Journal of Scientific & Industrial Research
Vol. 84, April 2025, pp. 421-434
DOI: 10.56042/jsir.v84i04.14040

PneuSwin: An Advanced X-Ray-Based Diagnostic System Integrating Ensemble Deep Learning Architectures and Swin Transformers for Pneumonia Detection

Sunil Kumar[1,2]* & Harish Kumar[1]

[1]Department of Computer Engineering, J.C. Bose University of Science and Technology, YMCA, Faridabad 121 006, Haryana, India

[2]Department of Information Technology, School of Engineering and Technology (UIET), CSJM University, Kalyanpur, Kanpur 208 024, Uttar Pradesh, India

Received 28 September 2024; revised 17 December 2024; accepted 14 February 2025

Pneumonia continues to pose a considerable global health concern, characterized by elevated fatality rates globally. X-rays are the primary radiological imaging technique for detecting pneumonia because of their widespread availability and inexpensive cost in medicine. Researchers have employed a variety of Deep Learning (DL)-based procedures to solve the issue, but only a small number of studies have amalgamated DL methods with Swin transformers. The Swin Transformer, distinguished for its ability to capture long-range dependencies and spatial associations, subsequently processes the refined features. This investigation introduces PneuSwin, a novel X-ray-based diagnostic system for efficient detection. The study utilized the PneuData dataset, a composite of three public X-ray datasets. It had Bacterial Pneumonia (BP), Viral Pneumonia (VP), and normal classes with balanced instances. Initially, the study paired various DL architectures (CapsNet, DenseNet-121, EfficientNet-B3, and ResNet-101) with the swin transformer. The DLs extracted relevant features from the PneuData images, sent them to Principal Component Analysis (PCA) to diminish their dimension and pick out the relevant features, and then inserted them as patches into the swin transformer for either binary or multi-class classification. Conversely, the PneuSwin concatenates the retained features, transferring them as a feature matrix to the PCA for relevant features and feeding them to the swin transformer. In both binary and multi-class classification, PnewSwin outperformed in comparison to ensemble DLs. In binary, PneuSwin had 97.21% accuracy, a 96.95% F1 score, and a 0.967 AUC, while in multi-class it had 97.67% accuracy, a 97.31% F1 score, and a 0.973 AUC. The results indicate that PneuSwin is proficient in detecting pneumonia in both binary and multi-class classifications.

Keywords: Classification, Convolutional neural network, Machine learning, Self-attention, Vision transformer

Introduction

Lower respiratory infections constitute a major worldwide health hazard, ranking as the fourth greatest cause of mortality. It encompasses several pulmonary disorders that affect the lungs, with pneumonia being the most renowned.[1,2] Various microorganisms, such as viruses, fungi, and bacteria, can cause pneumonia, which can result in fluid accumulation and inflammation in the respiratory system.[3] Pneumonia could be life-threatening in susceptible groups, including elderly individuals, young kids, and people with weaker immune systems. As a consequence of the fact that it is accountable for fourteen percent of the overall mortality rate among children under the age of five, which led to the deaths of nearly eight lakhs of children in 2019, and twenty-two percent of the total number of casualties included kids between the ages of one and five years old.[4]

Pneumonia is categorized into two primary types: Viral Pneumonia (VP) and Bacterial Pneumonia (BP). The two pneumonia forms exhibit comparable features, making distinction essential for precise diagnosis. VP is induced by several viruses, including Respiratory Syncytial Virus (RSV), influenza, adenoviruses, and prominently, SARS-CoV-2, which has been pivotal in the COVID-19 pandemic.[5] As reported by the World Health Organization (WHO), by 2024, COVID-19 had impacted over millions of individuals and resulted in approximately 7 million deaths.[6] In contrast, BP arises from bacterial infections that lead to inflammation and fluid accumulation in the lungs, often resulting in more severe symptoms than VP.

Radiography in medicine involves procedures and technology that visually represent the physique inside. Radio images allow Practitioners to diagnose and identify illnesses and their developmental states accurately. The healthcare sector relies on it for

*Author for Correspondence
E-mail: sunilymca24@gmail.com

diagnosis and surgery procedures. Various imaging methods, including chest X-rays (CXRs)[7], PET, CT scans[8], and MRI[9], are essential for diagnosing and evaluating lung disorders. CXRs serve as the primary imaging modality for initial patient assessment, providing a cost-effective and accessible method for identifying lung diseases.[10] Digital X-rays are the most common and contemporary lung imaging test.[7]

X-rays can detect pneumonia by identifying lung densities and white spots due to fluid or inflammation. BP and VP X-rays are similar but differ in consolidation, involvement pattern, and air bronchograms. VP may be patchy and impact multiple locations, while BP consolidates in one lobe or segment. BP often exhibits a localized pattern of involvement, impacting an entire lobe.[3–5] Air-filled airways are more prevalent in BP than VP.[8,9] In medicine, a clinician evaluates a single X-ray, while a Computer-Aided Diagnostic (CAD) system needs a bunch of X-ray (dataset) images to process.[11] Public access to X-ray datasets is crucial for medical research advancement and the development and validation of new CAD systems.[12] The analysis consolidated three publicly accessible datasets.[13–15] into a single, unique, and balanced dataset, PneuData, to amalgamate pneumonia and normal images. The binary classification of pneumonia and normal instances, as well as the multiclass classification of BP, VP, and normal instances, are illustrated in Fig. 1.

Utilizing DL architectures has the potential to enhance the computational structure, interpretability, and decision-making of the pneumonia CAD diagnostic system on X-rays.[8,16] Convolutional Neural Network (CNN) is a fundamental subcategory of DL and is often utilized in classification of images. CNNs are effective at using convolutional and pooling

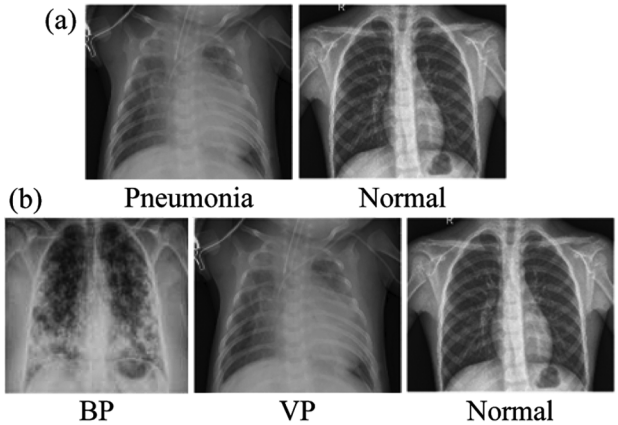

Fig. 1 — CXRs instances from the publicly accessible datasets[13–15]: (a) Binary classification, (b) Multi-class classification

methodologies to automatically extract valuable features from images.[17] The early diagnosis of pneumonia can be enhanced by CNNs because they are able to identify important CXR patterns based on computed features. CNNs excel in CXR processing due to extensive dataset training and their hierarchical feature learning capabilities, making them ideal for tasks involving spatially structured input data.[18]

Some major CNN designs used in pneumonia diagnosis utilizing X-rays are Capsule Network (CapsNet), DenseNet, EfficientNet, and ResNet. We analyzed the previously stated designs based on computational cost and performance and chose the traditional CapsNet[19], DenseNet-121[20], EfficientNet-B3[21], and ResNet-101[(22)] for the purpose.

The Transformer was first created for Natural Language Processing (NLP) tasks but has been adapted to fulfill computer vision tasks and introduces Vision Transformer (ViT).[23,24] Swin Transformer, a specific variation, is a highly effective DL approach that has gained significant popularity in the region. It is renowned for effectively handling long-range dependencies and utilizes a hierarchical architecture that partitions the input image into smaller patches, allowing it to capture both local and global features.[25] The Swin Transformer excels in pneumonia classification, requiring a comprehensive analysis of lung areas for precise determination. It uses its ability to capture complex patterns based on features at various scales to analyze X-rays and achieve precise classifications.[26,27]

Despite advancements in pneumonia detection through DL, there is a significant gap in integrating Swin Transformers with existing DL approaches. Swin Transformers' hierarchical representation and rapid computation of attention mechanisms could make it easier for models to identify relevant patterns in X-ray images, leading to more accurate diagnoses.[18] The limited exploration of ensemble models that combine DL methods and Swin Transformers presents an area for investigation, potentially leading to new solutions.[27] Coupling DL models with Swin Transformers is essential to leverage their complementary strengths, enhancing feature extraction and classification. This ensemble approach can lead to more robust and reliable diagnostic outcomes in the task.

