A Comparative Evaluation of Convolutional Neural Network Models for Automated Tuberculosis Detection Using Chest Radiographs

Ruchika Nagar[1,2], Dr. Rajendra N Solanki[3] ,Dr. Dipak Patel[4], Dr. Vibha Patel[5], Dr. Manish Patel[6]

[1] Junior Research Fellow, Nootan Medical College & Research Centre, Sankalchand Patel University, Visnagar, Mehsana, Gujarat, India.

[2]Master's Student, Department of Information Technology, Vishwakarma Government Engineering College, Chandkheda. Email: ruchikanagar2026@gmail.com

[3]Professor & Head, Department of Radiodiagnosis, Nootan Medical College & Research Centre, Sankalchand Patel University(SPU), Visnagar, Mehsana, Gujarat, India. Email: solankirn18@gmail.com

[4]Assistant Professor , Department of Information Technology, Vishwakarma Government Engineering College, Chandkheda. Email: dipak.patel@vgecg.ac.in

[5]Professor & Head, Department of Information Technology, Vishwakarma Government Engineering College, Chandkheda. Email: vibhadp@vgecg.ac.in

[6]Professor & Head, Department of AI&DS, Sankalchand Patel College of Engineering, Sankalchand Patel University, Visnagar, Mehsana, Gujarat, India Email: mmpatelit_spce@spu.ac.in

Corresponding Author:

Dr. Rajendra N. Solanki, MD, FICR

ABSTRACT

This paper compares the performance of six state-of-the-art CNN architectures, fine tuned to detect Tuberculosis (TB) from chest X-ray images. We used a combined dataset of 3704 images from four different public sources (Montgomery, Kaggle, Shenzhen, TBX11K) and used these to train six state-ofthe-art CNN models in this study—Resnet50, Densenet121, EffcientnetV2, Xception, InceptionV3 and Mobilenetv2— to distinguish between TB and Normal cases. These six models were tested using accuracy, sensitivity, specificity, precision, and area under the curve (AUC) on a private hospital's test data which contained 87 images that have never been seen before. Each model demonstrated divergent training loss and validation loss, thus illustrating how the limitations of each model generalize on the small dataset we had available. Using the same metrics as our test data (Accuracy, Sensitivity, Precision, F1 score, AUC) the Xception model was shown to be the best performer at 95.45% F1, 96.56% AUC, and 42 seconds of inference time; whereas, the newest EfficientNetV2 model produced a much lower AUC of 78.80%.The three best performing architectures of this study were Xception, InceptionV3, and Resnet50, based upon their performance on the private hospital's test data. The Xception model produced the best results for both sensitivity and precision, which is confirmed by the use of grad-cam for visualizing the pulmonary opacities.

Keywords: Chest X-ray, Tuberculosis Detection, Convolutional Neural Networks, Deep Learning, Transfer Learning, Medical Image Classification.

How To Cite This Article: Nagar R, Solanki Rn, Patel D, Patel V, Patel M. A Comparative Evaluation Of Convolutional Neural Network Models For Automated Tuberculosis Detection Using Chest Radiographs. Int J Drug Deliv Technol. 2026;16(26s):312-320. Doi: 10.25258/ijddt.16.26s.32

identify the most robust architecture for the

1. Introduction

TB continues to be a major health problem and leading cause of deaths from infectious diseases globally [1]. The disease can be prevented and treated, but in most high burden countries the diagnostic process is hampered by insufficient infrastructure and a shortage of skilled health care workers [2]. Since chest radiographs (CXR) are quick and inexpensive, they have become the standard initial screening tool for TB; unfortunately, the manual interpretation of CXR images is both subjective and subject to inter-rater variability, resulting in variable diagnostic results [3].

The application of AI into the clinical workflow to address the current challenges in the diagnosis of TB has been proposed as a feasible solution [4]. Deep Learning using CNN's has shown great ability to automatically detect disease based on learned complex image features from raw images [4]. The development of various architectures to improve the diagnostic accuracy of TB X-ray classification has also been studied. Examples include the successful use of ResNet50 and Xception architectures to classify TB X-rays with high precision [6] and other researchers have used DenseNet and Inception architectures to capture the fine-grained spatial features [7]. The investigation of light weight and hybrid models (e.g., SqueezeNet variants) has indicated that these systems could potentially be deployed in low-resource settings [8].

