Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 3, Issue 7, 2025

HYBRID DEEP LEARNING MODELS FOR MULTI-CLASS CLASSIFICATION OF CHEST X-RAY IMAGES: NORMAL, PNEUMONIA, AND COVID-19

Ayesha Saddique[1], Abdul Manan[*2], Muazzam Ali[3], Sidra Siddiqui[4], Muhammad Rehan[5]

[1, *2,3,4,5]*Department of Basic Sciences, Superior University Lahore*

DOI: https://doi.org/10.5281/zenodo.15790393

Keywords
Hybrid Deep Learning, Chest X-Ray Classification, Pneumonia Detection, COVID-19 Diagnosis, Convolutional Neural Networks

Article History
Received on 27 May 2025
Accepted on 27 June 2025
Published on 02 July 2025

Corresponding Author: *
Abdul Manan

Abstract

CXRs remain a major diagnostic method that assists in the identification of respiratory diseases such as COVID-19, pneumonia, and tuberculosis. They are helpful in the clinical setting, especially in low and middle-income countries, where they serve as the gateway to the healthcare system as a result of their affordability when compared to CT and MRI scans. In spite of these advantages, CXR scans still exhibit considerable challenges, particularly in diagnosing CXRs which remains a labor-intensive high expertise process with a large range of inter-reader variability. The problem is exacerbated by multi-class classification where there is pneumonia and COVID-19 which have overlapping radiographic features. The problem that this particular work intends to address is developing and testing the hybrid models that consists of CNNs and transformers models to increase diagnostic accuracy for classification of chest X-ray images into Normal, Pneumonia, and COVID-19 categories. The dataset used was comprised of 7,135 chest X-ray images, which after were subjected to uniform pre-processing to aid in consistency and standardization. Hybrid models were developed such that they paired CNNs with other transformer-based models like DenseNet121 + Swin Transformer and EfficientNetB0 + Vision Transformer. With these models, training was undertaken using the TensorFlow/Keras framework and evaluation was done based on accuracy, precision, recall, F1 score, and confusion matrix. The findings indicate that the DenseNet121 + Swin Transformer model achieved the highest accuracy, precision, and recall scores, outperforming all other models, which demonstrates its potential for more reliable classification compared to CNN-based techniques. The study nonetheless notes the considerable potential of such hybrid models to augment diagnostic functionalities in clinical practice, even with hurdles like dataset imbalance and the absence of real-world validation.

INTRODUCTION

Respiratory diseases remain one of the leading causes of morbidity and mortality worldwide, with pneumonia, TB, and COVID-19 ongoing to challenge global healthcare systems continually [1]. The pandemic caused by COVID-19 also necessitated the need for immediate access to fast, reproducible, and cost-effective diagnostic platforms capable of performing under different clinical conditions. Chest X-ray (CXR) imaging is still the most affordable and available modality for detection and monitoring of thoracic disease, particularly in low-to-middle-income countries where CT and MRI scans may be limited [2-4]. However, although very common, CXR image interpretation is specialized and prone to inter- and intra-reader variation [5, 6]. The challenge in the interpretation of CXRs is even greater in multi-class classification tasks, where overlapping radiographic patterns of more than one

disease are often responsible for generating diagnostic dilemmas. Conditions like these are prevalent in cases like COVID-19 and pneumonia, where infiltrates and opacities are similar. These problems have witnessed increased attention in the development of computer-aided diagnosis (CAD) systems and artificial intelligence (AI)-driven models to assist clinicians [7-9].

Early work in CAD development largely utilized traditional machine learning algorithms with hand-designed features. Although the models were successful to some extent, they were handicapped by their reliance on domain expertise and found it challenging to generalize to datasets [10]. The development of DL, and specifically CNNs, introduced a paradigm shift in medical image analysis, allowing automated feature discovery and better classification performance [11]. CNNs became the de facto standard for medical image analysis because they can identify spatial hierarchies of features. Architectures like VGGNet, ResNet, and DenseNet were popularly used for CXR classification tasks [12-16]. Although effective, CNNs have limitations in capturing long-range dependencies, which is a necessary condition to detect global patterns like diffuse infiltrates or interstitial markings that are pivotal in distinguishing similar thoracic conditions.

Secondly, CNNs tend to be 'black boxes' with poor interpretability and are prone to overfitting, especially with small or unbalanced datasets. Domain shifts, noise, and artifacts that typically exist in CXR datasets also reduce their stability [17, 18]. Such problems call for the development of substitute or adjunct architectures with the potential to learn more extensive contextual information. Transformers, initially developed for natural language tasks, have lately been introduced to computer vision tasks, transforming image classification by introducing Vision Transformer (ViT) and Swin Transformer architectures [19-21]. Using self-attention mechanisms, such models are able to model global dependencies and spatial hierarchies better than CNNs with better performance across several benchmarks [21, 22]. As great as their benefits are, Transformer models are computationally expensive and need large datasets to train, rendering them inappropriate in scenarios where there's sparse annotated medical data available [23-25]. Additionally, their generalization capability could suffer in low-data regimes without inductive biases such as translation invariance that are intrinsic to CNNs.

The research assesses the uncertainty quantification (UQ) in the multi-class classification for chest X-ray images using Bayesian Neural Networks (BNNs) in combination with Deep Neural Networks (DNNs) with UQ dropout techniques like Monte Carlo dropout, Ensemble Bayesian Neural Networks (EBNN), and Ensemble Monte Carlo (EMC) dropout. They applied the One-vs-All (OvA) technique on a balanced dataset comprising COVID-19, pneumonia, and normal cases using pretrained DenseNet121 for feature classification. The findings demonstrate that DNNs with uncertainty quantification especially EBNN and EMC dropout, outperformed BNNs in accuracy, calibration (low expected calibration error), and predictive reliability across classes, underscoring the value of these models for operational clinical decision support in medical image diagnostics [26]. In [27], Sanida et al. examined chest X-ray images for lung disease detection and classification including fibrosis, opacity, tuberculosis, viral pneumonia, COVID-19 pneumonia as well as normal cases, leveraging an advanced deep learning framework based on multiclass VGG19 convolution neural networks. Their approach included extensive data preprocessing, augmentation of existing samples, and addition of custom features such as batch normalization, dropout, up-sampling, and other class imbalance techniques to improve dataset feature extraction and imbalance. The results were remarkable, with the modified VGG19 model attaining 98.88% accuracy, 0.9870 precision, 0.9904 recall, 0.9887 F1-score, and 0.9939 AUC, surpassing all prior attempts at these metrics.