The experiment showcases a novel integration of the specified DL architectures with a Swin Transformer. Each DL architecture uses its own convolutional layers to extract complex features and

feed them to the PCA[28], streamlining it while keeping the relevant features for the swin transformer. The hierarchical structure and attention mechanism format of the Swin Transformer acquired features as patches, which improved CXR classification and demonstrated the effectiveness of the ensemble approach.

PneuSwin is an ensemble approach that combines features from specified DL architectures, PCA, and Swin transformers. It concatenates the features into a unified feature vector, then integrates it into a suitable feature matrix. We apply PCA to identify the most significant components and features from the supplied ones, then feed them into the Swin transformer. The Swin transformer processes the input as patches and learns to classify the data, either for binary or multi-class classification. This processing enhances the *PneuSwin's* ability to derive significant representations from combined features, capturing local and global data.

The investigation aims to tackle the substantial issue by developing and evaluating the proposed image classification diagnostic system to accurately predict pneumonia cases using X-ray images. The substantial contributions of the investigation are listed below:

- The investigation utilized amalgamated pneumonia and normal images by combining three publicly accessible datasets into only one distinctive Pneudata dataset.
- The investigation looked at advanced approaches like the ensembling of specified DL architectures with Swin transformers to evaluate the effectiveness of the proposed PneuSwin diagnostic system.
- The investigation comprehensively evaluates DL techniques such as CapsNet, DenseNet-121, EfficientNet-B3, ResNet-101, and Swin transformers, highlighting their effectiveness in accurately classifying pneumonia (binary and multi-class) and discussing essential factors to improve the diagnostic process.

The investigation is structured as follows: Section II discusses the task-related research. Section III elaborates on the applied approach, which includes explanations on the Pneudata dataset, preprocessing, DLs, the ensemble network, and the PneuSwin. Section IV elaborates on the results and discussions. Section V presents the conclusion and future directions

Related Work

The literature review highlights a significant gap in research regarding the integration of Swin Transformers with existing DLs for pneumonia identification. While there is a recognized need for this integration, limited studies on ensemble models that combine these approaches indicate a potential area for further exploration. Research in this domain could potentially yield innovative solutions. This section evaluates previous research on X-ray imagery for diagnosing pneumonia using DL architectures, transformers, and their ensembling in binary and multiclass classification. It provides a comprehensive analysis of the current research landscape, highlighting contributions and research gaps in the modern study's approach, illustrated in Table 1.

DL Architectures

In binary classification, the researchers developed a DL model that included a redesigned convolutional layer with an exclusive dropout, which was tested under various attrition rates from 10% to 50%. The study found that the 40% dropout rate yielded the highest success.[18] The explored study aimed to develop a DL diagnostic system capable of handling five degrees of Gaussian noise in images for six different types of CXRs. The DL system showed no significant decrease in effectiveness when comparing the original dataset with the other datasets.[29] The research used deep transfer learning to combine three DL models, GoogLeNet, ResNet-18, and DenseNet-121, using Kermany[13] and RSNA[30] datasets. However, the ensemble structure struggled to produce reliable forecasts.[31] The investigation used CNNs like DenseNet, VGGNet, and EfficientNet, using transfer learning techniques with pre-trained ImageNet. Data was preprocessed using contrast-limited adaptive histogram equalization (CLAHE) and bi-histogram equalization (BEASF). The employed dataset needed to be larger and tested with fewer performance metrics. Some CNN systems need improvement for effective decision-making, as they provided incorrect predictions.[32] The ChxCapsNet framework, which employed convolutional and capsule layers, can effectively assessed pediatric pneumonia, using abstract details and low-level hidden features for general predictions, and the InceptionV3-convolution model outperformed with others.[33]

Table 1 — Overview of the related existing investigations

Method	#Class	Ref.	Labeled X-rays	Employed Architectures	Performance Metrics (%)
DL Architectures (DL + Transfer Learning + Ensembling)	Binary	18	P: 4273 N: 1583	CNN + Varied & Modified Dropouts	Acc - 97.2, R – 97.30, Pr – 97.40, F1 – 97.40, AUC – 0.982
		29	P: 4273, N: 1583	CNN	Acc – 97.60, Sen –98.20, Sp – 98.70
		31	P: 4273 N: 1583	ResNet-18 + GoogLeNet + DenseNet-121	Acc - 98.81, R – 98.80 Pr – 98.82, F1– 98.35
		32	P: 747, N: 754	DenseNet121, VGGNet, EfficientNet	F1: 99.77
		33	P: 4275, N: 1583	ChxCapsNet: InceptionV3 + CapsNet	Acc - 94.84, R - 97.73, F1- 95.93
		34	BP: 2530, VP: 1345 COV: 797, N: 5510	17 CNN Architectures	Acc – 99.85
		11	BP:2780, VP: 1493 COV: 474, N: 1583	CNN & Transfer learning	F1 – 84.46
		35	VP: 1656 COV: 1281 N: 3270	Ensemble DNN Model: CJT & Transfer Learning Model: DETL	CJT: Acc – 98.22, Sen – 98.37, Sp – 99.79 DETL: Acc – 97.26, Sen – 98.37, Sp – 100
	Multi	36	P: 4273, COV: 576 N: 1583	CNN + XAI + Grad-CAM, LIME, SHAP	Acc – 95.94, Sp – 95.71 ± 1.55 Sen– 95.50 ± 1.72, F1 – 96.53 ± 0.95
		37	VP: 1485 COV: 423 N: 1579	MobileNetv2, VGG19, ResNet101, CheXNet, ResNet18, SqueezeNet, Inceptionv3, DenseNet201	Binary: Acc - 99.70, Sen – 99.70 Pr – 99.70, Sp – 99.55 Multi: Acc - 97.90, Sen – 97.95 Pr – 97.90, Sp – 98.80
		16	VP: 1345, COV: 3616 N: 10192	CovidDWNet	Application 3: Acc – 96.81
		38	K-COVID: P: 6836, COV: 659, N: 6,629	K-EfficientNet	Acc - 97.3
		39	P: 250, COV: 250 N: 250	DenseCapsNet: DenseNet121 + CapsNet	Acc - 90.70 F1 - 90.9
		40	P: 4273 N: 1583	ViT	Acc – 97.61, Sen – 95.00 Sp – 98.00, F1 – 95.00
Transformers	Binary	26	Dataset 1: P: 4273 N: 1583 Dataset 2: P: 1062 N:84312	Swin Transformer	Dataset 1:Acc – 87.30 Dataset 2:Acc – 97.20
		41	VP: 1342, COV: 3466 N: 10083	DBM-ViT	Acc - 97.25, Sp – 97.97 Pr, Sen, F1 - 97.24
	Multi	42	RSNA: P: 30225	Radio Transformer	F1 - 98.75, AUC - 99.85
		43	P: 10702, COV: 10819 N: 10314	COVID-Transformer	Acc - 92, Pr – 93, R – 89, F1– 91, AUC – 98
		27	T: 22585	Swin Transformer & TNT	Acc& Sen – 94.75, Sp - 95.09
Ensembling (DL + Transformers)	Binary	44	P: 4273 N: 1583	ResNet-50 + Attention Mechanism + Focal loss	Acc – 98.68, F1 – 98.25 R &Sp – 96.68, Pr – 98.89
		45	P: 4273 N: 1583	ViT + DenseNet169 + MobileNetV2	Acc - 93.91, R – 92.99, Pr - 93.96, F1 - 93.43
	Multi	46	Three: P: 11263 Cov: 11956 N: 10701 Four: BP: 1485 VP: 1485 COV: 1485 N: 1485	ResNet18 + ViT	Binary: Acc – 99.32 Three: Acc – 95.16 Four: Acc – 90.03
	Binary & Multi	47	Binary: VP: 4290 N: 3834 Multi: VP: 5000 BP: 5000 N: 5000	CNN Ensemble + Transformer Ensemble A: VGG16 + DenseNet201 + GoogleNet Ensemble B: Xception + DenseNet201 + InceptionResNetV2	Binary:Acc – 99.21, F1 – 99.21 Binary (Ensemble): A - Acc – 97.22, F1 – 97.14 Multi (Ensemble): A - Acc – 97.20 F1 – 95.80