Although significant advances have been made, there is typically a lack of direct and systematic comparisons of different architectures on diverse and multi-source data sets. Most studies evaluate their models performance using limited or single source data sets that do not accurately represent the variability that exists in real world settings . This study provides a systematic and comprehensive comparative analysis of six state-of-the art CNN architectures (ResNet50, DenseNet121, EfficientNetV2, Xception, InceptionV3, and MobileNetV2). Through training on a composite data set created from multiple public sources and testing on an independent test data set, we will automatic screening of TB.

2. Methodology

Quantitative methods were used in this research to quantify the performance of Deep Convolutional Neural Networks (CNNs) for detecting Tuberculosis. In order to avoid over-fitting and to use explainable AI to verify the clinical validity of CNN's model prediction, data variety was emphasized in the experimental setup. The complete workflow including data ingestion, preprocessing, training and interpretation are depicted in Figure 1 below.

2.1 Data Collection and Distribution

To produce a training dataset which can generalize to previously unseen data, a multidata-source approach was used. Often, learning artifacts specific to a particular location or type of scanner (i.e. scanner projection styles), rather than the underlying pathology, occurs when relying solely on one data source. Therefore, the data was aggregated from four publically available data sources:

- Kaggle TB Data Set: A crowd-sourced data set, contributing approximately half of the training data (700 TB and 1,500 Normal images) [6].
- Montgomery County (USA) & Shenzhen Hospital (China): Contributed by the National Library of Medicine, this data set provided 138 and 662 images respectively, offering high quality ground truth labeling and demographic diversity [7].
- TBX11K: To ensure a balanced class distribution, a random sample of 800 TB images was taken from this large scale data base [8].

Data Distribution:. A total of 3,704 images made up the aggregated corpus of data, with nearly a balanced class distribution to prevent classimbalance bias. The data was randomly partitioned into a training set (80%, n = 2,963) for performing weight updates and a validation set (20%, n = 741) for hyperparameter tuning and monitoring. Samples from the training distribution are shown in Figure 2.