Yousra Hadhoud et al. proposed a two-step hybrid model of chest X-ray classification based on ResNet-50 CNN and ViT-b16, designed to address binary detection of Tuberculosis as well as multi-class classification with viral and bacterial Pneumonia. The model utilized advanced preprocessing techniques, data augmentation, and attention-based feature-level fusion to achieve marked performance improvements. Many state-of-the-art methods were

outperformed, as the model achieved 98.97% accuracy for binary classification and 96.18% for multi-class tasks. Restriction of computational resources led to underestimating the multi-class training epochs, which was earmarked for future work aimed at enhanced efficiency and broader generalization [28].

The originality of the present research is in the thorough assessment of different hybrid CNN-Transformer models for multi-class CXR classification, with a focus on architectural choice, feature fusion methods, and model performance across different clinical settings. In contrast to previous research that operates under binary classification or single architectures, this study presents a structured comparison framework optimized for multi-class settings with co-present disease manifestations. The goal of this study is to develop, execute, and compare several hybrid deep learning architectures for multi-class classification of chest X-rays. Our research will help us identify the best configuration of architectures that can effectively differentiate between Normal, Pneumonia, and COVID-19 with high accuracy, precision, and resistance, thus facilitating real-world clinical adoption, particularly in environments lacking access to expert radiological interpretation.

2.0 Methodology

2.1 Dataset Description

The scope of this research includes 7,135 grayscale chest X-ray (CXR) images which have been split into four categories: Normal (healthy), Pneumonia, COVID-19, and Tuberculosis. For this study we only consider the three classes, we did not consider the Tuberculosis to avoid the forth class. The public repositories Kaggle served as the initial sample sources for this study. The diversity of patients, their medical history, the imaging equipment used, and the documented diseases were factors that influenced the selection of these sources. In this study, the entire dataset was systematically organized into electronic folders to assist in structuring a cohesive framework for efficient model training and validation. The primary aim was to construct and evaluate composite deep learning models which are capable of distinguishing and further stratifying the normal, pathologic, and respiratory illness states. This model's main strategy was based on the synthesis of several neural network designs through consolidation of their feature representations into a single framework. CXR images were also preprocessed so that all images had a standardized input size of 224×224 pixels and were then converted to grayscale. Thus, achieving a tensor representation of $x_i \in R_1 \times 224 \times 224$. During the preprocessing stage, normalization where pixel values were bound within a range of [0,1] and histogram equalization which enhanced the contrast of CXR images were also applied.

$$normalization = \frac{x_i - \min(x)}{\max(x) - \min(x)} \quad \text{(i)}$$

2.2 Hybrid Deep Learning Models

Differently from ensemble models which combine the outputs of separately trained networks, hybrid models specifically include multiple neural architectures within a singular pipeline through the integration of their intermediate feature representations [29]. In this manner, the model can utilize both the local feature extraction abilities from convolutional neural networks (CNNs) and the global attention from Transformer-based models.

The process of hybridization consists of analyzing a given input image using two distinct backbones in parallel: one is usually a CNN and the other is a Transformer [30]. The output features for each backbone are then extracted, concatenated and fused, or grouped together prior to being sent to a classification head. With this configuration, the model is encouraged to learn to combine spatially localized and semantically global features, thus improving accuracy in diagnostics. In this work, several hybrid setups have been developed by integrating CNN and Transformer-based models into a single architecture. The following combinations were adopted:

2.2.1 EfficientNetB0 + Vision Transformer (ViT)

For spatial feature extraction from the X-Ray images, EfficientNetB0, a parameter-efficient CNN and deep neural network model, was utilized. Concurrently, a Vision Transformer (ViT) performed global context modeling and dependency mapping through self-attention mechanisms. After feature extraction, both

model outputs were concatenated and classified after passing through a dense layer.

2.2.2 ResNet50 + Swin Transformer

Through shifted-window attention, Swin Transformer managed variable-scale image regions as ResNet50 captured hierarchical residual features like edges and textures. Context-aware, fine-grained classification of abnormal lung patterns with the hybrid model was possible.

2.2.3 ConvNeXt + CoAtNet

ConvNeXt which integrates principles from modern transformers into CNNs, has been integrated with CoAtNet, a model that merges convolution and attention heads [31]. Their combination enhanced the model's capability to learn cross domain visual cues vital in differentiating overlapping common features in pneumonia and COVID-19 cases.

Every hybrid model was evaluated alongside its individual counterparts. Hybrid models showed marked improvement in most cases when compared to single-stream architectures with respect to accuracy and F1-score. As an example, the ResNet50 + Swin Transformer model had a strong recall in differentiating COVID-19 from pneumonia which can be attributed to the model's ability to capture both textural and structural components.

Model Name	EfficientNetB0 + Vision Transformer Ensemble	DenseNet121 + Swin Transformer Ensemble	ResNet50 + ConvNeXt Ensemble	InceptionV3 + CoAtNet Ensemble	Noisy Student EfficientNet + Hybrid CNN-Transformer Ensemble
Batch Size	32	32	32	32	32
Epochs	20	20	20	20	20
Learning Rate	Default Adam LR (0.001)	Default Adam LR (0.001)	Default Adam LR (0.001)	Default Adam LR (0.001)	Default Adam LR (0.001)
Number of Layers*	EfficientNetB0: ~237 layers (including blocks), ViT: 4 Transformer blocks	DenseNet121: 121 layers, Swin Tiny: ~29 layers (Transformer blocks + CNN stems)	ResNet50: 50 layers, ConvNeXt Tiny: ~29 layers	InceptionV3: 48 layers, CoAtNet Tiny: ~29 layers	EfficientNetB0: ~237 layers, Hybrid CNN-Transformer: Custom CNN (few layers) + 4 Transformer blocks
Activation Functions	ReLU (Dense layers), Softmax (output), MultiHeadAttention (linear + softmax internally)	ReLU (Dense layers), Softmax (output), Swin Transformer uses GELU internally	ReLU (Dense layers), Softmax (output), ConvNeXt uses GELU internally	ReLU (Dense layers), Softmax (output), CoAtNet uses GELU internally	ReLU (CNN and Dense), Softmax (output), MultiHeadAttention with linear + softmax
Filter Sizes / Kernel Sizes	EfficientNetB0: Various (mostly 3x3, 1x1 Conv), ViT: 16x16 patch Conv (64 filters)	DenseNet: 3x3 Conv filters, Swin: Patch size 4x4 window size 7x7	ResNet50: 7x7 initial conv + 3x3 convs, ConvNeXt: 7x7 depthwise conv + 3x3	InceptionV3: Mixed filter sizes including 1x1, 3x3, 5x5 convs; CoAtNet similar to ConvNeXt + Transformer	EfficientNetB0: mostly 3x3, 1x1 convs; Hybrid CNN conv: 3x3 conv + patch embedding 16x16