Legends- Acc: Accuracy, BP: Bacterial pneumonia, COV: COVID-19, F1: F1 Score, LIME: Local interpretable modelagnostic explanation, N: Normal, P: Pneumonia, Pr: Precision, R: Recall, Sen: Sensitivity, SHAP: SHapley additive explanation, Sp: Specificity, T: Total, TNT: Transformer in transformer, VP: Viral pneumonia, ViT: Vision transformer, XAI: Explainable AI

In multiclass classification, combining 17 CNN votes to optimize data efficiency was innovative. Among classifiers A, B, C, D, and E, A contrasted pneumonia with normal. B separated VP and BP. C differentiated COVID-19 from normal. D separated COVID-19 from BP. E compared normal to COVID-19.[34] Twelve DL models were trained to predict CXR images of healthy and VP/BP individuals. Model classification evaluations were conducted using different training data ratios, with 50%, 20%, and 10% being used. The unique Condorcet's Jury Theorem (CJT) approach determined ensemble classifier vote scores. Models in the voter pool could enhance majority accuracy using CJT. A CJT-based ensemble classification system was compared against a distinctive Domain Extended Transfer Learning (DETL) ensemble learner using a soft voting ensemble to determine which was better.[35] The implemented model is an explainable AI (XAI) and CNN fusion. Utilizing convolutional feature extraction, high-level and object-based data were gathered.[36] DLs, trained with image augmentation, outperformed shallow networks, while DenseNet-based CheXNet achieved comparable results without image augmentation, demonstrating the effectiveness of transfer learning and image augmentation.[37] Residual blocks resulting from feature reuse and deepness-wise expanded convolutional component portions comprise the CovidDWNet architecture. The obtained feature maps were efficient by employing the gradient boosting method in conjunction with the COVIDWNet architecture.[16] K-EfficientNet is an expanded iteration of EfficientNet. The K-COVID database is curated for analysis by amalgamating six publically accessible datasets.[38] DenseCapsNet is a framework for DL that integrates DenseNet with CapsNet, thereby minimizing the need for CNNs on extensive input.[39]

Transformers

In binary classification, global context and spatial connections derived from images are obtained using the ViT model, which blends self-attention methods and transformer layout. The ViT model was appropriate for global context, spatial connections, and multi-resolution images.[40] The analysis used the Swin Transformer, tailored for specific features, as the backbone network, and compared its experimental outcomes to CNNs.[26]

In multiclass classification, through depth wise convolutions, the DBM-ViT model effectively extracted global information from images, while it obtained local information by feeding feature maps with combined sequences.[41] RadioTransformer, a student-teacher transformer system, leveraged eye-gaze observation to assess cognitive complexity and recorded global and local image features. The system utilized a cascaded global-focal transformer structure to understand the visual search patterns of radiologists in their 'human visual attention regions'.[42] COVID-Transformer presented a ViT-based pipeline with baseline DLs. The process of creating the customized MLP block involved removing the pre-trained MLP prediction block and adding untrained feed-forward layers.[43] The swin transformer and the transformer in transformer were combined in an ensemble technique.[27]

Ensembling

In binary classification, to mitigate the impact of unbalanced training instances, the proposed method allocated greater weight to minority classes throughout the training process. Furthermore, the researchers incorporated improved focal loss into the suggested architecture. Findings demonstrated that the spatial and channel attention modules outperformed ResNet-50.[44] DenseNet169, MobileNetV2, and ViT are CNNs that have been fine-tuned. The outcomes are derived by integrating the retrieved features from the three CNNs during the feature engineering, followed by classification.[45]

In multiclass classification, PneuNet, a VIT-based approach, used channel-based attention for image diagnosis, focusing multi-head attention on channel patches rather than features. The 64:16:20 dataset distributions for training, validation, and testing may cause biases.[46]

The recommended fusion approach combined the DL and transformer encoder. Ensemble Model A combined DenseNet201, VGG16, and GoogleNet; Ensemble B included DenseNet201, Inception ResNetV2, and Xception. The foundation used two independent ways to extract significant features from CXRs, whilst the transformer encoder relies on the MLP self-attention mechanism to achieve correct diagnosis.[47]

Material and Methods

The methodology presents the detailed use of the created PneuData dataset, exploration of DL architectures with the Swin transformer, integration of

the DL architectures with the Swin transformer, and the recommended PneuSwin framework in developing an effective pneumonia diagnostic system.

Dataset

The generated PneuData dataset employed in the investigation has a balanced number of instances from multiple publicly accessible databases, such as LDOCTCXR[13], Curated[14], and Balanced Augmented Covid CXR[15] (Table 2). In order to achieve balance in PneuData, we selected 2632 instances of viral pneumonia and 6709 normal instances from the Balanced Augmented Covid CXR Dataset. Although the PneuData instances are balanced, they do not address data unbalancing issues, which is the principal concern for this sort of assignments.

Only the (c) and (d) columns of Table 2 are employed for binary classification, which comprises labeled pneumonia instances from bacterial, viral pneumonia, and normal instances. For the Multi-class classification (a), (b), and (d) of Table 2 is used, which includes bacterial, viral pneumonia, and normal instances.

Preprocessing

This research investigates various preprocessing techniques for diagnosing pneumonia, including noise reduction, image resizing, CLAHE, and normalization. Noise reduction is carried out using a Gaussian filter, which effectively reduces noise and improves the quality of visual data.[48] Images of differing dimensions were resized to 224 × 224 pixels for the study. The CLAHE technique adjusts intensity levels using local histograms, improving contrast and revealing essential structural features. All of the images were subjected to normalization using a ratio of 1.0/255.[(49)]

Deep Learning Architectures

CapsNet[19], DenseNet-121[(20)], EfficientNet-B3[(21)], and ResNet-101[(22)] DL architectures have been significant in the area of pneumonia identification utilizing CXR images. Through the use of distinctive design concepts and sophisticated features, our objective was to augment the efficacy of the proposed diagnostics system.

Table 2 — PneuData Instances

Dataset	BP (a)	VP (b)	P = BP + VP (c)	Normal (d)	Total
LDOCTCXR[13]	2780	1493	4273	1583	5856
Curated[14]	3001	1656	4657	3270	7927
Balanced[15]	—	2632	2632	6709	9341
Total	5781	5781	11562	11562	23124

CapsNets enhance image processing by using "capsules" of neurons to encode features and capture spatial relationships.[19] They employ routing algorithms for effective communication between capsules, adapting based on agreement. CapsNets utilize output vectors for image attributes, accommodating variations in perspective, scale, and deformation, unlike traditional CNNs.[33,39]

DenseNet-121 is a dense convolutional network that promotes feature reuse and accelerates deep learning model development. It employs direct connections between nodes, reducing parameters and enhancing feature distribution.[20,31] The architecture includes convolutional and transition layers to lower dimensionality and complexity, effectively addressing the problem of blurred gradients in deep networks.[32, 47]

EfficientNet-B3 employs compound scaling to optimize the network's depth, breadth, and resolution, resulting in compact, accurate models.[21] It integrates advanced design elements like inverted bottleneck and squeeze-and-excitation blocks, enhancing efficiency and performance while minimizing computational demands, making it ideal for resource-constrained environments such as edge computing.[32,38]

Residual network-101 (ResNet-101) is a deep learning architecture that employs shortcut connections for residual learning, addressing degradation issues and improving accuracy with increased depth. It consists of 101 layers, including multiple residual blocks and convolutional layers, which effectively capture complex data features. By focusing on residual mapping rather than direct input mapping, ResNet-101 allows for efficient training.[22,31] The ResNet-101 design is advantageous due to its capacity to acquire the identity mapping (i.e., $Y = X$) when necessary. The architecture maintains consistent parameters and computational costs with traditional networks, and it can approximate identity mapping, optimizing the training process through weight adjustments in convolutional layers.[37,49]

Swin Transformer

The Swin Transformer, a breakthrough in ViTs, enhances image comprehension efficiency and has shown promising results in pneumonia detection. Its robust attention mechanism and hierarchical feature representation[25] accurately detect intricate patterns and anomalies in X-ray images, detecting subtle infection indications like infiltrates and consolidation. After extensive training, it has demonstrated remarkable reliability in distinguishing healthy lungs

from those affected by pneumonia.[26–28] The diagnostic process of pneumonia is presented through Fig. 2.

