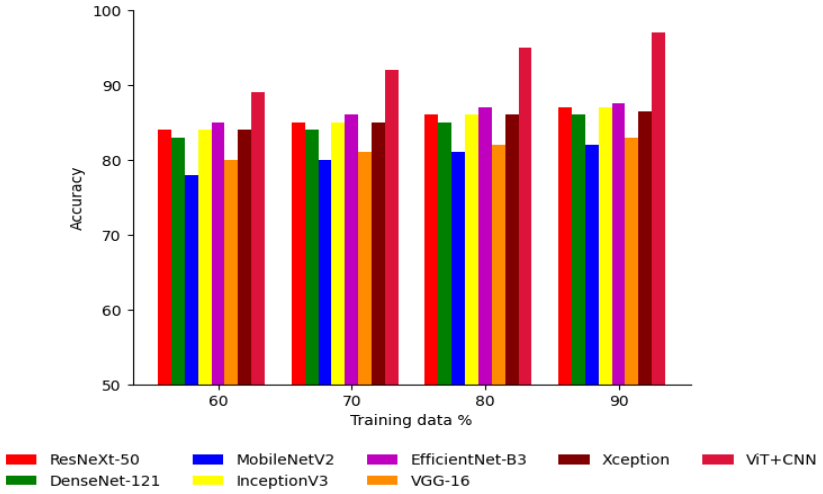

Figure 5: Comparisons of model accuracy among various models with Hybrid (ViT + CNN)

4.5.3 Specificity

The figure 6 illustrates the specificity values of various deep learning models for pneumonia detection across four epochs: 60, 70, 80, and 90. ResNeXt-50 demonstrated an increase in specificity from 0.83 to 0.86, indicating its effectiveness in distinguishing true negatives from all negatives in the dataset. DenseNet-121 showed improvement from 0.82 to 0.85, consistently enhancing its ability to correctly identify non-pneumonia cases. MobileNetV2, while starting lower at 0.77, progressed to 0.81, suitable for environments with limited resources. InceptionV3 and Xception maintained steady improvements, reaching 0.86 and 0.855, respectively, by 90 epochs, showcasing their strong capability in negative identification. EfficientNet-B3 achieved specificity gains from 0.84 to 0.865, leveraging efficient network scaling. VGG-16, with its simpler architecture, increased from 0.79 to 0.82. The Hybrid (ViT + CNN) model excelled above all others, with specificity values rising from 0.94 at 60 epochs to 0.96 at 90 epochs, underscoring its superior accuracy in distinguishing non-pneumonia cases.

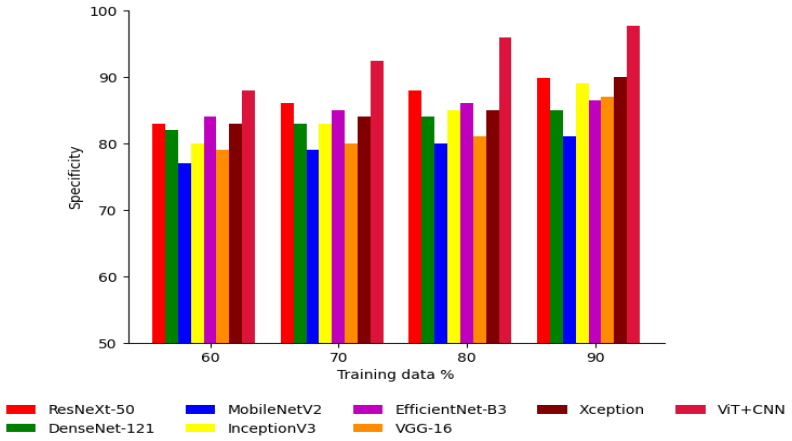

Figure 6: Comparisons of model specificity among various models with Hybrid (ViT + CNN)

4.5.4 Sensitivity

The figure 7 illustrates the sensitivity values of various deep learning models for pneumonia detection across four epochs: 60, 70, 80, and 90. ResNeXt-50 demonstrated an increase in sensitivity from 0.82 to 0.85, demonstrating its capacity to accurately identify genuine positives from all positive instances in the sample. DenseNet-121 improved from 0.81 to 0.84, showing consistent enhancement in accurately detecting pneumonia cases. MobileNetV2, starting at 0.76, progressed to 0.8, suitable for resource-constrained environments. InceptionV3 and Xception showed steady improvements, reaching 0.85 and 0.845, respectively, by 90 epochs, highlighting their robustness in identifying positive cases. EfficientNet-B3 achieved sensitivity gains from 0.83 to 0.855, leveraging its efficient network design. VGG-16 improved from 0.78 to 0.81 with its simpler architecture. The Hybrid (ViT + CNN) model outperformed all others, with sensitivity values increasing from 0.96 at 60 epochs to 0.98 at 90 epochs, demonstrating superior accuracy in detecting pneumonia cases.