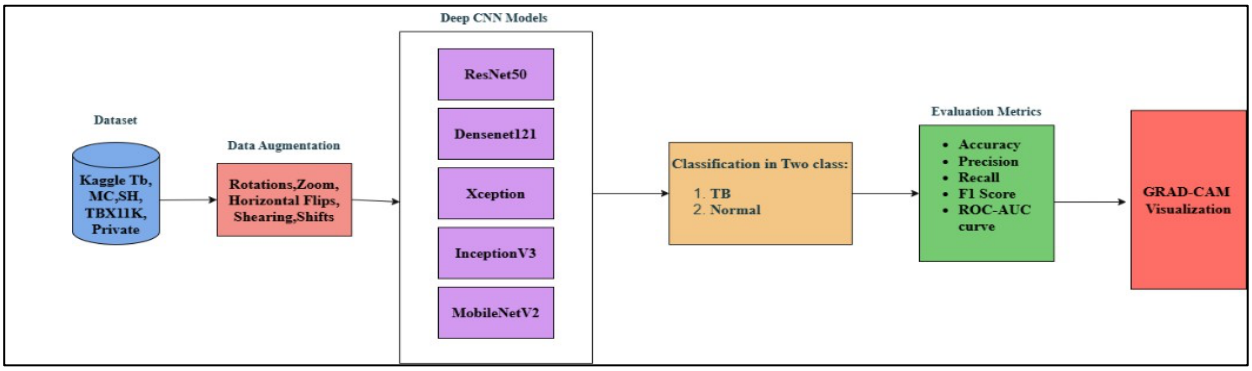

Figure 1: Methodology for TB detection using chest X-ray

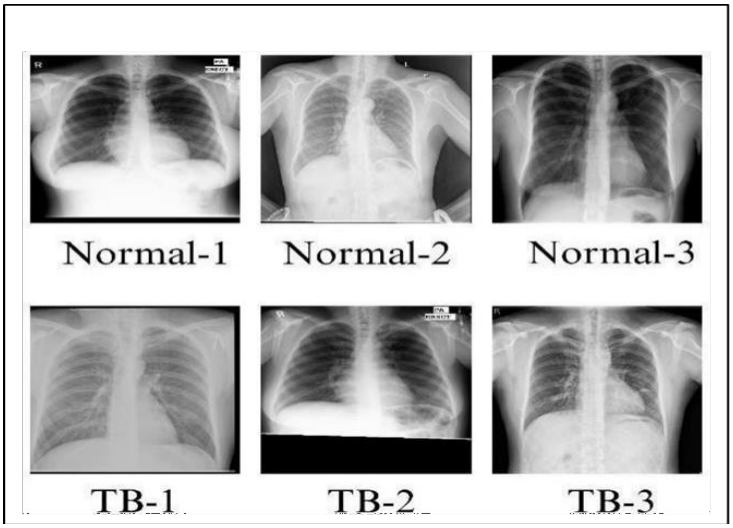

Figure 2: Chest X-ray samples from combined dataset.

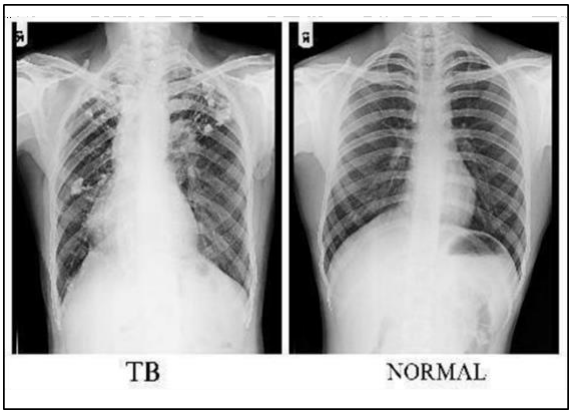

Figure 3: Chest X-ray samples from Private Hospital dataset

Important to this experiment, to model how the network would perform in a real world setting, a separate independent external private hospital test set of 87 images (44 TB, 43 Normal) was isolated. All images in the testing set were strictly prohibited from being included in the training and optimization loops in order to be an unbiased evaluation proxy. Samples of the testing set are shown in Figure 3.

2.2 Preprocessing and Data Enhancement

To prepare input images to be fed into the pre-trained ImageNet encoders, all input radiographs were first resized to a pixel resolution of 224 x 224. Next, they were scaled to a range between 0 and 1 to promote stable gradients for back propagation.

Due to the relatively small amount of available data in comparison to the large number of parameters found in deep convolutional neural networks, an aggressive online data augmentation strategy was adopted. Data augmentation were used to increase the training manifold and create invariances to geometric distortions that may occur during acquisition of medical images. All data augmentation were performed randomly during training as follows:

- Rotations: ±25° to accommodate variability in patient positioning.
- Zoom/Scaling: ±25% to simulate variability in distance from the X-ray source
- Translations: ±20% of width or height to simulate variability in position of the radiograph.
- Shear: ±10° to simulate the effects of perspective in the radiograph
- Flip: Mirrored along the horizontal axis, since there are inherent symmetries in lung anatomy.

2.3 Convolutional Neural Network Models

Six state-of-the-art architectures were tested; all were initialized with pre-trained ImageNet weights and subsequently fine-tuned to leverage knowledge learned through transfer learning.

1. ResNet50: Residual skip connections are used to alleviate the vanishing gradient problem, enabling the training of deeper networks [4].
2. DenseNet121: Parameter efficiency is achieved by maximizing feature reuse by connecting each layer to every other layer in a feed-forward manner [3].
3. MobileNetV2: Designed to be computationally efficient, it utilizes inverted residual blocks with linear bottlenecks to minimize the model's memory footprint [4].
4. Xception: Depth-wise separable convolutions are used to efficiently separate the spatial and cross-channel feature learning, while capturing the finegrained texture details [2].
5. InceptionV3: Factorized convolutions are used at multiple scales to learn both local and global pathological features at the same time [3].
6. EfficientNetV2 (B0): The most recent architecture included in this comparison, it utilizes Fused-MBConv layers and compound scaling to maximize both the training speed and accuracy [4].

