Binary Classification of Tuberculosis CXR Images Across Diverse Range of CNN Architectures: A Comparative Study

Syeda Meraj
Department of Information Systems
Kulliyyah of Information and communications Technology
Selangor, Malaysia
syedashaizadi@gmail.com

Asadullah Shah
Department of Information Systems
Kulliyyah of Information and communications Technology
Selangor, Malaysia
asadullah@iium.edu.my

Ahsiah Ismail
Department of Computer Science
Kulliyyah of Information and communications Technology
Selangor, Malaysia
ahsiah@iium.edu.my

Tengku MT Sembok
Department of Computer Science
Kulliyyah of Information and communications Technology
Selangor, Malaysia
tmts@iium.edu.my

Syed Shadab
Department of Information Technology
Al Musanna College of Technology
Oman
sgmshadab@gmail.com

Syed Aftab
Department of Aeronautical Engineering Technology
Higher Colleges of Technology
Al-Ain, UAE
saftab@hct.ac.ae

2024 IEEE 9th International Conference on Engineering Technologies and Applied Sciences (ICETAS) | 979-8-3503-6314-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICETAS62372.2024.11119808

***Abstract*—This paper investigates the performance of widely used pre-trained CNN architectures (VGG16, MobileNetV3, DenseNet121, and RegNet040) across diverse datasets, particularly focusing on tuberculosis (TB) detection using Chest X-Rays (CXRs). Deep learning (DL) techniques applied to CXRs aid radiologists in promptly and accurately identifying TB, which is especially critical in low-income regions with constrained diagnostic resources. The research reveals that MobileNetV3 consistently demonstrates superior performance compared to other architectures.**

Keywords—Artificial Intelligence (AI), Convolutional Neural Networks (CNNs), Deep Learning (DL), Machine Learning (ML), Tuberculosis (TB), Pre-trained models.

I. Introduction

Tuberculosis (TB) remains a significant global health threat, particularly in low-income countries with limited resources for diagnosis. Conventional methods like sputum and culture tests take 5-8 weeks, leading to delayed treatment and increased mortality [1]. A report by the World Health Organization (WHO) estimated over 10 million TB cases were diagnosed in 2022 [2]. In low resource areas, there's a scarcity of both equipment and skilled radiologists which results in delayed diagnoses of many TB cases. Often radiologists are overworked, understaffed, under tremendous pressure to accurately interpret medical images such as chest X-rays, which are vital for TB screening.[3].

This is where Deep Learning (DL) techniques can play a transformative role. By developing DL systems trained on vast datasets of Chest X-Rays (CXRs), we can create a helpful assistant for radiologists. Similar to how radiologists learn through experience, this DL would identify patterns indicative of TB. Functioning as a second opinion, it would not replace radiologists but rather offer valuable insights and support in confirming diagnoses. This aligns with WHO recommendations for using chest X-rays in TB screening and contributes to the UN's Sustainable Development Goal 3 of ending the TB epidemic by 2030. By employing AI as a friendly assistant, radiologists can become more efficient and accurate in their fight against TB.

This paper focuses on using a diverse range of CNN architectures and leverages the smaller SH and MC datasets for research. Le, Nguyen-Tat, & Ngo (2022) [4], utilized various pre-trained architectures, making their study suitable for comparison. We aim to conduct a binary classification of TB across various range of CNN models.

II. Methodology

The TB CXR Images are obtained from U.S. National Library of Medicine provides two datasets of CXRs from Montgomery County, Maryland, USA, and Shenzhen No. 3 People's Hospital, China, for TB diagnosis, including normal and abnormal X-rays with radiologist readings [5]. Table 1 summarizes these datasets.

TABLE I. Description of the Datasets

Datasets Name	Normal CXRs	TB CXRs	Total CXRs	Image Type
Montgomery (MC Dataset)	80	58	138	PNG
Shenzhen (SH Dataset)	326	336	662	PNG
Both (MC + SH) Datasets	406	394	800	PNG

Fig.1. illustrates the end-to-end pipeline for efficiently processing the CXR images through pre-processing, feature extraction, and classification.