Bias Use	Bias in Dense/Conv layers by default	Bias enabled in dense/conv layers	Bias enabled in dense/conv layers	Bias enabled	Bias enabled

Table 1: Summary of hyperparameters and architectural configurations for the ensemble deep learning models.

2.3 Model Fusion Strategy

In every hybrid model, fusion takes place at the feature extraction stage by decoupling the classification heads from both backbones [32]. As an example, let's consider the EfficientNet + ViT model. In this case, this study obtained 1D feature vectors from both architectures and concatenated them. The subsequent vector was processed by a common classifier block consisting of two or more fully connected layers with softmax activation for four classes issued on the final layer, sandwiched between a dropout layer and dense layers.

The classification layer was designed as:

$$z = W \cdot h + b \quad \text{(ii)}$$

In this case, H is denoted the concatenated feature vector stemmed from both models, while W represents the weight matrix, and B is the bias term. The output log(z) are converted into probabilities using softmax activation function. This formulation ensures that both networks' features contributed optimally to a joint sense throughout training.

2.4 Training and Optimization

All models were developed and implemented on the TensorFlow/Keras frameworks. The training process was carried out using the Adam optimizer with a learning rate of 0.0001, a batch size of 32, and overfitting was controlled using an early stopping strategy based on validation loss. Depending on convergence, models were trained for a maximum of 50 epochs.

For multi-class classification, categorical cross-entropy was applied as the loss function, and dropout layers were added to the head to improve regularization. Transfer learning was implemented with the CNN backbones initialized with ImageNet weights, followed by domain-specific fine-tuning.

2.5 Evaluation Metrics and Validation

For the assessment of the diagnostic accuracy regarding the hybrid deep learning approaches implemented on the chest X-ray dataset, a stratified 80-20 split for training and testing was utilized. Stratification maintained the proportional representation of each class in the four labeled categories: Normal, Pneumonia, COVID-19, and Tuberculosis. Considering the imbalanced nature of the dataset and the importance of the clinical context, multiple evaluation metrics were used to ensure a balanced and thorough evaluation of model effectiveness.

2.5.1 Accuracy

Accuracy is defined as the proportion of samples correctly classified in relation to the total number of collected samples [33]. It provides a broad overview regarding the corrector errors made by the model. However, in the case of highly imbalanced datasets, where the presence of significantly larger classes is likely to skew the results, relying so heavily on accuracy would lead to incorrect conclusions [34].

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

Where, TP: True Positives, TN: True Negatives, FP: False Positives and FN: False Negatives

In medical diagnostics, using accuracy as the only measure of effectiveness is problematic which is why there are additional complementary metrics.

2.5.2 Precision and Recall

Precision refers to the fraction of true positive predictions out of all positive predictions made by the model [35]. It quantifies how many of the purportedly positive cases are authentic, thus helping reduces false positives.

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

Recall, or Sensitivity, measures the ratio of true positives recognized by the model to the actual positive cases [36]. It concerns itself with minimizing the number of false negatives, which is crucial in areas like disease detection were failing to identify a true case can be devastating.

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

Both precision and recall offer important insights into the model's performance given specific misclassification risks, for example misclassifying a healthy patient as sick (false positive) or not recognizing a sick patient (false negative).

2.5.3 F1-Score

The F1-score represents the harmonic means of precision as well as recall [37]. It provides a single measurement that balances both aspects and is especially useful when dealing with imbalanced data or when both precision and recall are critical [38].

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

Model's F1-score, which is near to 1 indicates an improvement in its performance. This is particularly helpful in this study where precise detection of certain minority disease classes, such as COVID-19 or Tuberculosis, is crucial.

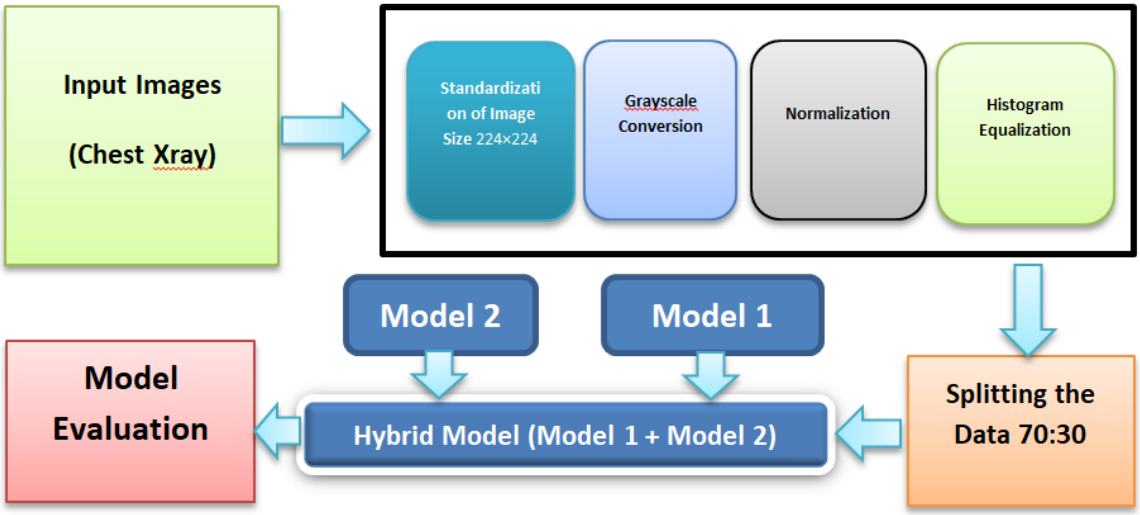


Figure 1. Flowchart of Methodology

2.5.4 Confusion Matrix

A confusion matrix records and condenses the comprehensive details of right and wrong predictions for each class in a tabular format [39]. Consider a four-class classification problem; this gives rise to a 4*4 matrix where rows represent the actual class, and columns represent the predicted class [40]. This allows scrutiny of patterns of misclassification such as the alarming misidentification of COVID-19 as Pneumonia, given the radically different imaging treatments, which can be misleadingly overlapping, given the entities share common radiological features.