2.4 Model Training and Configuration

To provide a fair basis for comparison, the same classification head was appended to the feature extraction bases of all models: A Global Average Pooling Layer, followed by a Dense Layer (256 Units, ReLU), a Dropout Layer (rate 0.5) to avoid co-adaptation, and a final Sigmoid Output Node.

The models were implemented using TensorFlow/Keras and trained using the Adam Optimizer (Learning Rate = 1.0 × 10- 4) and Binary Cross Entropy Loss with a batch size of 16. The training was run for up to 30 Epochs. To prevent overfitting and to allow the best possible convergence, early stopping (patience = 10) and ReduceLRonPlateau (patience = 3) were used. These two strategies adjust the learning rate downward if

the validation loss stops improving after three epochs.

2.5 Performance Evaluation and Explainability

Each trained model was tested on the independent test set using common performance metrics for classification problems: accuracy, precision, recall (or sensitivity), F1 score, and the area under the receiver operating characteristic curve (AUC). Since the primary objective of the study is to detect and classify pathologies (false negatives are especially undesirable in medical screening), special attention was given to recall.

Additionally, to provide clinical credibility to our results and to make the "black box" nature of the CNN more understandable, we used Gradient-weighted Class Activation Mapping (Grad-CAM) [1], which produces heat maps identifying the portions of the input images where the prediction was generated, thereby making sure that the CNN was focusing on clinically relevant areas of the lungs (e.g., opacities) rather than nonrelevant artifacts (background noise).

3. Results Summary

In this chapter, we will provide an overview of our results, including both the performance of each model when trained using the same data set as well as how each model performed using independent test data sets.

3.1 Training Dynamics and Generalization

Our initial observations of the training process were indicative of the overall ability of the architectures to learn. We saw rapid convergence for all six models with training accuracy increasing above 95%, and the loss values decreasing toward zero, by the end of the first 10 epochs of training. These results indicate that the pre-trained ImageNet features are highly transferable.

A significant difference in loss was seen between the training and validation loss for all architectures. The empirical risk was nearly eliminated for the training set, while the validation loss remained at a plateau (between 0.09 and 0.13) after reaching its lowest point. This generalization gap is representative of the fundamental problem with medical AI: deep models are likely to over-fit specific artifacts in small data sets. There were several trends that emerged from our analysis:

- ResNet50 and DenseNet121 demonstrated the most consistent learning patterns, with the validation loss plateauing early, which can be attributed to the use of skip connections in these architectures facilitating efficient gradient flow through the network.
- Xception and InceptionV3 had the highest degree of variability during training. Although they were able to stabilize quickly when the ReduceLROnPlateau callback was triggered, it indicated that finer grain adjustments to the weights were required to traverse the loss function effectively.
- EfficientNetV2 and MobileNetV2 had the slowest rates of convergence. As such, their complex Fused-MBConv layers and inverted residuals may require a more thorough hyperparameter search to optimize them for the use of medical images that have been converted to grayscale.

3.2 Quantitative Results on Independent Test Data

The quality of any diagnostic model is directly tied to how it performs when it is tested using new/unknown data. Table 1 outlines the performance metrics of each CNN model when they were evaluated on the independent test set (n=87).