The Swin Transformer is renowned for its hierarchical methodology and dynamic shift-based window system. To get localized processing and modest computing in hierarchical methodology, the input image I ∈ℝ^ (h × w × c), where ℝ represents the computed value and h, w, and c are the height, width, and channels, respectively, they are fragmented into p × p patches that don't overlap, resulting in a grid of patches with dimensions (h' × w'), where h' = h/p and w' = w/p.

Then linearly project each patch with a weight matrix $w_e \in \mathbb{R}^{\wedge}$ ($D_{in} \times D_{emb}$) to get the patch embeddings $E \in \mathbb{R}^{\wedge}$ (h' × w' × D_{emb}), where Din is the input dimension and $D_{emb}$is the embedding dimension. E is associated with position embedding $P_E \in \mathbb{R}^{\wedge}$ (h' × w '× D_{emb}). The E is added element-wise with the P_E, resulting a modified E':

$$E' = E + P_E \quad \ldots (1)$$

These patches are intelligently merged as the model advances through its phases to create a hierarchical visual representation. The computation of subsequent Swin Transformer blocks is achieved via the shifting window partitioning technique Eqs (2) – (5).[25]

$$\hat{f}^{b} = \mathrm{WMSA}\left(\mathrm{LNorm}\left(f^{b-1}\right)\right) + f^{b-1} \quad \ldots (2)$$

$$f^{b} = \mathrm{MLP}\left(\mathrm{LNorm}\left(\hat{f}^{b}\right)\right) + \hat{f}^{b} \quad \ldots (3)$$

$$\hat{f}^{b+1} = \mathrm{SWMSA}\left(\mathrm{LNorm}\left(f^{b}\right)\right) + f^{b} \quad \ldots (4)$$

$$f^{b+1} = \mathrm{MLP}\left(\mathrm{LNorm}\left(\hat{f}^{b+1}\right)\right) + \hat{f}^{b+1} \quad \ldots (5)$$

Let $\hat{f}^{b}$and f^{b} represent the output features of the WMSA and SWMSA module and the MLP module for block b, respectively. LNorm is stand for linear normalization.

An essential element is introduced in succeeding layers of the Swin Transformer: the windows are consciously "shifted." By performing this ostensibly uncomplicated transition, tokens located in distinct windows can interact in communication and share data throughout the whole of the image, thereby surpassing the constraints associated with attention mechanisms (Eq.6) that are solely localized.[25, 44]

$$\mathrm{Attention}(Q, K, V) = \mathrm{SoftMax}\left(\frac{QK^{T}}{\sqrt{D_{emb}}}\right)V \quad \ldots (6)$$

Here, the query, the key, and the value that are represented by the indices Q, K, and V, accordingly. Next, The Q_i, K_i, and V_i matrix associated with each head is subsequently utilized to calculate the related attention score using Eq. (7).

$$\mathrm{head_i} = \mathrm{Attention}\left(Q * W_i^{Q}, K * W_i^{K}, V * W_i^{V}\right) \quad \ldots (7)$$

Here the matrices with dimensions, D_{emb} equal to the number of heads (h) multiplied by the respective W_i^{Q}, W_i^{K}, and W_i^{V} multiples.

The MHA layer produces its final result by combining the outputs of all heads and performing a matrix-like complete conjunction, as outlined in Eq. (8):

$$\mathrm{MultiHead}(Q, K, V) = \mathrm{Merge}\left(\mathrm{head}_1, \mathrm{head}_2, \ldots, \mathrm{head_i}\right) \quad \ldots (8)$$

To achieve the outputs of the Swin Transformer, the residual link could be applied on before and after the MLP layers. This layer of the complete model is often formed by iteratively and consecutively stacking several transformers. The Swin Transformer incorporates the fundamental principle of windows multi-head self-attention (W-MSA), whereby the feature map is partitioned into distinct window areas that do not overlap. W-MSA refines patch

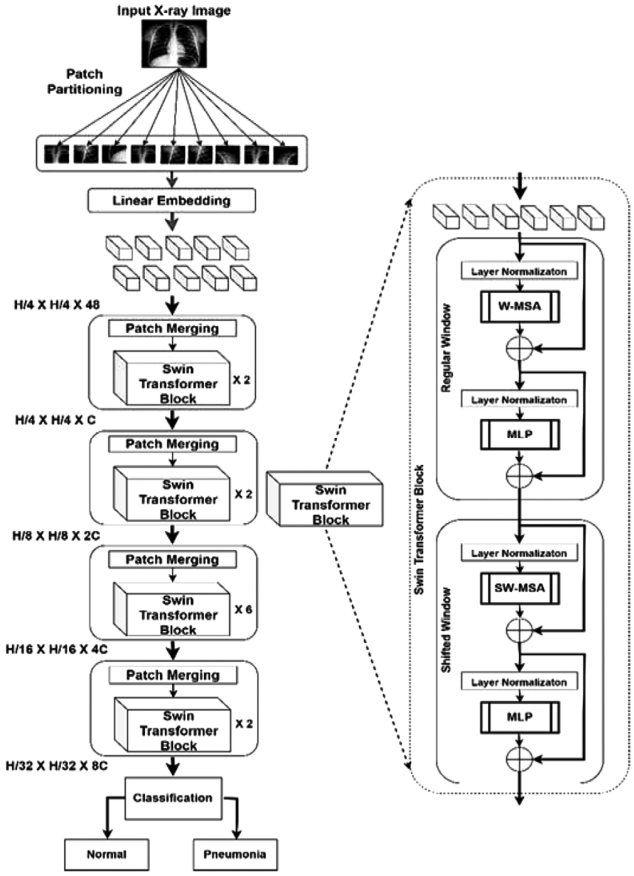

Fig. 2 — Pneumonia diagnosis through Swin Transformer

representations by identifying local connections in each window. Each window executes a distinct multi-head self-attention (MSA) operation. This significantly decreases the computational burden of performing MSA procedures on the feature map. To facilitate the exchange of information between neighboring windows, the Swin Transformer incorporates Shifted Windows Multi-Head Self-Attention (SW-MSA). The SW-MSA uses window shifting to convey global information across the image, promoting context awareness. Producing multi-scale feature maps involves sequentially down-sampling imagery at 4x, 8x, 16x, and 32x for the hierarchical processing, after each stage.[25–27]

Ensembling

The research employs the separate ensembling of each DL architecture i.e. CapsNet, DenseNet-121, EfficientNet-B3, and ResNet-101 with the Swin transformer. The DL architectures used in the investigation are used to get feature vectors for each image by removing the top classification layer. The Swin transformer is then used to classify. The DL architectures do feature extraction by passing input CXRs through their convolutional layers, specifically intended to extract more intricate features at various degrees of abstraction. These features are considered as patches, to capture the patterns and structures in the CXRs that are significant for classification.[18,27] After extracting the features (*F_Ci* – Features extracted by CapsNet, *F_Di* – Features extracted by DenseNet-121, *F_Ei* – Features extracted by EfficientNet-B3, *F_Ri* – Features extracted by ResNet-101), they were passed to the PCA for dimensionality reduction.[28] After that, the achieved outcomes could be transformed into a proper format for input into the Swin Transformer by flattening them. The Swin Transformer uses and manipulates these extracted features via its hierarchical design. The attention mechanism in the Swin Transformer lets it focus on essential aspects and relationships in the extracted features, which improves the classification (binary or multi-class) of the input CXRs.[27] The ensembling of DL architecture with the Swin transformer for the task is presented through Fig. 3.