979-8-3503-6314-2 /24/$31.00 ©2024 IEEE

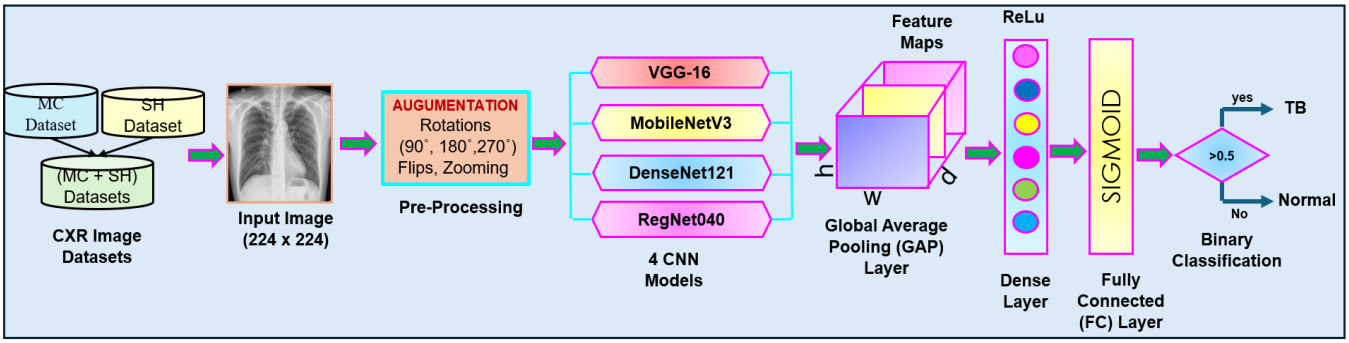

Fig. 1. Illustrates the end-to-end chest x-ray classification methodology with four pre-trained CNN model architectures.

A. Data Preprocessing

CXRs are normalize and resize to 224x224 input image size. Next, data augmentation techniques such as rotations, flips and zooming are applied to increase the dataset size. A setup data generators is created to efficiently preprocess batches of images during training.

B. Dataset Partition

Split data into training, validation, and test sets, allocating 70% for training, 15% for validation, and 15% for testing. Address class imbalance in the training set by oversampling the minority class.

C. Model Preparation

Modify models for binary classification by adding dense layers with sigmoid activation. Compile models using 'adam' optimizer, 'binary_crossentropy' loss function, and 'accuracy' metric. Define callbacks like ModelCheckpoint and EarlyStopping to save the best model and enable early stopping.

Diverse range of models have been considered similar to this study because of easy comparision purposes. A few details about the models utilized are discussed below.

VGG16: VGG16 is a renowned convolutional neural network (CNN) architecture known for its simplicity and effectiveness in image classification. With 16 convolutional layers and fully connected layers, it offers a straightforward structure and strong performance on diverse datasets. However, its large parameter count leads to higher computational costs compared to newer architectures. [6].

MobileNetV3: MobileNetV3 is a lightweight convolutional neural network architecture optimized for mobile and embedded devices. It incorporates efficient building blocks like inverted residuals and linear bottlenecks to achieve this balance. Advantages of MobileNetV3 include its small size, fast inference speed, and suitability for deployment on resource-constrained devices [7]. However, its main limitation might be slightly lower accuracy compared to larger and more complex models like VGG16 or DenseNet121.

DenseNet121: DenseNet121 is a densely connected CNN architecture recognized for its dense connectivity patterns, promoting feature reuse and gradient flow enhancement. It excels in image classification with efficient parameter utilization and resilience to overfitting [8]. Its dense connections contribute to increased memory usage, potentially limiting its deployment on memory-constrained devices compared to lighter models like MobileNetV3.

RegNet040: RegNet040 is a member of the RegNet family of neural network architectures designed for efficient and scalable performance in computer vision tasks. RegNet models emphasize simplicity, regularity, and scalability, making them easy to train and deploy across different hardware platforms. Advantages of RegNet040 include its simplicity, efficiency, and scalability, allowing for fast training and inference on various tasks and datasets [9]. However, its main limitation may be its performance compared to more complex architectures like VGG16 or EfficientNetB7, especially on datasets with more complex patterns and structures. In this stage each model extracts different feature from the input CRXs.

D. Global Average Pooling Layer

Global Avergare Pooling (GAP) layer reduces feature maps into a single value by taking the average. Thereby, minimizing overfitting and retaining necessary spacial information.

E. Dense Layer with ReLU activation

The features from the GAP layer are fed into the Dense layer with ReLU (Rectified Linear Unit) activation function. This layer helps in consolidating different feature activations and interpreting features for the classification.

F. Fully Connected (FC) Layer

The features from the previous layer are passed through FC layer with sigmoid activation function. The sigmoid activation function transforms the input values into output probability values between 0 (Zero) and 1 (one).

G. Binary Classification

In the end a binary classification decision is made based on the sigmoid output. If the probability of the input CXR is > 0.5, then the CRX is classified as TB (TB positive). If the probabillity of the CXR is ≤ 0.5 then the CXR is classified as Normal (TB negative).