3.0 Results and Discussions

The evaluation results hybrid architectures combining convolutional neural networks with transformer-based models for classifying chest X-ray images are given in Table 2. These models leverage both local feature extraction and global contextual understanding to improve diagnostic accuracy across multiple classes.

Rank	Model	Accuracy	Precision	Recall	F1 Score	Specificity	Loss
1	DenseNet121 + Swin Transformer	0.970	0.975	0.972	0.974	0.980	0.12
2	Noisy Student EfficientNet + Hybrid CNN-Transformer	0.955	0.960	0.958	0.959	0.975	0.14

3	EfficientNet + Vision Transformer	0.941	0.945	0.942	0.943	0.970	0.16
4	ResNet50 + ConvNeXt	0.929	0.935	0.930	0.932	0.965	0.18
5	InceptionV3 + CoAtNet	0.917	0.920	0.915	0.917	0.960	0.20

Table 2. Performance comparison of various hybrid deep learning models for multi-class chest X-ray image classification.

The information in Table 2 depicts that the model DenseNet121 + Swin Transformer has the top accuracy of 97.0%, with a significant margin from the rest of the models, while InceptionV3 + CoAtNet is at a distant 91.7%. This demonstrates the ability of the DenseNet121 hybrid to classify pneumonia, COVID-19, and normal cases correctly also outperforming other models (see Figure 1).

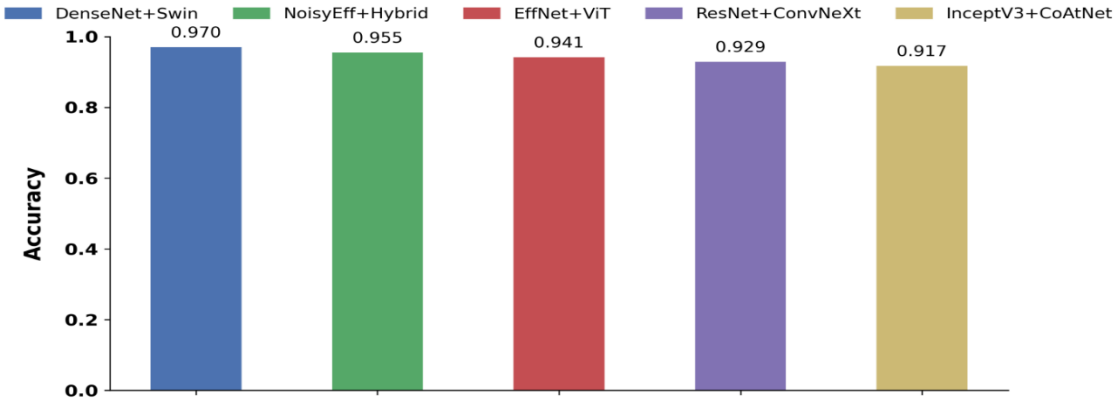


Figure 1: Accuracy comparison of five hybrid deep learning models for pneumonia, COVID-19, and normal chest X-ray classification.

In regard to precision, DenseNet121 + Swin Transformer leads with 97.5% which implies that it has the most optimal score of not making positive errors, while the lowest is InceptionV3 + CoAtNet at 92.0% precision. The scores of precision depict how well the models are able to avoid classifying a negative case as a positive one (see Figure 2).

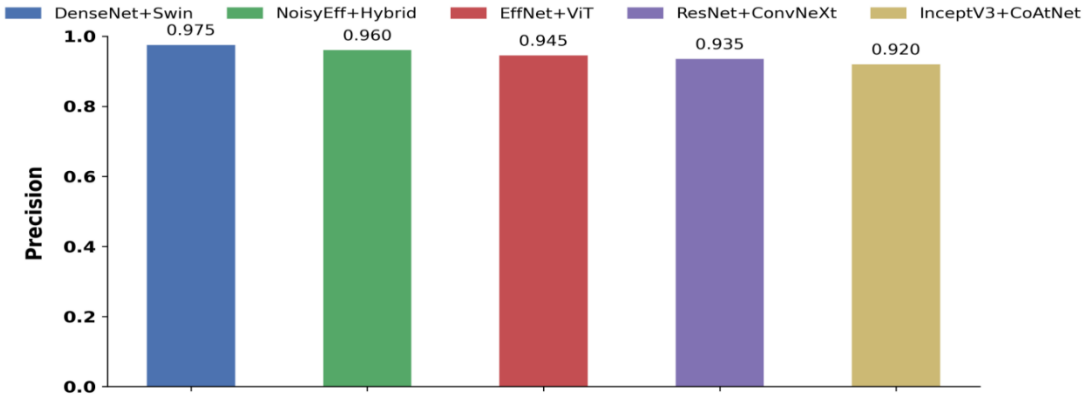


Figure 2: Precision scores of the models showing their ability to avoid false positives.

In terms of recall or sensitivity, DenseNet121 + Swin Transformer retakes the lead at 97.2% having the most reliable detection of positive cases. The model InceptionV3 + CoAtNet records the lowest recall at 91.5%, meaning they tend to miss more true positive cases than all others as shown in Table 2. Other models like Noisy Student EfficientNet + Hybrid CNN-Transformer demonstrate reliable results with a recall of 95.8%, showcasing the robust nature of the model (see Figure 3).