PneuSwin

The proposed PneuSwin is an ensemble approach that combines the features (as patches) extracted by specified DL architectures, fed to PCA, and classification by Swin transformers. The PneuSwin methodology for the task is illustrated through Fig. 4.

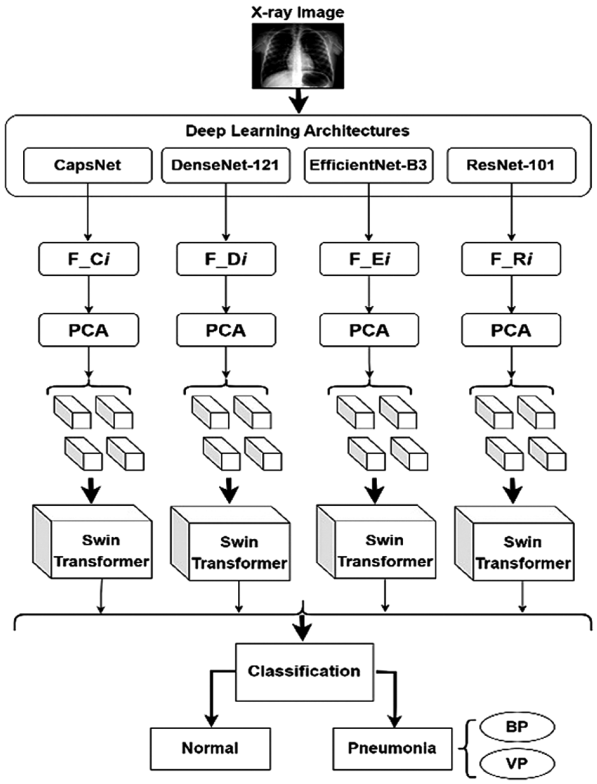

Fig. 3 — Ensembling of DL architecture with Swin transformer

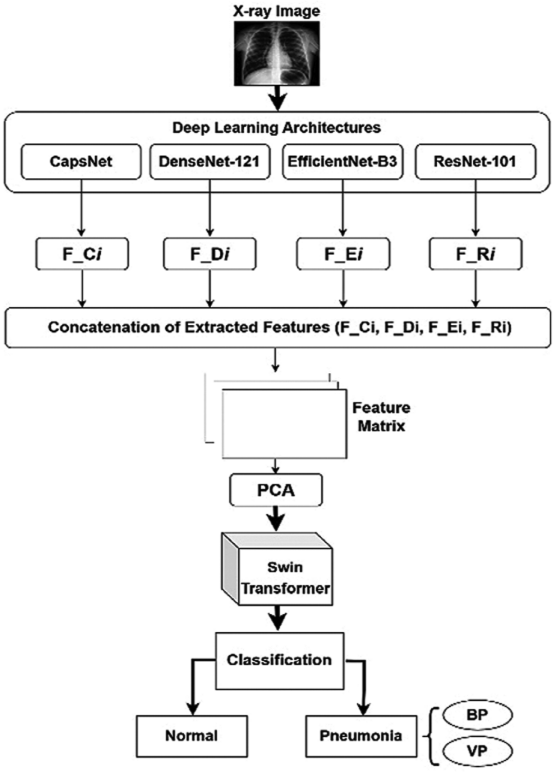

Fig. 4 — PneuSwin methodology for the task

PneuSwin concatenates the features calculated by each specified DL to form a unified feature vector. The feature vectors are tied into a suitable feature matrix in which each row corresponds to each image while each columnre presents a particular feature. The research applies PCA to the concatenated feature matrix to decrease dimensionality. PCA is a technique that identifies the most significant components of a feature matrix while simultaneously lowering the overall dimensionality. The outcomes of the PCA, a feature matrix with decreased dimensions, will be sent to the Swin transformer. The Swin transformer processes the feature matrix as input and learns to classify the data (either for binary or multi-class). The hierarchical processing of patches in Swin Transformer can offer advantages in integrating features retrieved by DL architectures. It enables capturing local and global data, thus enhancing the model's capacity to derive significant representations from the combined features. PneuSwin's algorithm is presented as Algorithm 1 for understanding the functioning of the framework as described above.

Algorithm 1: PneuSwin's training algorithm

PneuData: D = [X_k] //Input an image dataset
For each image a ∈ D = X_b //where b = [1:k]
1. a → ((Gaussian filter) AND (Resize (224 × 224))
AND(CLAHE) AND (Normalize i∈ (0, 1))) //Preprocessing
2. pre-trained DLs (a)→top_layer = False
3. For each Extracted_Features = DLs (a)
F_Ci=CapsNet(X_a),F_Di= DenseNet-121(X_a),F_Ei = EfficientNet-B3(X_a), F_Ri = ResNet-101(X_a)
4. fused_features = concatenation (F_Ci, F_Di, F_Ei, F_Ri) //Concatenation
5. fused_features→feature_matrix//PCA
feature_centered = feature_matrix - Mean (feature_matrix,axis)
Cov_mat = (Σ) (feature_centered)
λ, U = Eigen_Decomposition (Cov_mat)
Principal_components = Cov_mat > threshold
Selected_PC = top_eigenvectors (Principal_components)
reduced_features = U [: , Selected_PC]
6. Reduced_features→ PCA(feature_matrix)
7. PneuSwin = Swin_transformer(Reduced_features)
8.class_probabilities = Softmax (logits)
9. Obtained class predictions respective to the class.
Binary classification: Threshold >= class prediction
Multi-class classification: class = Highest class prediction

Performance Metrics

Performance metrics assess and quantify research model effectiveness. Different metrics for performance are utilized, including accuracy, precision, recall, F1 score, and Area Under the Receiver Operating Characteristic Curve (AUC).The study evaluates the accuracy of classification algorithms by calculating the proportion of true positives (TPs) and false positives (TNs) in total positive estimates. Precision is assessed by dividing the proportion of genuine positives by actual positives to assess sensitivity and the number of relevant samples.[48,49] The F1 score integrates recall and precision, while the AUC metric shows the relationship between true positive rate (TPR) and false positive rate (FPR) at different levels. The AUC variability can be estimated for multi-class classification using one-to-the-rest or average pairwise AUCs.[10] The metric and its mathematical foundation presented through Eqs (9) – (12).

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

$$\text{Recall} = \frac{TP}{TP+FN} \quad \ldots (10)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad \ldots (11)$$

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

where, TP: True Positive, TN: True Negative, FP: False Positive, FN: False Negative.

Results and Discussion

Hyper-parameters for Implementation

The designated DL models used the following hyper-parameters:

- The optimizer used is AdamW.
- A learning rate of 0.002 is observed.
- The loss function used is Categorical Cross-Entropy.
- There are 32 batches and 100 epochs for training.

AdamW, an Adam variant, addresses weight decay in DL models to reduce over fitting. A momentum-adaptive optimizer uses adaptive learning rates. The categorical cross-entropy is applicable and suitable for both binary and multi-class classification applications. The use of a batch size of 32 often results in a trade-off between model stability and computing performance. While it is generally acceptable to conduct 100 training epochs, it is crucial to regularly watch the validation loss in order to mitigate the risk of over fitting. In such a case, early pausing is applied.

Evaluation

This analysis included two forms of classification: binary classification of pneumonia and normal, and multi-class classification in the BP vs. VP vs. normal classes.

Binary classification

The binary classification test, which shows the dissimilarity between pneumonia and normal

instances, demonstrates a range of performance levels on the test section, i.e., 20% of the PneuData. Table 3 illustrates the outcomes for ensemble DLs with Swin transformers in contrast to PneuSwin.

In ensembling, the results suggested that the Resnet-101 with Swin Transformer (ResNet-101+ST) model achieved a significant degree of fastidiousness in correctly classifying. The accuracy, F1, and AUC of the ensemble method in distinguishing between normal and pneumonia instances were 95.14%, 95.71%, and 0.946, respectively. EfficientNet-B3 + ST ranks as the third-highest performer in terms of effectiveness, demonstrating strong performance. Although CapsNet + ST exhibit satisfactory performance, on the contrary, DenseNet-121 + ST exhibit the least desirable performance. The illustration of the ROC curve for each of the DLs can be seen in Fig. 5.