Our methodology ensures a comprehensive approach to training and evaluating multiple deep learning models on a TB image datasets, highlighting key steps from data preparation to final evaluation. In comparison to study by Le et al., (2022) [4] they have utilized three TB datasets (MC, SH, & India), while we have considered only two TB dataset (MC & SH) which are evaluated separately and later by combining them. Furthermore, we have considered only four

CNN models for performance evaluation of the datasets whereas their study incorporates five models for performance evaluation.

III. Results & Discussion

The confusion matrices as shown in Fig. 2 for VGG16, MobileNetV3, DenseNet121, and RegNet040 on the MC dataset reveal their classification performance. VGG16 correctly identifies all 12 Normal cases but misclassifies all 9 TB cases as Normal. MobileNetV3 performs better, correctly classifying 12 Normal and 4 TB cases, though 5 TB cases are misclassified. DenseNet121 correctly identifies 12 Normal and 1 TB case but misclassifies 8 TB cases. RegNet040 shows the best performance, accurately classifying 11 Normal and 6 TB cases, with fewer misclassifications, making it the most balanced model.

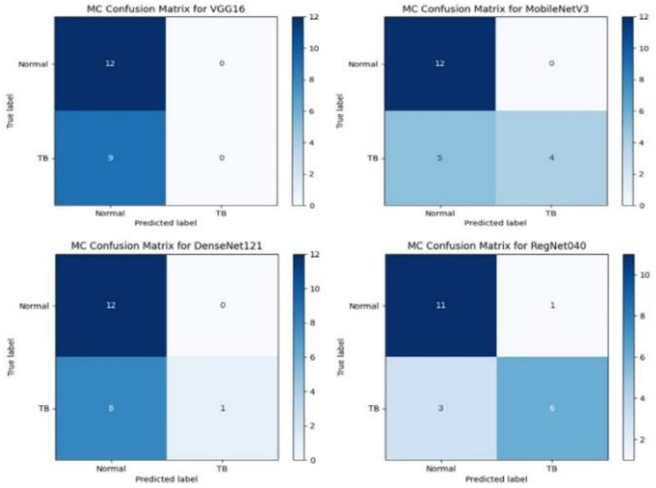

Fig. 2. Confusion matrices obtained using the MC dataset on four pre-trained CNNs

From Table 2, we can infer that our model performance for DenseNet121 and RegNet040 outperforms the results from Le et al. (2022), while VGG16 and MobileNetV3 perform worse.

TABLE II. Comparision of results on MC Dataset

MC Dataset				
Pre-trained Models	[4]		**Our Results**	
	Accuracy	**F1-Score**	**Accuracy**	**F1-Score**
VGG16	64.38%	68.66%	57.00%	42.00%
MobileNetV3	77.81%	78.92%	67.00%	60.00%
DenseNet121	60.94%	70.65%	76.00%	75.00%
RegNet040	71.56%	75.32%	81.00%	80.00%

In Fig. 3, confusion matrices, VGG16 correctly classifies 35 Normal and 40 TB cases but misclassifies 14 Normal and 11 TB cases. MobileNetV3 performs well with 45 correct Normal and 36 correct TB classifications, though it misclassifies 15 TB cases as Normal. DenseNet121 shows similar performance with 42 Normal and 41 TB cases correctly classified, but 10 TB and 7 Normal cases are misclassified. RegNet040 demonstrates the best performance, accurately identifying 44 Normal and 42 TB cases, with the fewest misclassifications among the models. Overall, RegNet040 stands out as the most balanced and accurate model.

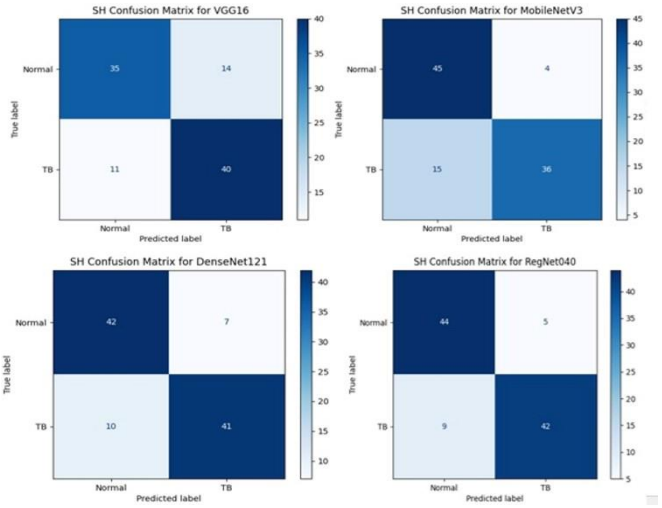

Fig. 3. Confusion matrices obtained using the SH dataset on four pre-trained CNNs