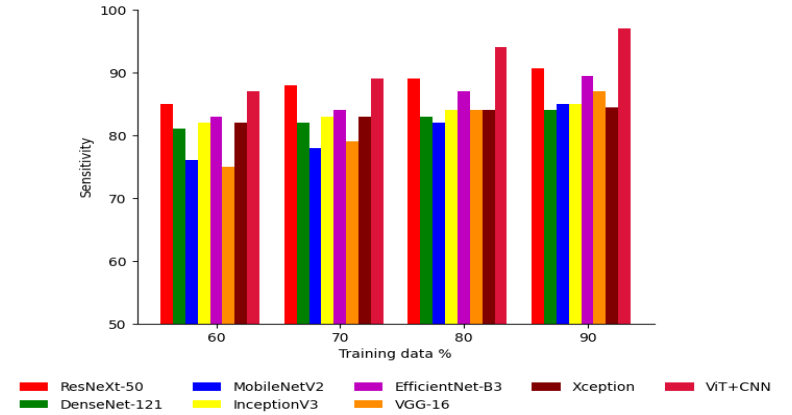

Figure 7: Comparisons of model sensitivity among various models with Hybrid (ViT + CNN)

4.5.5 F1_Score

The figure 8 illustrates the F1-score values of various deep learning models for pneumonia detection across four epochs: 60, 70, 80, and 90. ResNeXt-50 demonstrated an increase in F1-score from 0.825 to 0.855, indicating its effectiveness in achieving a balance between precision and recall. DenseNet-121 improved from 0.815 to 0.84, showing consistent enhancement in overall model performance. MobileNetV2, although starting lower at 0.765, progressed to 0.805, making it appropriate for settings with little processing power. InceptionV3 and Xception displayed steady improvements, reaching 0.855 and 0.85, respectively, by 90 epochs, indicating their robust performance across epochs. EfficientNet-B3 achieved F1-score gains from 0.835 to 0.86, benefiting from its efficient architecture. VGG-16 improved from 0.785 to 0.815 with its simpler design. The Hybrid (ViT + CNN) model consistently outperformed all others, with F1-score values consistently above 0.95 across all epochs, demonstrating superior performance in achieving both precision and recall in pneumonia detection tasks.

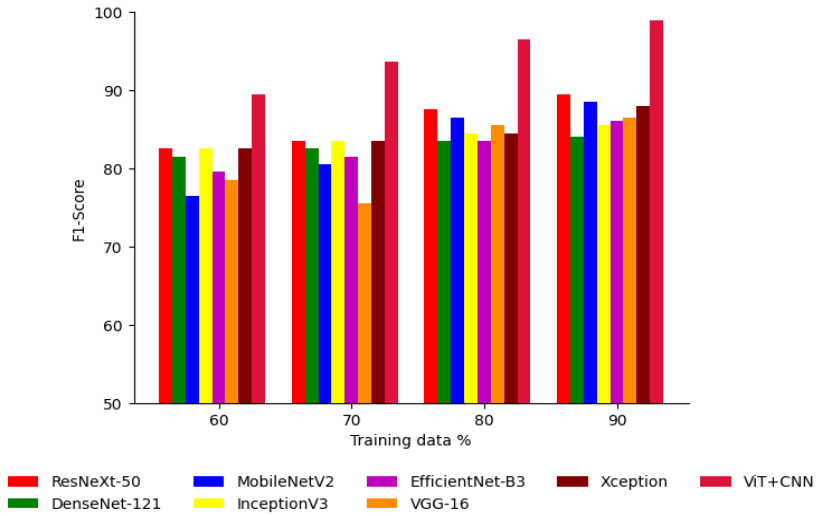

Figure 8: Comparisons of F1_Score among various models with Hybrid (ViT + CNN)

4.5.6 ROC

The ROC curves showing how different deep learning models perform at different thresholds for pneumonia diagnosis are shown in Figure 9. The trade-off between specificity (true negative rate) and sensitivity (true positive rate) is shown by each curve. ResNeXt-50 starts with a sensitivity of 0.05 and specificity of 0.2 at a threshold of 0.2, increasing to 0.5 sensitivity and 0.96 specificity at the highest threshold of 0.96. DenseNet-121 shows a similar trend, beginning with 0.04 sensitivity and 0.18 specificity at a threshold of 0.18, rising to 0.48 sensitivity and 0.95 specificity at 0.95 threshold. MobileNetV2, InceptionV3, EfficientNet-B3, VGG-16, Xception, and Hybrid (ViT+CNN) also exhibit varying degrees of sensitivity and specificity across thresholds. Notably, the Hybrid (ViT+CNN) model consistently achieves higher sensitivity and specificity compared to other models across the entire range of thresholds, indicating its superior performance in detecting pneumonia from chest X-ray images.