In contrast to these ensemble DLs, PneuSwin outperformed. The accuracy of our suggested PneuSwin model was found to be 97.21%. Additionally, the F1 and AUC for the pneumonia class were determined to be 96.95% and 0.967, respectively. The performance of class-wise PneuSwin on the PneuData dataset is shown in Table 4, which grounds the confusion matrix presented in Fig. 6. The accuracy and loss of the PneuSwin model is demonstrated in Fig. 7.

Table 3 — Outcomes of the DLs for the binary classification

DLs	Acc	Pr	R	F1	AUC
CapsNet + ST	91.74	89.98	96.03	92.62	0.906
DenseNet-121 + ST	89.85	87.89	87.02	86.72	0.869
EfficientNet-B3 + ST	93.67	92.54	93.76	92.96	0.935
ResNet-101 + ST	95.14	95.06	96.04	95.71	0.946
PneuSwin	97.21	97.22	96.71	96.95	0.967

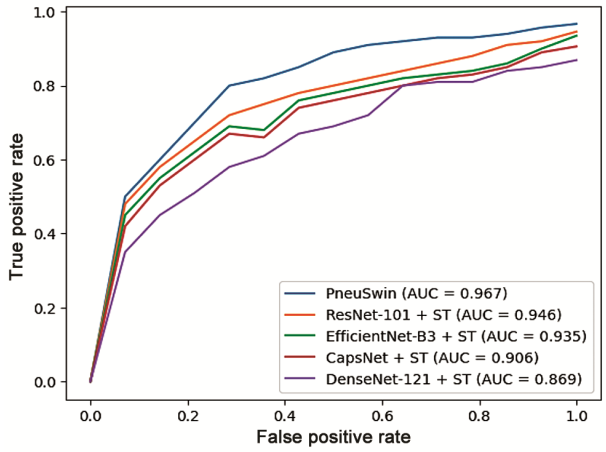

Fig. 5 — ROC graph for binary classification

Multi-class Classification

The findings of the multi-class classification test indicate the differentiation between BP, VP, and normal instances. 20% of the PneuData for each class is the test portion. DLs learned to recognize BP and VP forms in images. The outcomes for ensemble DLs in contrast to PneuSwin are illustrated in Table 5.

In ensembling, the findings of the study indicate that the EfficientNet-B3 + ST model demonstrated a notable level of precision in the classification process. The ensemble technique achieved an accuracy of 96.56%, an F1 score of 96.14%, and an AUC of 0.951 in differentiating between BP, VP, and normal instances.

The ResNet-101 + ST model has robust performance, positioning it as the third-highest performer in terms of performance. In contrast to DenseNet-121 + ST, CapsNet + ST demonstrate suboptimal performance. The illustration of the ROC curve for each of the DLs can be seen in Fig. 8.

The performance of PneuSwin surpassed that of these ensemble DLs. The PneuSwin model demonstrated an average accuracy, F1, and AUC values of 97.67%, 97.31%, and 0.973, respectively. The performance of class-wise PneuSwin for multi-

Table 4 — PneuSwin's Performance in the binary classification (%)

Class	Pr	R	F1	AUC
Normal	96.83	96.63	96.72	0.969
Pneumonia	97.61	96.78	97.19	0.965
Average	97.22	96.71	96.95	0.967
Overall Accuracy : 97.21				

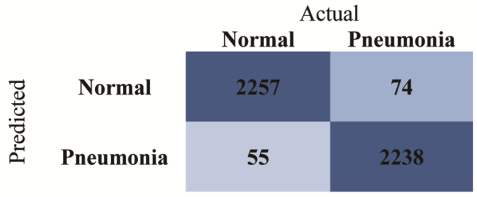

Fig. 6 — Confusion matrix of the PneuSwin (binary classification)

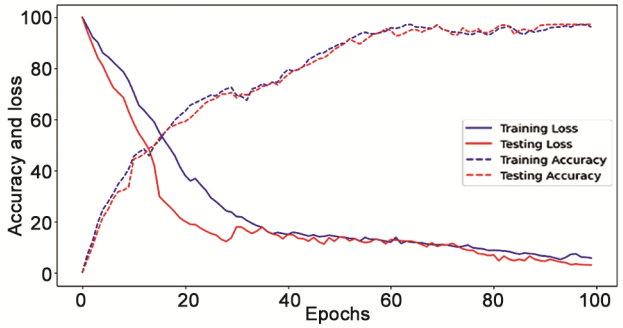

Fig. 7 — Accuracy and loss graph of the PneuSwin (binary classification)

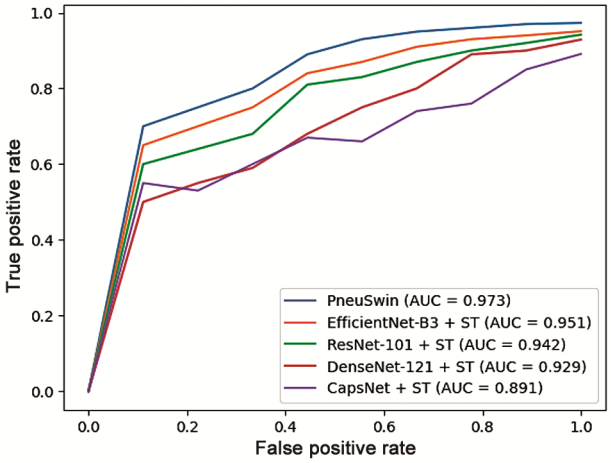

Fig. 8 — ROC graph for multi-class classification

Table 5 — Outcomes of the DLs for the multi-class classification (%)

DLs	Acc	Pr	R	F1	AUC
CapsNet + ST	92.98	86.01	96.59	91.02	0.891
DenseNet-121 + ST	93.64	88.61	94.88	91.62	0.929
EfficientNet-B3 + ST	96.56	94.88	97.34	96.14	0.951
ResNet-101 + ST	95.45	91.65	98.87	95.25	0.942
PneuSwin	97.67	97.37	97.35	97.31	0.973

Table 6 — PneuSwin's Performance in the multi-class classification (%)

Class	Pr	R	F1	AUC
Normal	98.58	98.53	98.54	0.964
BP	96.44	95.86	96.21	0.968
VP	97.08	97.74	97.29	0.981
Average	97.37	97.35	97.31	0.973
Overall Accuracy : 97.67				

class classification on the PneuData dataset displayed in Table 6 is providing the foundation for the confusion matrix in Fig. 9. The accuracy and loss of the PneuSwin model is demonstrated in Fig. 10.

Existing research shows various flaws that prompted PneuSwin. First, researchers focus on limited datasets for investigation, such as LDOCTCXR dataset, rather than recently produced datasets. This constraint may favor established datasets, which may not adequately reflect real-world data. Many studies concentrate on binary or multi-class classification problems, but few cover both concurrently. This gap is crucial because real-world applications generally demand models to classify data into several groups and discern binary outcomes. The transformer assembly of DLs, which may improve model performance, is understudied. The suggested research must fill gaps in the literature due to the limited number of studies on these issues.