TABLE III. Comparision of results on SH Dataset

SH Dataset				
Pre-trained Models	[4]		**Our Results**	
	Accuracy	**F1-Score**	**Accuracy**	**F1-Score**
VGG16	64.38%	72.71%	75.00%	75.02%
MobileNetV3	67.19%	74.86%	81.00%	80.00%
DenseNet121	70.00%	71.57%	83.00%	83.00%
RegNet040	62.19%	68.43%	86.00%	86.00%

Table 3 indicates that our model performance for all four models (VGG16, MobileNetV3, DenseNet121, and RegNet040) outperforms the results from Le et al. (2022) on the SH dataset.

Fig. 4, confusion matrices have the best overall performance, accurately identifying both Normal and TB cases with minimal misclassifications. VGG16 performs well but tends to misclassify some TB cases as Normal. MobileNetV3 excels at identifying Normal cases but struggles significantly with TB detection. RegNet040 is highly accurate for Normal cases but has the highest rate of misclassifying TB cases. Overall, DenseNet121 is the most reliable model, while MobileNetV3 and RegNet040 need improvement in TB detection.

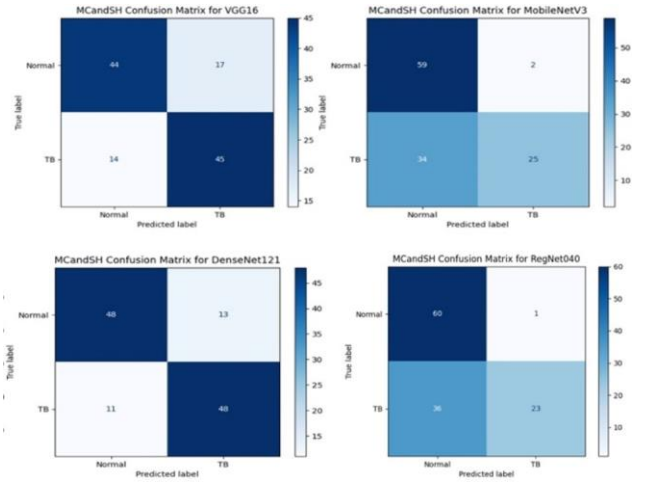

Fig. 4. Confusion matrices obtained using both the MC & SH datasets on four pre-trained CNNs

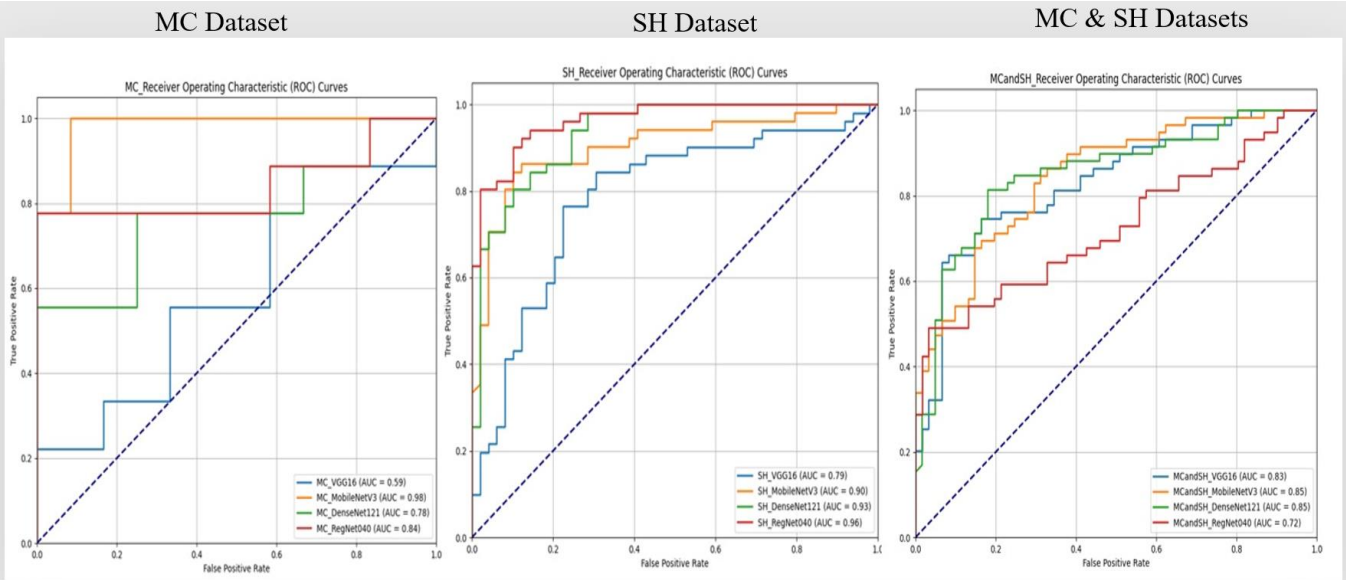

Fig. 5. Comparison of ROC curve performance for four CNN models evaluated on the MC, SH and combined (MC & SH) datasets.