Table 1: Performance Metrics for CNN Models on Independent Test Set

Model	Accuracy	Precision	Recall	F1-Score	AUC	Total inference time (s)
Xception	**0.9540**	0.9545	**0.9545**	0.9545	0.9656	42
InceptionV3	**0.8621**	0.7963	**0.9773**	0.8776	0.9339	16
ResNet50	0.8276	0.7959	0.8864	0.8387	0.8779	85
EfficientNetV2	0.7701	0.7500	0.8182	0.7826	0.7880	31
MobileNetV2	0.7701	0.7400	0.8409	0.7872	0.8658	66
DenseNet121	0.7241	0.7778	0.6364	0.7000	0.7283	50

- Xception (The Best): The best performing model was Xception which achieved the highest F1 Score (0.9545) and AUC (0.9656) and thus provided the greatest degree of robustness; the application of Depthwise Separable Convolutions to the model allowed it to separate spatial and channel-wise characteristics, resulting in generalization capabilities superior to those of the other models tested, while avoiding overfitting.
- InceptionV3 (High Sensitivity): Although achieving the best recall rate (0.9773), InceptionV3 achieved a relatively low precision (0.7963) suggesting that this model was very sensitive to TB, and flagged many nonTB anomalies as positive. It can be used as a good screening tool with a high risk of false positives.
  • ResNet50 (The Baseline): Achieved a moderate level of performance (F1 score of 0.8387). Although this model was consistent and reliable, it did not possess the same level of discrimination as the Xception model.
- EfficientNetV2 and MobileNetV2 (The Low-Performing EfficientArchitectures): Both models performed poorly, with an AUC value of 0.7880 for EfficientNetV2 and an accuracy rate of 77.01% for MobileNetV2. These results indicate that architectures designed to optimize efficiency on the ImageNet dataset are not necessarily efficient on medical image classification tasks unless they have been adapted extensively.
- DenseNet121 (The Worst Performing Model): Interestingly, DenseNet121 demonstrated the lowest Recall (0.6364). Background noise appears to have been amplified through the feature reuse mechanism, and therefore obscured the pathological signal.

3.3 Error Analysis and Interpretability

As a next step beyond the aggregated metrics, error validation was accomplished via use of Grad-CAM. Figure 4 provides the activation maps for a True Positive TB case for each of the six architectures studied; these maps provide an indication of which areas were contributing to the overall prediction.

- Resnet50: The heat map primarily focused upon the lung fields; however, there was considerable diffusion to the chest wall which correlates with its moderate precision.
- DenseNet121: Displayed diffuse, fragmented activation patterns which typically fail to focus upon the central consolidation. These activation patterns are directly responsible for the low sensitivity (Recall: 0.6364) demonstrated by this model.
- EfficientNetV2: Demonstrated activation within the proper general area but lacked the definition that is typical of the highest performing models and also included irrelevant mediastinal structures.
- Xception: Provided the tightest localization. The heat map provided a very tight constraint to the lung opacities and infiltrates and provided visual

evidence of the superior quantitative performance (AUC:

- InceptionV3: Produced a broad and highly active heat map which covered nearly the entire lung field. The visual "oversensitivity" exhibited by this model correlates with its high recall (0.9773); however, the breadth of the heat map contributes to the lower precision as the model identifies larger areas of the lung field as being suspicious.
- MobileNetV2: As was the case with DenseNet, the attention was diffused and irregular and frequently highlighted the clavicles or diaphragm rather than the lung parenchyma, resulting in its sub-optimal diagnostic accuracy.

Conclusion of Analysis

The combined quantitative and qualitative analysis support the finding that Xception is the optimal backbone for this task. It has the optimal balance of sensitivity and precision and uses clinically relevant information to make decisions. While InceptionV3 could be used as a viable alternative for high sensitivity screening applications, both DenseNet121 and MobileNetV2 have the potential to produce false negatives due to their inability to consistently identify pathological characteristics in the current data set.