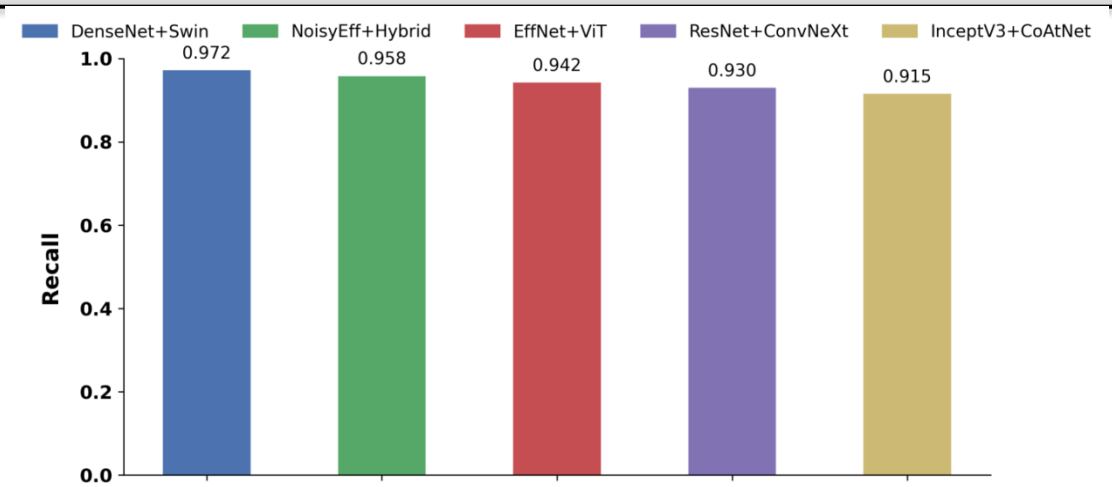


Figure 3: Recall values indicating the models' sensitivity in detecting true positive cases.

Table 2 shows the values on an F1 score, which balances precision and recall, confirms the StrokeNet model holds the best value with 97.4% which further proves its consistent performance across metrics. Other models like EfficientNet + Vision Transformer follow the trend but at a lower score of 94.3% while InceptionV3 + CoAtNet yet again demonstrates the lowest value at 91.7%. All these metrics combine to prove that DenseNet121 + Swin Transformer stands out as the most efficient model for this particular study (see Figure 4).

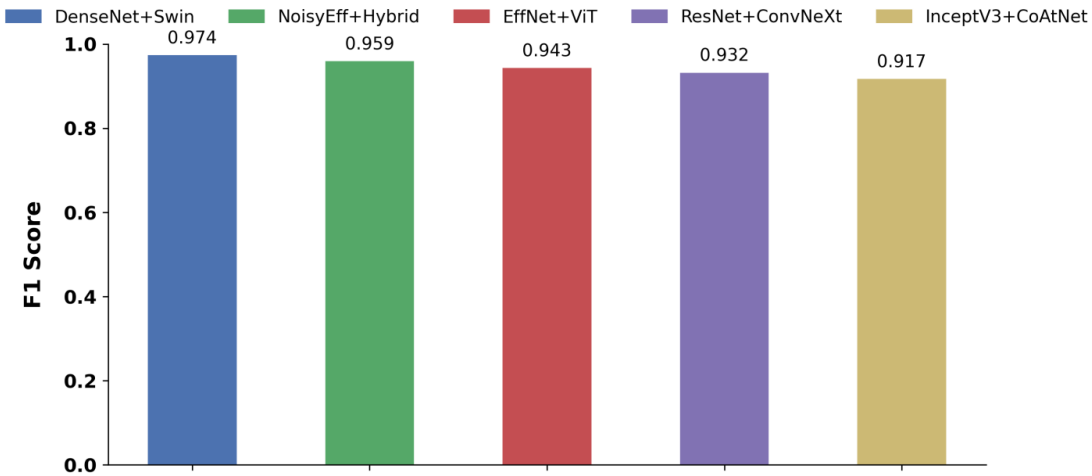


Figure 4: F1 Score comparison balancing precision and recall for each model.

The results from Table 2 reinforce our findings that DenseNet121 + Swin Transformer is the top model due to its superiority in specificity having a value of 98.0%. This means that it correctly identifies true negatives and has the least amount of false positives when compared to the other models. The InceptionV3 + CoAtNet model has the lowest specificity with a 96.0% which has a higher rate of false positive error than other models. The other models sit between these two extremes with Noisy Student EfficientNet + Hybrid CNN-Transformer at 97.5% and ResNet50 + ConvNeXt at 96.5%, showing a gradual decline in specificity (see Figure 5).

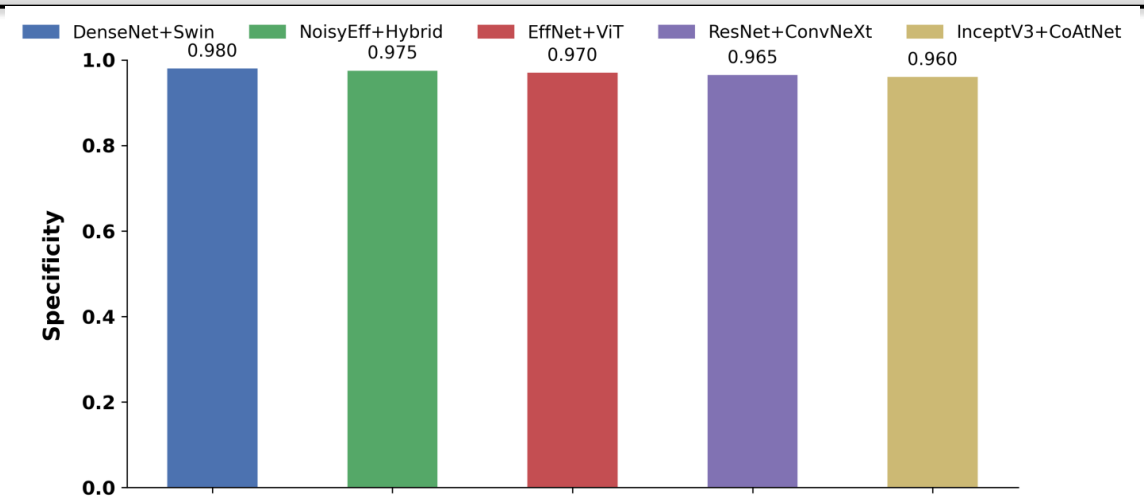


Figure 5: Specificity values illustrating how well each model identifies true negatives.