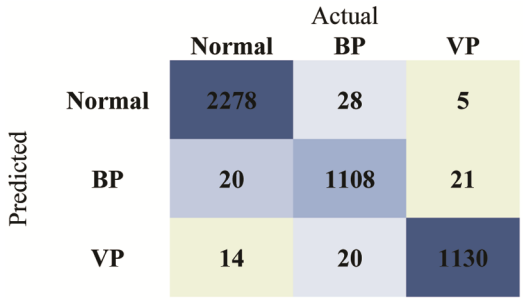

Fig. 9 — Confusion matrix of PneuSwin for multi-class classification

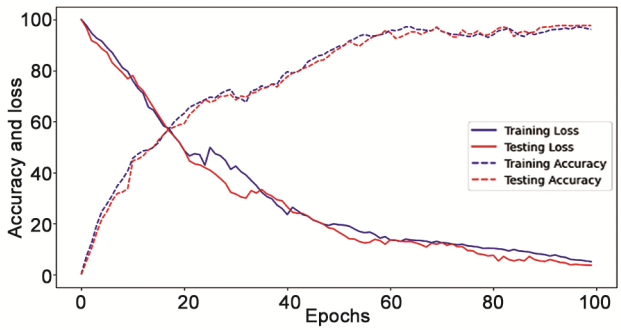

Fig. 10 — Accuracy and loss graph of the PneuSwin (Multi-class classification)

Ablation Study

The backbone of the proposed PneuSwin model is the Swin Transformer, which is well-suited for processing images. Different parts of the PneuSwin method that work well together are DL architectures like CapsNet, DenseNet-121, EfficientNet-B3, and ResNet-101's feature extraction and fusion, PCA, and Swin transformer for classification. Putting together different DL architectures with the Swin transformer is crucial because the multiple architectures allow the model to collect varied aspects and features and benefit from their complementary abilities by removing the top classification layer to obtain image features. Features are patches that encapsulate classification-relevant patterns and structures. By extracting features at different abstraction levels, the model can capture low-level and high-level image attributes, improving class discrimination. PCA reduces dimensionality to reduce computing complexity and over fitting; it prioritizes key components while preserving key data. It also eliminates redundant or noisy elements that impair classification. After dimensionality reduction, the Swin transformer learns to categorize data using the feature matrix. The Swin transformer integrates DL architectural features via hierarchical patch processing. This lets the model collect local and

Table 7 — PneuSwin's Performance in the multi-class classification (%)

DLs	#Instances	#Class	ACC	F1	Ref.
Swin-T	*D1*- P: 4273, N: 1583 *D2*- P: 1062, N:84312	2	*D1:*87.30 *D2:*97.20	—	26
DBM-ViT	VP: 1342, COV: 3466, N: 10083	3	97.25	97.24	41
ViT	P: 4273, N: 1583	2	97.61	95.00	40
COVID-Transformer	P: 10702, COV:10819, N: 10314	3	92.00	91.00	43
ViT + DenseNet169 + MobileNetV2	P: 4273 N: 1583	2	93.91	93.43	45
PneuSwin: DLs + PCA + Swin-T	P: 11562, BP: 5781, VP: 5781, N: 11562	2 & 3	Binary:97.21 Multi:97.67	Binary:96.95 Multi:97.31	Our

Legends: Swin-T: Swin Transformer, D: Dataset

global information from coupled features, improving representation learning and classification. The research demonstrates the significance of every element in the PneuSwin and emphasizes their individual contributions to the overall classification accuracy.

Comparative Analysis

An empirical analysis of pneumonia diagnosis through DLs, demonstrating that our investigation attained the highest level of effectiveness in contrast to other investigations is presented through Table 7.

A wide range of research focused on examining binary or multiclass classification in isolation. Although PneuSwin exhibited exceptional performance in binary and multi-class classification tasks, most of the research concentrated on the DLs, Swin Transformer, DBM-ViT, COVID-Transformer, and ViT variations, which also demonstrated remarkable performance. A noteworthy observation is that PneuSwin maintained elevated levels of accuracy and F1 scores consistently. PneuSwin exhibited outstanding performance in binary classification, successfully distinguishing between pneumoniaous and non-pneumonious instances. PneuSwin exhibited exceptional performance in multi-class classification, underscoring its effectiveness in accurately classifying instances of BP, VP, and normal conditions. The results underscored the potential of the PneuSwin model as a reliable and accurate diagnostic system for the identification of pneumonia.

Conclusions

To address pneumonia, we developed a novel diagnostic system based on an ensemble approach called PneuSwin using the PneuData dataset, which combines BP, VP, and normal X-ray images. PneuSwin adeptly employs sophisticated ensembling methods and concatenates features using multiple deep learning methods to optimize feature engineering. PCA decreases the range of features to facilitate relevant feature selection. The use of the Swin Transformer, using a hierarchical patch processing approach, enhances its capability to represent features effectively and facilitate accurate classifications. Conversely, deep learning ensemble methods such as the ResNet-101 combined with the Swin Transformer has shown enhanced efficacy in binary classification. The EfficientNet-B3 combined with the Swin Transformer model produced reliable outcomes for multi-class classification. In spite of the study's success, there are still constraints. PCA has shown use in feature engineering; nevertheless, the integration of newer methodologies might substantially improve model performance. Future research should explore PneuSwin's generalizability across various imaging modalities and datasets, as well as the impact of hyper-parameter optimization on its overall effectiveness.

References

1 World Health Organization, The top 10 causes of death, https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death (1 April 2024).

2 Pietrangelo A, The Top 10 Deadliest Diseases,Healthline, (2023), https://www.healthline.com/health/top-10-deadliest-diseases#respiratory-illness (10 April 2024).

3 Mackenzie G, The definition and classification of pneumonia, *Pneumonia*, **8(1)** (2016) 14, doi:10.1186/s41479-016-0012-z.

4 World Health Organization, Pneumonia in children, (2022), https://www.who.int/news-room/fact-sheets/detail/pneumonia (14 April 2024).

5 Causes and Risk Factors, NHLBI, NIH, (2022), https://www.nhlbi.nih.gov/health/pneumonia/causes(15 April 2024).

6 COVID-19 deaths, WHO COVID-19 dashboard, https://data.who.int/dashboards/covid19/deaths?n=c (1 April 2024).

7 Ng K H, & Rehani M M, X ray imaging goes digital, *BMJ*, **333(7572)** (2006) 765–766, doi: 10.1136/bmj.38977.669769.2c.

8 Domingues I, Pereira G, Martins P, Duarte H, Santos J & Abreu P H, Using deep learning techniques in medical imaging: A systematic review of applications on CT and PET, *Artif Intell Rev*, **53(6)** (2019) 4093–4160, doi: 10.1007/s10462-019-09788-3.

9 Jiang W, Ong F, Johnson K M, Nagle S K, Hope T A & Lustig M, Motion robust high resolution 3D free-breathing pulmonary MRI using dynamic 3D image self-navigator, *Magn Reson Med*, **79(6)** (2017) 2954–2967, doi: 10.1002/mrm.26958.

10 Kumar S, Kumar H, Kumar G, Singh S P, Bijalwan A & Diwakar M, A methodical exploration of imaging modalities from dataset to detection through machine learning paradigms in prominent lung disease diagnosis: A review, *BMC Med Imaging*, **24(1)** (2024) 30, doi: 10.1186/s12880-024-01192-w.

11 Avola D, Bacciu A, Cinque L, Fagioli A, Marini M R & Taiello R, Study on transfer learning capabilities for pneumonia classification in chest-x-rays images, *Comput Methods Programs Biomed*, **221** (2022) 106833, doi: 10.1016/j.cmpb.2022.106833.

12 Yusuf M, Atal I, Li J, Smith P, Ravaud P, Fergie M, Callaghan M, Selfe J, Reporting quality of studies using machine learning models for medical diagnosis: a systematic review, *BMJ Open*, **10(3)** (2020)e034568, e034568, doi:10.1136/bmjopen-2019-034568.

13 Kermany D, Zhang K & Goldbaum M, Large dataset of labeled optical coherence tomography (OCT) and chest X-ray images, Mendley data, **172(5)** (2018) 1122–1131, doi:10.17632/rscbjbr9sj.3.

14 Sait U, Gokul L K, Sunny P P, Bhaumik R, Kumar T & Shivakumar S, Curated dataset for COVID-19 posterior-anterior chest radiography images (X-Rays), *Mendeley Data*,V3, (2021), doi:10.17632/9xkhgts2s6.3(16 April 2024).

15 Balanced Augmented Covid CXR Dataset, *Kaggle*, (2022), https://www.kaggle.com/datasets/tr1gg3rtrash/balanced-augmented-covid-cxr-dataset(18 April 2024).

16 Celik G, Detection of Covid-19 and other pneumonia cases from CT and X-ray chest images using deep learning based on feature reuse residual block and depthwise dilated convolutions neural network, *Appl Soft Comput*, **133** (2022) 109906, doi: 10.1016/j.asoc.2022.109906.