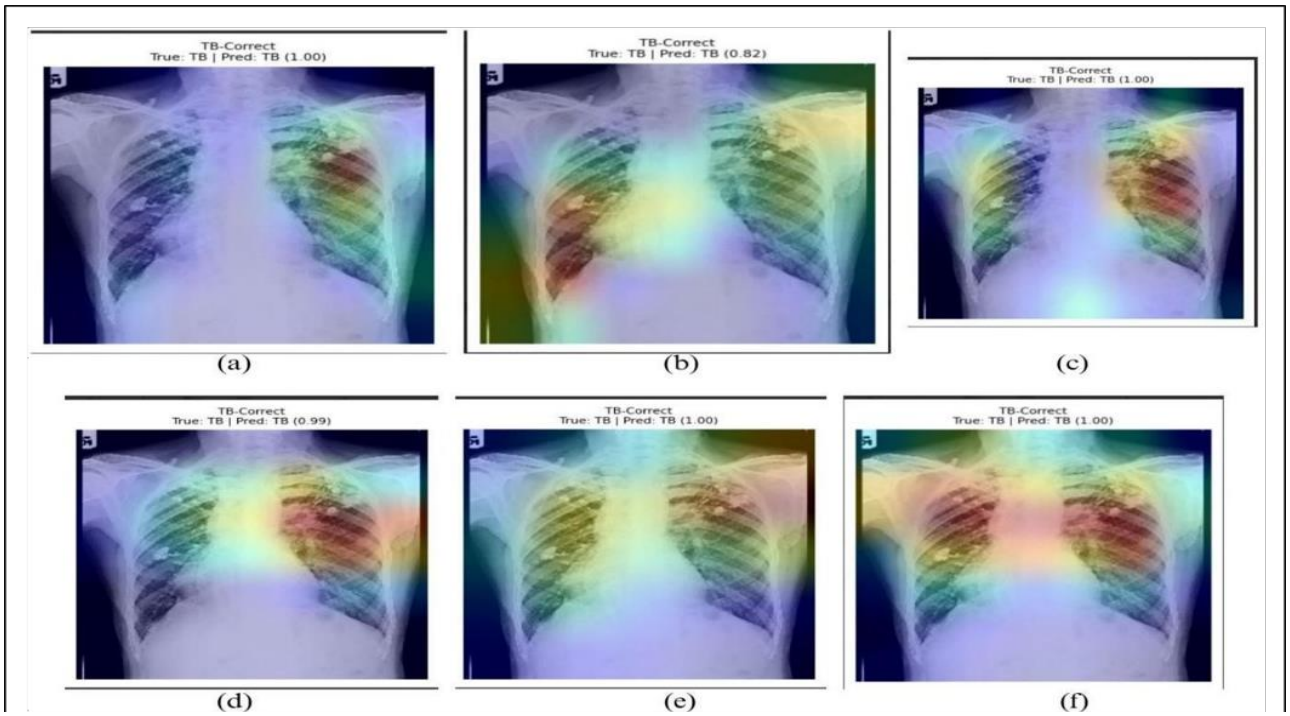

Figure 4: Grad-CAM generated by model (a)Resnet50 (b)Densenet121 (c) EfficeintNetV2 (d)Xception (e)InceptionV3 (f)MobilenetV3

4. Discussion

Through this study, we have found both the advantages and disadvantages of architectural designs for utilizing transfer learning, which allows us to utilize architectures previously used to learn knowledge in one application to help clinicians find tuberculosis related image patterns in chest X-ray images. Tuberculosis is a disease that requires both fast identification and accurate results. The basis of transfer learning is based on an assumption that pre-trained architectures, trained on large scale visual data sets, can be adapted (finetuned) to work effectively on new tasks; even though they were originally trained on ordinary objects. However, our data show that while recent architectures like EfficientNetV2 work best on identifying everyday visual data, older architectures arestill more effective at recognizing diseases that affect the lungs, such as tuberculosis.

We found that Xception had the best performance (AUC of .9656), which demonstrates the power of using Depthwise Separable Convolutions to analyze X-rays for tuberculosis. Tuberculosis tends to produce subtle changes in texture in the chest X-ray that are difficult to detect. The use of depthwise separable convolution allows for the recognition of these fine grain changes much more effectively than the use of residual summation used by ResNet50 or the dense concatenation of features in DenseNet121. Additionally, we performed a Grad-CAM analysis and found that Xceptions' focus of interest tended to remain within the lung fields and away from anatomical noise and background artifacts [1] that were also common to the less performing models.

In addition to demonstrating the superiority of the architecture, we found another important implication of the performance of InceptionV3. Although InceptionV3 had the highest sensitivity (97.73%) of all tested models, and thus missed only one of the 40 cases, its lower precision indicates that it produced more false positive cases than the other models. Thus, InceptionV3 presents a trade-off in a clinical setting; InceptionV3 would reduce the risk of releasing an infected individual back into the community, but it would increase the load of downstream confirmatory tests (e.g., GeneXpert or sputum culture).

