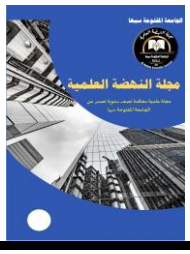
مجلة النهضة العلمية

Al-Nahda Scientific Journal

Journal homepage :https://ous.edu.ly/Journal/index.php/OUSJ



Visual Explanations of Deep Learning Approach for Tuberculosis Detection

Abdelkader Alrabai
Physics Department, Faculty of Education, Wadi Alshatti University, Alshatti – Libya
*Corresponding Email: a.alrabai@wau.edu.ly

Abstract

Tuberculosis (TB) continues to pose a significant global health threat, especially in low-resource regions where timely diagnosis is often challenging. Recent advancements in AI and deep learning have shown great promise in automating disease detection through analysis of chest X-ray images, a widely used and cost-effective diagnostic method. This study employed a deep convolutional neural network (CNN) specially VGG16 architecture—pre-trained on the ImageNet dataset—to automatically classify chest radiographs as TB or normal. The model was evaluated using standard performance metrics and achieved an impressive accuracy of 99.93%, demonstrating strong capability in identifying TB from X-ray images. To promote transparency and clinical trust, integrated gradients was incorporated as an explainable AI technique. Integrated gradients visualizations pinpoint the lung regions most influential to the model's predictions, enabling healthcare professionals to better understand and validate the AI's decisions. Ultimately, these results presented a promising, interpretable, and scalable approach for automated TB screening—particularly valuable in resource-limited healthcare settings—and support the potential integration of such systems into clinical decision support tools.

Keywords: CNN, Explainable, Integrated gradients, Tuberculosis.

التفسيرات البصرية لنهج التعلم العميق في الكشف عن مرض السل

عبدالقادر الربيعي
قسم الفيزياء، كلية التربية، جامعة وادي الشاطئ، الشاطئ – ليبيا
البريد الإلكتروني للمراسلة: a.alrabai@wau.edu.ly

الملخص

لا يزال مرض السل يشكل تهديدا كبيرا للصحة العالمية، لا سيما في المناطق ذات الموارد المنخفضة حيث غالبا ما يكون التشخيص في الوقت المناسب صعبا. أظهرت التطورات الحديثة في الذكاء الاصطناعي والتعلم العميق وعدا كبيرا في أتمتة الكشف عن الأمراض من خلال تحليل صور الأشعة السينية للصدر، وهي طريقة تشخيصية مستخدمة على نطاق واسع وفعالة من حيث التكلفة. في هذه الدراسة استخدمت شبكة عصبية تلافيفية عميقة (VGG16) مدربة مسبقا على مجموعة بيانات (ImageNet)–لتصنيف الصور الشعاعية للصدر تلقائيا على أنها مصابة أو سليمة. تم تقييم النموذج باستخدام مقاييس الأداء القياسية وحقق دقة عالية بلغت 99.93٪، مما يدل على قدرتها القوية في تحديد مرض السل من صور الأشعة السينية. لتعزيز الشفافية والثقة السريرية، تم دمج التدرجات المتكاملة كأسلوب ذكاء اصطناعي يمكن تفسيره. تحدد تصورات التدرجات المتكاملة مناطق الرئة الأكثر تأثيرا على تنبؤات النموذج، مما يمكن المتخصصين في الرعاية الصحية من فهم قرارات الذكاء الاصطناعي والتحقق منها بشكل أفضل. في نهاية المطاف، قدمت هذه النتائج نهجا واعدا وقابل للتفسير وقابل للتطوير لفحص السل الآلي – وهو ذو قيمة

Received 15 December 2025 - Accepted 04 January 2026
DOI: 10.65588/wscyy775

خاصة في إعدادات الرعاية الصحية المحدودة الموارد ودعم التكامل المحتمل لهذه الأنظمة في أدوات دعم القرار السريري.

الكلمات المفتاحية: التدرجات المتكاملة، شبكة عصبية تلافيفية، التفسير، مرض السل.