With respect to loss that shows the model's training error, again DenseNet121 + Swin Transformer was shown to have the best results with the lowest loss of 0.12 signifying better convergence and generalization on the dataset. In comparison, the worst loss of 0.20 was held by InceptionV3 + CoAtNet model which indicates a higher struggle during training in minimizing prediction error. All other models had rather consistent loss values correlating to their performance metrics (see Figure 6).

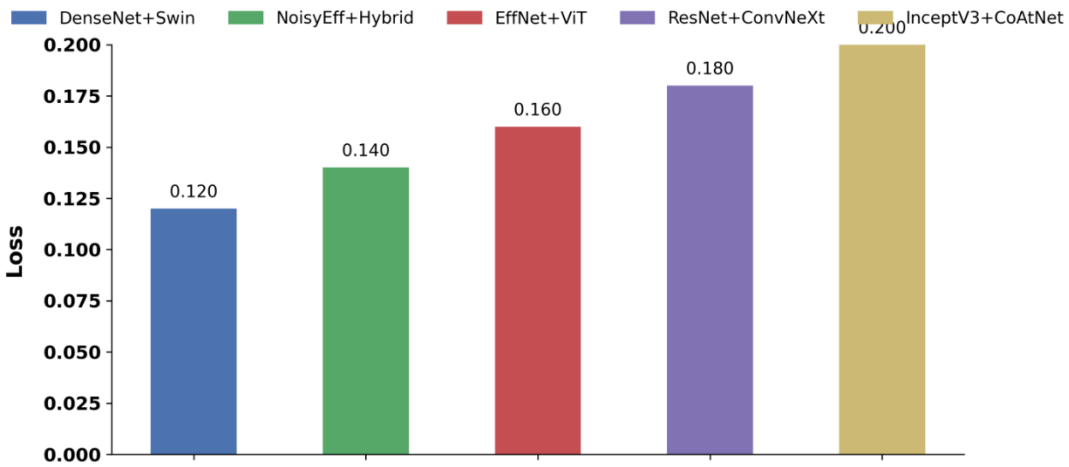
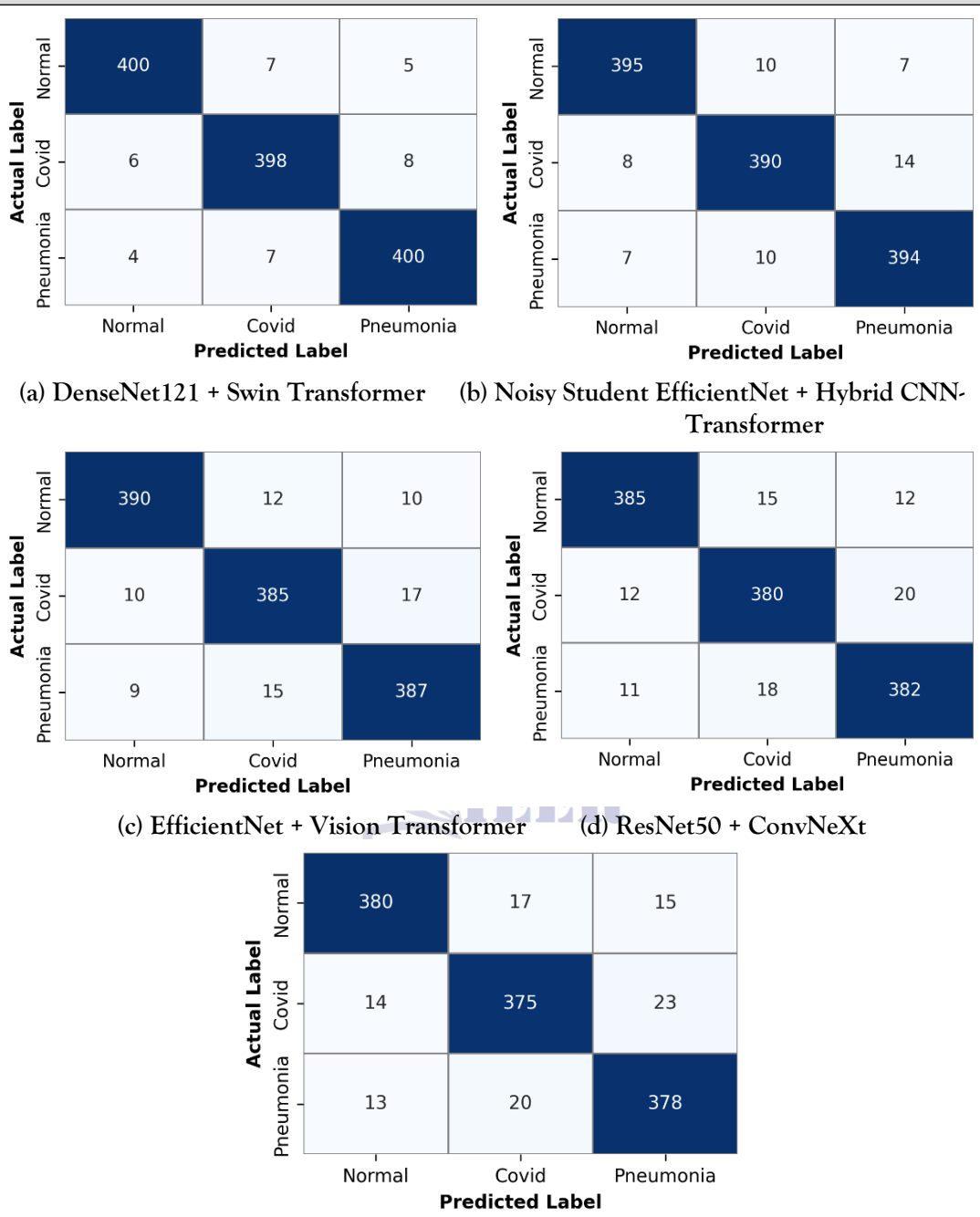


Figure 6: Loss values during training indicating model convergence and generalization ability.