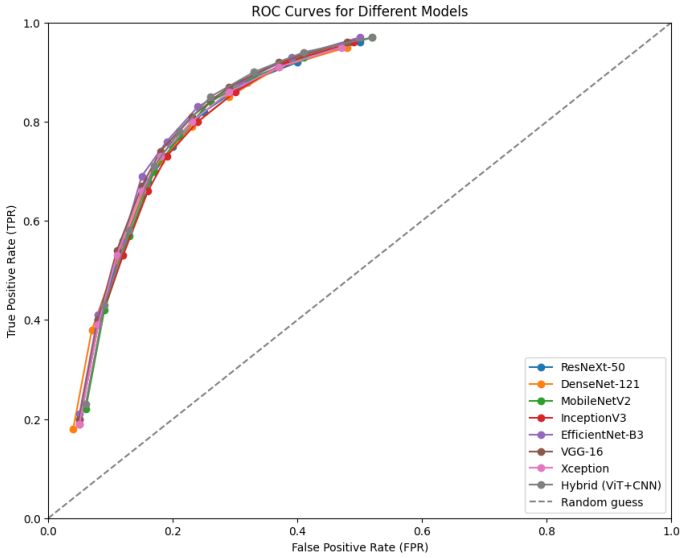

Figure 9: Comparisons of roc among various models with Hybrid (ViT + CNN)

5. Conclusion

Using chest X-ray pictures, this work systematically assessed the effectiveness of several deep learning models for pneumonia identification during several training epochs (60, 70, 80, and 90). ResNeXt-50 and DenseNet-121 demonstrated consistent accuracy improvements, showcasing their robust learning capabilities. MobileNetV2, designed for efficiency, showed gradual accuracy gains suitable for resource-constrained environments. InceptionV3 and Xception consistently achieved high accuracy, highlighting their effective feature extraction capabilities. EfficientNet-B3 exhibited notable performance enhancements owing to its scalable architecture. VGG-16, with a simpler structure, demonstrated steady but moderate accuracy improvements. The Hybrid (ViT + CNN) model emerged as the top performer, surpassing others with accuracy consistently exceeding 0.90 across epochs. This underscores the efficacy of integrating global and local feature extraction methods. The results are indicative of the potential of deep learning models, in particular the hybrid methods for increasing the accuracy of pneumonia detection from chest X-ray images and methods for clinical diagnostic workflow improvement.

Acknowledgment

I would like to express my very great appreciation to the co-authors in this manuscript, since their suggestions, either valuable or constructive, came in handy during the planning and development of this research work.

Conflicts of Interest

The authors declare no conflict of interest.

Ethical Approval

No humans or any living beings are involved in this research.

Consent to Participate

No humans or any living beings are involved in this research.

Funding Information

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

1. Abdulahi, A.T., Ogundokun, R.O., Adenike, A.R., Shah, M.A. and Ahmed, Y.K., 2024. PulmoNet: a novel deep learning based pulmonary diseases detection model. BMC Medical Imaging, 24(1), p.51.
2. Singh, S., Kumar, M., Kumar, A., Verma, B.K., Abhishek, K. and Selvarajan, S., 2024. Efficient pneumonia detection using Vision Transformers on chest X-rays. Scientific Reports, 14(1), p.2487.
3. Qiu, J., Mitra, J., Ghose, S., Dumas, C., Yang, J., Sarachan, B. and Judson, M.A., 2024. A Multichannel CT and Radiomics-Guided CNN-ViT(RadCT-CNNViT) Ensemble Network for Diagnosis of Pulmonary Sarcoidosis. Diagnostics, 14(10), p.1049.
4. Heidari, M., Azad, R., Kolahi, S.G., Arimond, R., Niggemeier, L., Sulaiman, A., Bozorgpour, A., Aghdam, E.K., Kazerouni, A., Hacihaliloglu, I. and Merhof, D., 2024. Enhancing Efficiency in Vision Transformer Networks: Design Techniques and Insights. arXiv preprint arXiv:2403.19882.
5. Rudnicka, Z., Proniewska, K., Perkins, M. and Pregowska, A., 2024. Health Digital Twins Supported by Artificial Intelligence-based Algorithms and Extended Reality in Cardiology. arXiv preprint arXiv:2401.14208.
6. Alharbi, A.H. and Hosni Mahmoud, H.A., 2022, May. Pneumonia transfer learning deep learning model from segmented X-rays. In Healthcare (Vol. 10, No. 6, p. 987). MDPI.
7. Lafraxo, S., El Ansari, M. and Koutti, L., 2024. A new hybrid approach for pneumonia detection using chest X-rays based on ACNN-LSTM and attention mechanism. Multimedia Tools and Applications, pp.1-23.
8. Masud, M., Bairagi, A.K., Nahid, A.A., Sikder, N., Rubaiee, S., Ahmed, A. and Anand, D., 2021. A pneumonia diagnosis scheme based on hybrid features extracted from chest radiographs using an ensemble learning algorithm. Journal of Healthcare Engineering, 2021(1), p.8862089.
9. Yunusa, H., Qin, S., Chukkol, A.H.A., Yusuf, A.A., Bello, I. and Lawan, A., 2024. Exploring the Synergies of Hybrid CNNs and ViTs Architectures for Computer Vision: A survey. arXiv preprint arXiv:2402.02941.
10. Nalluri, S. and Sasikala, R., 2024. Pneumonia screening on chest X-rays with optimized ensemble model. Expert Systems with Applications, 242, p.122705.
11. Mann, P.S., Panchal, S.D., Singh, S., Saggu, G.S. and Gupta, K., 2024. A hybrid deep convolutional neural network model for improved diagnosis of pneumonia. Neural Computing and Applications, 36(4), pp.1791-1804.
12. Avola, D., Bacciu, A., Cinque, L., Fagioli, A., Marini, M.R. and Taiello, R., 2022. Study on transfer learning capabilities for pneumonia classification in chest-x-rays images. Computer Methods and Programs in Biomedicine, 221, p.106833.
13. Ranpariya, D., Parikh, P., Patel, M.I. and Gajjar, R., 2022, February. A CNN based Hybrid