17 Sharma P, Nayak D R, Balabantaray B K, Tanveer M & Nayak R, A survey on cancer detection via convolutional neural networks: Current challenges and future directions, *Neural Netw*, **169** (2024) 637–659, doi: 10.1016/j.neunet.2023.11.006.

18 Szepesi P & Szilágyi L, Detection of pneumonia using convolutional neural networks and deep learning, *J Appl Biomed*, **42(3)** (2022) 1012–1022, doi: 10.1016/j.bbe.2022.08.001.

19 Sabour S, Frosst N & Hinton G E, Dynamic routing between capsules, *arXiv*, (2017), doi:10.48550/arxiv.1710.09829.

20 Huang G, Liu Z, Van Der Maaten L & Weinberger K Q, Densely connected convolutional networks, *arXiv*, (2016), doi: 10.48550/arxiv.1608.06993.

21 Tan M & Le Q, Efficient net: Rethinking model scaling for convolutional neural networks, *arXiv*, (2019), doi: 10.48550/arxiv.1905.11946.

22 He K, Zhang X, Ren S & Sun J, Deep residual learning for image recognition, *arXiv*, (2015), doi: 10.48550/arxiv.1512.03385.

23 Jaderberg M, Simonyan K, Zisserman A & Kavukcuoglu K, Spatial transformer networks, *arXiv*, (2015), doi: 10.48550/arxiv.1506.02025.

24 Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X & Unterthiner T, An image is worth 16 × 16 words: Transformers for image recognition at scale, *arXiv*, (2020), doi: 10.48550/arxiv.2010.11929.

25 Liu Z, Lin Y, Cao Y, Hu H, Wei Y & Zhang Z, Swin transformer: Hierarchical vision transformer using shifted windows, *arXiv*, (2021), doi: 10.48550/arxiv.2103.14030.

26 Ma Y & Lv W, Identification of pneumonia in chest X-ray image based on transformer, *Int J Antennas Propag*, **2022** (2022) 1–8, doi: 10.1155/2022/5072666.

27 Jiang J & Lin S, COVID-19 detection in chest x-ray images using swin-transformer and transformer in transformer, *arXiv*, (2021), doi: 10.48550/arxiv.2110.08427.

28 Qu Y, Meng Y, Fan H & Xu R X, Low-cost thermal imaging with machine learning for non-invasive diagnosis and therapeutic monitoring of pneumonia, *Infrared Phys Technol*, **123** (2022) 104201, doi: 10.1016/j.infrared.2022.104201.

29 Kusk M W & Lysdahlgaard S, The effect of gaussian noise on pneumonia detection on chest radiographs using convolutional neural networks, *Radiography*, **29(1)** (2022) 38–43, doi: 10.1016/j.radi.2022.09.011.

30 RSNA pneumonia detection challenge,*Kaggle*, https://www.kaggle.com/competitions/rsna-pneumonia-detection-challenge/data (20 April 2024).

31 Kundu R, Das R, Geem Z W, Han G T & Sarkar R, Pneumonia detection in chest X-ray images using an ensemble of deep learning models, *PLoS One*, **16(9)** (2021), e0256630, doi: 10.1371/journal.pone.0256630.

32 Özdemır Z & Keleş H Y, Covid-19 detection in chest X-ray images with deep learning, in *29th Signal Processing and Communications Applications Conference* (SIU, Istanbul, Turkey) 09–11 June 2021, https://doi.org/10.1109/siu53274.2021.9478028.

33 Bodapati J D & Rohith V N, ChxCapsNet: Deep capsule network with transfer learning for evaluating pneumonia in paediatric chest radiographs, *Measurement*, **188** (2022) 110491, doi: 10.1016/j.measurement.2021.110491.

34 Li D & Li S, An artificial intelligence deep learning platform achieves high diagnostic accuracy for Covid-19 pneumonia by reading chest X-ray images, *iScience*, **25(4)** (2022) 104031, doi: 10.1016/j.isci.2022.104031.

35 Srivastava G, Pradhan N & Saini Y, Ensemble of deep neural networks based on condorcet's jury theorem for screening Covid-19 and pneumonia from radiograph images, *Comput Biol Med*, **149** (2022) 105979, doi: 10.1016/j.compbiomed.2022.105979.

36 Bhandari M, Shahi T B, Siku B & Neupane A, Explanatory classification of CXR images into COVID-19, Pneumonia and Tuberculosis using deep learning and XAI, *Comput Biol Med*, **150** (2022), 106156, doi: 10.1016/j.compbiomed.2022.106156.

37 Chowdhury M E H, Rahman T, Khandakar A, Mazhar R, Kadir M A & Mahbub Z B, Can AI help in screening viral and COVID-19 pneumonia?, *IEEE Access*, **8** (2020) 132665–132676, doi: 10.1109/access.2020.3010287.

38 Diallo P A K K & Ju Y, Accurate detection of COVID-19 using k-efficientnet deep learning image classifier and K-COVID chest X-ray images dataset, in *2020 IEEE 6th*

International Conference on Computer and Communications (ICCC, Chengdu, China) 11–14 December 2020, https://doi.org/10.1109/iccc51575.2020.9344949.

39 Quan H, Xu X, Zheng T, Li Z, Zhao M & Cui X, DenseCapsNet: Detection of COVID-19 from X-ray images using a capsule neural network, *Comput Biol Med*, **133** (2021) 104399, doi: 10.1016/j.compbiomed.2021.104399.

40 Singh S, Kumar M, Kumar A, Verma B K, Abhishek K & Selvarajan S, Efficient pneumonia detection using vision transformers on chest X-rays, *Sci Rep*, **14(1)** (2024) 2487, doi: 10.1038/s41598-024-52703-2.

41 Hao Y, Zhang C & Li X, DBM-ViT: A multiscale features fusion algorithm for health status detection in CXR / CT lungs images, *Biomed Signal Process Control*, **87** (2023) 105365, doi: 10.1016/j.bspc.2023.105365.

42 Bhattacharya M, Jain S & Prasanna P, Radio transformer: A cascaded global-focal transformer for visual attention-guided disease classification, *arXiv*, (2022), doi: 10.48550/arxiv.2202.11781.

43 Shome D, Kar T, Mohanty S, Tiwari P, Muhammad K & AlTameem A, COVID-Transformer: Interpretable COVID-19 detection using vision transformer for healthcare, *Int J Environ Res Public Health*, **18(21)** (2021) 11086, doi: 10.3390/ijerph182111086.

44 Li D, Attention-enhanced architecture for improved pneumonia detection in chest X-ray images, *BMC Med Imaging*, **24(1)** (2024) 6, doi: 10.1186/s12880-023-01177-1.

45 Mabrouk A, Redondo R P D, Dahou A, Elaziz M A & Kayed M, Pneumonia detection on chest X-ray images using ensemble of deep convolutional neural networks, *Appl Sci*, **12(13)** (2022) 6448, doi: 10.3390/app12136448.

46 Wang T, Nie Z, Wang R, Xu Q, Huang H & Xu H, PneuNet: Deep learning for COVID-19 pneumonia diagnosis on chest X-ray image analysis using vision transformer, *Med Biol Eng Comput*, **61(6)** (2023) 395–1408, doi: 10.1007/s11517-022-02746-2.

47 Ukwuoma C C, Qin Z, Heyat M B B, Akhtar F, Bamisile O & Muaad A Y, A hybrid explainable ensemble transformer encoder for pneumonia identification from chest X-ray images, *J Adv Res*, **48** (2022) 191–211, doi: 10.1016/j.jare.2022.08.021.

48 Cadena L, Zotin A, Cadena F & Espinosa N, Noise removal of the x-ray medical image using fast spatial filters and GPU, *Appl Digital Imgage Processing XLI*, **10752** (2018) 568–577, doi: 10.1117/12.2319327.

49 Ioffe S & Szegedy C, Batch normalization: Accelerating deep network training by reducing internal covariate shift, *arXiv*, (2015), doi: 10.48550/arxiv.1502.03167.