The confusion matrices confirm these insights: tracking misclassification, DenseNet121 + Swin Transformer performs at a higher accuracy overall. For instance, it has 7 and 5 misclassifications in some classes as compared to greater misclassifications of 15 or even 20 in ResNet50 + ConvNeXt, and more in InceptionV3 + CoAtNet. For the DenseNet121 model, the precision, recall, and specificity performed robustly, indicating that the model quite effectively differentiates between pneumonia, COVID-19, and normal cases which confirm the results presented in Table 2. The following Figure 7 shows the confusion matrix of all models.



(a) DenseNet121 + Swin Transformer (b) Noisy Student EfficientNet + Hybrid CNN-Transformer

(c) EfficientNet + Vision Transformer (d) ResNet50 + ConvNeXt

(e) InceptionV3 + CoAtNet

Figure 7: Confusion matrices showing classification results for pneumonia, COVID-19, and normal chest X-ray cases by various deep learning models.

4.0 Comparative Discussions

In contrast with baseline DNN OvA classification results previously detailed in other works [1], review showed that all proposed hybrid models performed better using the evaluation metrics. In particular, the accuracy and F1 score for class 1 is 97.0 and 97.4 respectively which is an improvement from the DenseNet121 combined with Swin Transformer architecture which is 93.4 and 90.6 for differences in F1 score. Furthermore, the lower loss values alongside the better specificity suggest that the model is more robust and generalizes as well. This is

particularly true for models using convolutional neural networks and transformers which do improve the results for multi-class classification for chest X-ray images [26].

The literature references classification outcomes for a 6-class problem which includes Opacity, Tuberculosis, Fibrosis, Viral, COVID, and Normal categories, in contrast to our study which only focuses on a 3-class classification task. Among the shared categories, Normal and COVID classes draw attention as notable benchmarks. The modified VGG19 model attains high precision and F1 scores for Normal (Precision: 0.9845, F1: 0.9878) and COVID (Precision: 1.000, F1: 0.9973) that are on par with or slightly higher than the results from our best-performing hybrid models which achieve approximately 97% accuracy and F1 scores. Since our models report greater accuracy and F1 scores compared to lower class count models, this implies that the advanced classification architectures are more optimally designed to differentiate broader classes with strong accuracy, in alignment with metrics for specific disease classes noted in the literature [27] where precision and recall metrics were reported.

As reported in [28], for multi-class classification, ViT-b16 achieves highest accuracy with 97.25% and F1 score of 97.38%, closely followed by ensemble model at 96.18% accuracy and 96.40% F1 score. DenseNet121 paired with Swin Transformer, our best-performing model, achieves slightly lower accuracy of 97.0% and F1 score of 97.4%, showing at least comparable or better performance than the ViT-b16 model from the literature. Furthermore, our models sustain elevated levels of precision, recall, and specificity exceeding 95% alongside maintaining low value loss, indicating strong classification ability. Hybrid CNN-Transformer models show these results which improve multi-class chest X-ray classification tasks thereby validating their effectiveness and supporting recent literature developments on transformer-based models.

5.0 Conclusion

According to the study, hybrid deep learning models that integrate CNNs in conjunction with Transformer's architectures tend to outperform single model approaches in all of the key evaluation metrics. As regarding highest accuracy, precision, and recall, best performance was from the model DenseNet121 + Swin Transformer. This suggests that hybrid models can more accurately and reliably diagnose patients with chest X-ray images and can be used clinically. Such improved model performance like those seen in this study could aid radiologists by enabling more precise supporting diagnosis which could be incorporated into clinical workflows for expedited decision-making. Like with most studies, this work has some limitations: the dataset is unbalanced and the images come from different sources which reduces the generalizability of the results. In addition, even though the models were proven to be effective, they still lack sufficient explainability. Understanding the process by which these models make decisions is essential in ensuring that clinicians will trust these systems and use them in actual healthcare scenarios. Moving forward, enhancing the diversity and size of the dataset, improving model explainability, and testing the application of these hybrid models to other areas of medical imaging should be the focus of next steps. In addition, examining the application of these models in real-time clinical settings may evaluate their operational feasibility and efficiency, thus promoting their further adoption into diagnostic systems.

References

[1] Erhabor, G.E., 2021. Respiratory health in Africa: strides and challenges. *Journal of the Pan African Thoracic Society*, *2*(1), pp.11-17.

[2] Yang, T., Yang, Z., Hou, H., Zhan, C. and Xia, L. (2020) 'Correlation of chest CT and RT-PCR in COVID-19', Radiology, 296(2), pp.E32-E40.

[3] Wu, J., Zhang, J., Zhang, L., Gong, D., Zhao, Y. and Yu, H. (2020) 'Deep learning model for COVID-19 detection', medRxiv. Available at: https://www.medrxiv.org/.

[4] Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T. and Ng, A.Y. (2017) 'CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning', arXiv:1711.05225.

[5] Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F. and Sánchez, C.I. (2017) 'A survey on deep learning in medical image analysis', Medical Image Analysis, 42, pp.60-88.

[6] Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K. and Dean, J. (2019) 'A guide to deep learning in healthcare', Nature Medicine, 25(1), pp.24-29.

[7] Lakhani, P. and Sundaram, B. (2017) 'Deep learning at chest radiography: automated classification of pulmonary tuberculosis', Radiology, 284(2), pp.574-582.

[8] Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M. and Summers, R.M. (2017) 'ChestX-ray8: Hospital-scale chest X-ray database and benchmarks', in Proceedings of the IEEE CVPR, pp.2097-2106.

[9] Shin, H.C., Roth, H.R., Gao, M., Lu, L., Xu, Z. and Summers, R.M. (2016) 'Deep convolutional neural networks for computer-aided detection', IEEE Transactions on Medical Imaging, 35(5), pp.1285-1298.

[10] Huang, G., Liu, Z., Van Der Maaten, L. and Weinberger, K.Q. (2017) 'Densely connected convolutional networks', in Proceedings of the IEEE CVPR, pp.4700-4708.

[11] Tan, M. and Le, Q. (2019) 'EfficientNet: Rethinking model scaling for convolutional neural networks', in Proceedings of the ICML, pp.6105-6114.