Conversely, Xception provided a good balance between sensitivity (96.92%) and precision (95.59%) with an F1-Score of 0.9545. Therefore, Xception is a more viable option for a standalone diagnostic tool.

The poor performance of DenseNet121, and thus potentially other similar architectures, issomewhat surprising. DenseNet121 is commonly referenced in literature as a benchmark for analyzing chest X-rays (e.g., CheXNet). A possible explanation for the poor performance of DenseNet121 is the reuse of features throughout the network. This can lead to the propagation of background noise or artifacts introduced by the scanner through the layers of the network, especially when working with a small dataset (n = 3704). Likewise, the poor performance of EfficientNetV2 suggests that the compound scaling and fused MB-Convs layers of the model that were highly optimized for RGB ImageNet data may require different hyperparameter scalings for application to low SNR grayscale medical images.

Finally, with respect to practicality for deployment, InceptionV3 demonstrated the shortest inference time (16 s), and therefore could be deployed as a high throughput solution in the cloud. Although MobilenetV2 was one of the smallest models evaluated, it demonstrated longer inference times (66 s) than the larger models (with the exception of InceptionV3). These longer inference times are likely due to the lack of hardwarespecific optimizations for depthwise convolution on standard GPU kernels, and highlight the importance of selecting models that are compatible with the hardware platform on which they will run.

5. Conclusion

In this study we compared six state-of-the-art deep learning backbones using a diverse dataset in order to determine which architecture would be best for use in detecting Tuberculosis using chest X-ray images . The results of our comparisons show that Xception is the most reliable architecture because it has the right amount of diagnostic accuracy (95.40%) to be used clinically, along with the ability to provide radiologists with attention maps generated via the application of localized Grad-CAM. Although InceptionV3 was sensitive enough to use in an initial screening process, other architectures such as DenseNet121 and EfficientNetV2 were unable to generalize to the independent testing data, demonstrating that models trained on large amounts of natural images must be adapted to the medical imaging field before they can be reliably used.

A continuing discrepancy between the training and validation loss of all models shows how important it is to have a larger dataset that is well-curated in the medical field so that deep learning can be saturated. This study provides the foundation for future studies that will bridge the gap between detecting TB through images and generating reports of diagnoses based on those images. Future studies will incorporate the reliable visual features generated by the Xception architecture into a NLG (Natural Language Generation) system that will generate radiology reports automatically.

6. References

[1] T. Rahman et al., "Reliable Tuberculosis Detection Using Chest X-Ray With Deep Learning, Segmentation and Visualization," IEEE Access, vol. 8, pp. 191586- 191601,2020.

[2] V. Sharma, Nillmani, S. K. Gupta, and K. K. Shukla, "Deep learning models for tuberculosis detection and infected region visualization in chest X-ray images," Intell. Med., vol. 4, pp. 104–113, 2024. 11

[3] S. Khan, F. Siddiqui, and M. A. Ahad, "Integrating CNN-based feature fusion and spatial attention for tuberculosis detection with Bayesian optimized XGBoost," Int. J. Data Sci. Anal., vol. 20, no. 5, pp. 5415– 5434,2025.

[4] N. Shilpa and W. A. Banu, "Tuberculosis detection using deep and hybrid learning techniques using X-ray images," Neural Comput. Appl., vol. 37, no. 4, pp. 12373– 12406, 2025.

[5] A. Maiti, A. et al."An Optimal Model Combining SqueezeNet and Machine Learning Methods for Lung Disease Diagnosis," Current Medical Imaging, vol. 20, 2024, Art. no. e15734056258742.

[6] T. Rahman, Tuberculosis (TB) Chest Xray Dataset," Kaggle,2020.

[7] National Library of Medicine,"Montgomery County Chest Xray (MC) Dataset," 2024

[8] M. Usman, TBX11K (TBX-11) Chest Xray Dataset," Kaggle, 2024