Model for Pneumonia Classification Using Chest X-ray Images. In 2022 2nd International Conference on Artificial Intelligence and Signal Processing (AISP) (pp. 1-7). IEEE.
14. Ukwuoma, C.C., Qin, Z., Heyat, M.B.B., Akhtar, F., Bamisile, O., Muaad, A.Y., Addo, D. and Al-Antari, M.A., 2023. A hybrid explainable ensemble transformer encoder for pneumonia identification from chest X-ray images. Journal of Advanced Research, 48, pp.191-211.
15. Asif, S., Wenhui, Y., ur-Rehman, S., ul-ain, Q., Amjad, K., Yueyang, Y., Jinhai, S. and Awais, M., 2024. Advancements and Prospects of Machine Learning in Medical Diagnostics: Unveiling the Future of Diagnostic Precision. Archives of Computational Methods in Engineering, pp.1-31.
16. An, Q., Chen, W. and Shao, W., 2024. A Deep Convolutional Neural Network for Pneumonia Detection in X-ray Images with Attention Ensemble. Diagnostics, 14(4), p.390.
17. Fathia, A., 2024. Hybrid Deep Learning Models for Enhanced Medical Image Classification.
18. Wang, Z., Chetouani, A., Jarraya, M., Hans, D. and Jennane, R., 2024. Transformer with Selective Shuffled Position Embedding and key-patch exchange strategy for early detection of Knee Osteoarthritis. Expert Systems with Applications, 255, p.124614.
19. Lafraxo, S., El Ansari, M. and Koutti, L., 2024. A new hybrid approach for pneumonia detection using chest X-rays based on ACNN-LSTM and attention mechanism. Multimedia Tools and Applications, pp.1-23.
20. Yunusa, H., Qin, S., Chukkol, A.H.A., Yusuf, A.A., Bello, I. and Lawan, A., 2024. Exploring the Synergies of Hybrid CNNs and ViTs Architectures for Computer Vision: A survey. arXiv preprint arXiv:2402.02941.
21. Pacal, I., 2024. MaxCerVixT: A novel lightweight vision transformer-based Approach for precise cervical cancer detection. Knowledge-Based Systems, 289, p.111482.
22. Rajaraman, S., Zamzmi, G., Yang, F., Liang, Z., Xue, Z. and Antani, S., 2024. Semantically redundant training data removal and deep model classification performance: A study with chest X-rays. Computerized Medical Imaging and Graphics, 115, p.102379.
23. Butt, M.H.F., Li, J.P., Ahmad, M. and Butt, M.A.F., 2024. Graph-infused hybrid vision transformer: Advancing GeoAI for enhanced land cover classification. International Journal of Applied Earth Observation and Geoinformation, 129, p.103773.
24. Kaya, M. and Çetin-Kaya, Y., 2024. A novel ensemble learning framework based on a genetic algorithm for the classification of pneumonia. Engineering Applications of Artificial Intelligence, 133, p.108494.
25. Varahagiri, S., Sinha, A., Dubey, S.R. and Singh, S.K., 2024. 3D-Convolution Guided Spectral-Spatial Transformer for Hyperspectral Image Classification. arXiv preprint arXiv:2404.13252.