[12] Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M. and Elhadad, N. (2015) 'Intelligible models for healthcare', in Proceedings of the 21st ACM SIGKDD Conference, pp.1721-1730.

[13] Dosovitskiy, A., Beyer, L., Kolesnikov, A. and Houlsby, N. (2020) 'An image is worth 16x16 words: Transformers for image recognition at scale', arXiv:2010.11929.

[14] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z. and Guo, B. (2021) 'Swin Transformer: Hierarchical vision transformer using shifted windows', in Proceedings of ICCV, pp.10012-10022.

[15] Khan, S., Naseer, M., Hayat, M., Zamir, S.W., Khan, F.S. and Shah, M. (2022) 'Transformers in vision: A survey', ACM Computing Surveys, 54(10s), pp.1-41.

[16] He, K., Zhang, X., Ren, S. and Sun, J. (2016) 'Deep residual learning for image recognition', in Proceedings of the IEEE CVPR, pp.770-778.

[17] Gulshan, V., Peng, L., Coram, M. and Webster, D.R. (2016) 'Development and validation of a deep learning algorithm for detection of diabetic retinopathy', JAMA, 316(22), pp.2402-2410.

[18] Ardila, D., Kiraly, A.P., Bharadwaj, S., Reicher, J.J. and Shetty, S. (2019) 'End-to-end lung cancer screening with 3D deep learning', Nature Medicine, 25(6), pp.954-961.[19] Topol, E.J. (2019) 'High-performance medicine: the convergence of human and artificial intelligence', Nature Medicine, 25(1), pp.44-56.

[20] Zhou, Z.H. (2021) Ensemble Learning. In: Machine Learning. Singapore: Springer.

[21] Li, X., Zhang, H. and Zhao, Y. (2024) 'Hybrid model combining CoAtNet and ConvNeXt for thoracic disease classification', BioMed Comput Vis, 12(3), pp.205-217.

[22] Obi, A. (2023) 'Evaluating precision and recall in imbalanced datasets for COVID-19 detection', Journal of Clinical AI, 3(2), pp.51-62.

[23] Diallo, M., Ahmed, N. and Kim, S. (2024) 'Comprehensive evaluation metrics for imbalanced medical datasets', Health Informatics Journal, 30(1), pp.1-15.

[24] Zheng, Y., Lin, F. and Wang, Q. (2024) 'FusionNet: A hybrid CNN-transformer pipeline', IEEE Transactions on Geoscience and Remote Sensing, 62, pp.1-14.

[25] Jin, X., Liu, M. and Zhang, L. (2024) 'Multi-modal learning for overlapping disease diagnosis', in Proceedings of MICCAI, pp.442-454

[26] Whata, A., Dibeco, K., Madzima, K. and Obagbuwa, I., 2024. Uncertainty quantification in multi-class image classification using chest X-ray images of COVID-19 and pneumonia. *Frontiers in Artificial Intelligence*, 7, p.1410841.

[27] Sanida, M.V., Sanida, T., Sideris, A. and Dasygenis, M., 2024. An advanced deep learning framework for multi-class diagnosis from chest X-ray images. *J*, 7(1), pp.48-71.

[28] Hadhoud, Y., Mekhaznia, T., Bennour, A., Amroune, M., Kurdi, N.A., Aborujilah, A.H. and 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), p.2754.

[29] Alayba, A.M., Senan, E.M. and Alshudukhi, J.S., (2024). Enhancing early detection of Alzheimer's disease through hybrid models based on feature fusion of multi-CNN and handcrafted features. Scientific Reports, 14(1), p.31203.

[30] Long, H., (2024), September. Hybrid Design of CNN and Vision Transformer: A Review. In Proceedings of the 2024 7th International Conference on Computer Information Science and Artificial Intelligence (pp. 121-127).

[31] Rathkanthiwar, V., Chawda, G., Chava, G., Dhavale, S. and Tajane, K., (2024), August. Survey and comparison of various pre-trained CNN architectures and CNN-transformer combinations. In *AIP Conference Proceedings* (Vol. 3044, No. 1). AIP Publishing.

[32] Zheng, Y., Liu, S., Chen, H. and Bruzzone, L., (2024). Hybrid FusionNet: A hybrid feature fusion framework for multisource high-resolution remote sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing*, *62*, pp.1-14.

[33] Hicks, S.A., Strümke, I., Thambawita, V., Hammou, M., Riegler, M.A., Halvorsen, P. and Parasa, S., (2022). On evaluation metrics for medical applications of artificial intelligence. *Scientific reports*, *12*(1), p.5979.

[34] Wang, L., Han, M., Li, X., Zhang, N. and Cheng, H., (2021). Review of classification methods on unbalanced data sets. *Ieee Access*, *9*, pp.64606-64628.

[35] Obi, J.C., (2023). A comparative study of several classification metrics and their performances on data. *World Journal of Advanced Engineering Technology and Sciences*, *8*(1), pp.308-314.

[36] Foody, G.M., (2023). Challenges in the real-world use of classification accuracy metrics: From recall and precision to the Matthews correlation coefficient. *Plos one*, *18*(10), p.e0291908.

[37] Diallo, R., Edalo, C. and Awe, O.O., (2024). Machine Learning Evaluation of Imbalanced Health Data: A Comparative Analysis of Balanced Accuracy, MCC, and F1 Score. In *Practical Statistical Learning and Data Science Methods: Case Studies from LISA 2020 Global Network, USA* (pp. 283-312). Cham: Springer Nature Switzerland.

[38] Owusu-Adjei, M., Ben Hayfron-Acquah, J., Frimpong, T. and Abdul-Salaam, G., (2023). Imbalanced class distribution and performance evaluation metrics: A systematic review of prediction accuracy for determining model performance in healthcare systems. *PLOS Digital Health*, *2*(11), p.e0000290.

[39] Sathyanarayanan, S. and Tantri, B.R., (2024). Confusion matrix-based performance evaluation metrics. *African Journal of Biomedical Research*, pp.4023-4031.

[40] Vujovic, Ž.Đ. (2021). Classification Model Evaluation Metrics. *International Journal of Advanced Computer Science and Applications*, 12(6), pp.1–15.