TABLE IV. RESULTS ON MC & SH DATASET

COMBINED MC&SH DATASET		
PRE-TRAINED MODELS	**Accuracy**	**F1-Score**
VGG16	74.00%	74.00%
MOBILENETV3	70.00%	67.65%
DENSENET121	80.00%	80.00%
REGNET040	69.00%	65.67%

Analyzing Table 4, we can conclude that DenseNet121 demonstrates the best performance on the combined MC & SH dataset, achieving the highest accuracy and F1-Score at 80.00%. VGG16 also performs well with balanced metrics of 74.00% for both accuracy and F1-Score. MobileNetV3, despite having a decent accuracy of 70.00%, shows a very low F1-Score of 67.65%, indicating issues with class handling or imbalance. RegNet040 has moderate performance, with an accuracy of 69.00% and an F1-Score of 65.67%, slightly lower than VGG16.

IV. ROC CURVES

Receiver Operating Characteristic (ROC) Curve is a graphical representation to evaluate the performance of a binary classification. The ROC curve aids in assessing the trade-off between the true positives rate (TPR) and false positive rate (FPR) [10]. In the Fig. 5, ROC curves obtained from the datasets are discussed. Each plot in the Fig. 5 illustrates the performance of four different CNN classifiers. The Left plot shows the ROC curves obtained from the MC dataset. In the plot only MobileNetV3 shows a good performance result in comparison with the other three CNN models. The decrease in the performance would be due to the class imbalance present in the MC dataset. The plot in the middle is obtained from the SH dataset. It has the best performance results in comparison with the other two datasets. This may be due to the good quality CXR images present in the SH dataset with almost similar class balance of the CXR images in the dataset. The plot on the right is obtained from processing both the MC & SH as one dataset. Here, we can see that due to class imbalance and slightly low quality CXR images the performance of the combined dataset decreases in contrast with SH dataset. However, it has still better performance results than MC dataset.

V. CONCLUSION

In summary, MobileNetV3 consistently performs well, on all three datasets, especially on the MC dataset. The SH dataset provides better overall performance results, likely due to balanced classes and better quality images. Lastly, the combination of MC & SH dataset gives modest performance scores not surpassing the SH dataset results. Therefore, indicating that image quality and data balance play highly significant role in achieving high AUC scores. Furthermore, to enhance model performance, we can consider fine-tuning model parameters, and exploring different architectures. Employing regularization techniques can also aid in improving generalization. Experimenting with ensemble methods to combine predictions from multiple models may further boost overall performance.

REFERENCES

[1] WHO, *WHO operational handbook on tuberculosis. Module 1: prevention-infection prevention and control*. World Health Organization, 2023.

[2] WHO, "Global Tuberculosis Report 2023," 2023.

[3] A. Wong, J. R. H. Lee, H. Rahmat-Khah, A. Sabri, A. Alaref, and H. Liu, "TB-Net: a tailored, self-attention deep convolutional neural network design for detection of tuberculosis cases from chest X-ray images," *Frontiers in Artificial Intelligence*, vol. 5, 2022.

[4] T.-M. Le, B.-T. Nguyen-Tat, and V. M. Ngo, "Automated evaluation of tuberculosis using deep neural networks," *EAI Endorsed Transactions on Industrial Networks and Intelligent Systems*, vol. 9, no. 30, pp. e4–e4, 2022.

[5] S. Jaeger, S. Candemir, S. Antani, Y.-X. J. Wáng, P.-X. Lu, and G. Thoma, "Two public chest X-ray datasets for computer-aided screening of pulmonary diseases," *Quantitative imaging in medicine and surgery*, vol. 4, no. 6, p. 475, 2014.

[6] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[7] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, and others, "Searching for mobilenetv3," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1314–1324.

[8] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[9] I. Radosavovic, R. P. Kosaraju, R. Girshick, K. He, and P. Dollár, "Designing network design spaces," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10428–10436.

[10] P. Shankar, "Parametric modeling of receiver operating characteristics curves," *Model Assisted Statistics and Applications*, vol. 19, no. 2, pp. 211–221, 2024.