1. Introduction

Tuberculosis (TB) remains one of the leading global health challenges, contributing significantly to illness and death. It is caused by the bacterium Mycobacterium tuberculosis, which released into the air by infected individuals. Roughly 25% of the global population shows immune system signs of exposure to the infection. This exposure may stay inactive or eventually develop into a symptomatic illness. Individuals who carry the bacteria without showing symptoms are classified as having latent TB, or (TB infection). In contrast, those who develop symptoms and clinical signs are said to have TB disease [1]. TB remains a major global health concern, affecting millions each year. Prompt diagnosis is essential for effective treatment and limiting transmission. Early detection not only improves outcomes and prevents complications but also helps contain the disease by identifying and isolating cases. However, delays in diagnosis persist due to limited access to testing, socio-economic challenges, low awareness, and the disease's often nonspecific symptoms. Although diagnostic tools have improved in accuracy, overcoming barriers such as stigma and healthcare access is still key to controlling TB effectively [2]. While the lungs are most frequently involved in TB (pulmonary TB), the infection can also spread to other areas such as the pleura, lymph nodes, skin, bones, genitourinary system, abdomen, joints, and the meninges — a form referred to as extrapulmonary TB [3]. Active TB is diagnosed using chest radiographs along with microscopic analysis and culture of bodily fluids. In contrast, latent TB is identified through either a tuberculin skin test or specific blood-based assays [4]. Traditional chest X-rays remain the primary tool for screening, diagnosing, and monitoring treatment in both pulmonary and extrapulmonary tuberculosis. However, this longstanding method has limitations in accuracy. Recent advances, such as digital radiography and computer-aided diagnosis (CAD) have transformed TB detection. Over the past few years, innovations including AI and machine learning have further enhanced diagnostic processes by improving data management and image analysis. These technologies are increasingly being applied in medicine, offering significant potential for enhancing TB diagnosis and care [5]. For many years, medical image analysis and interpretation have been carried out by humans. However, the rapid progress of AI has led to a growing adoption of computer-assisted tools in healthcare to enhance diagnostic accuracy and efficiency. These technologies enable real-time disease prediction and detailed evaluation of treatment alternatives, while reducing issues like inconsistencies between different observers, errors due to variability in disease presentations, and the fatigue experienced by human specialists [6].

CAD systems use AI to analyze radiological images and address the shortage of radiologists, particularly in developing regions. These systems are commonly employed to detect various diseases from medical images, with machine learning and deep learning being the primary AI techniques used for analyzing chest X-rays. The rapid growth of medical imaging data has made manual interpretation increasingly challenging for radiologists. When bacteriological tests are inconclusive, radiology—and specifically CAD analysis of chest X-rays—plays a critical role

in TB diagnosis. Deep learning, especially through deep CNNs, has become a leading approach in radiology, enabling effective feature extraction and accurate classification of chest X-rays as normal or abnormal for TB detection [7]. Understanding the reasoning behind a deep learning model's decision, especially in disease diagnosis, is crucial for assessing its trustworthiness. Many doctors and radiologists are hesitant to rely on deep learning models because they cannot observe what the network is focusing on to generate predictions. This limitation, often called the black box problem, arises because the internal processes within the hidden layers are not transparent. When the connection between inputs and outputs is unclear, even a single incorrect prediction can have serious, potentially fatal consequences. Visualization methods help by revealing whether the model is concentrating on the relevant regions of X-ray images or mistakenly using irrelevant information for its classification [8].

This work focuses on applying deep learning methods to detect TB from chest X-ray images, with particular attention given to understanding how the model reaches its decisions. In conjunction with achieving strong diagnostic accuracy, explainability techniques are incorporated to reveal the image regions and features that influence predictions. By making the model's behavior more transparent, the study aims to strengthen clinical confidence, improve interpretability, and support the practical use of AI-based TB diagnostic systems in routine healthcare practice.

2. Related works

Numerous studies have explored the use of AI, especially deep learning, to enhance TB detection and diagnosis using chest X-ray images. Alongside these efforts, researchers have developed approaches to make these models more interpretable, with certain investigations concentrating specifically on explanation methods tailored to TB detection.

Mirugwe et al. [9] evaluated six CNN architectures (VGG16, VGG19, ResNet50/101/152, and Inception-ResNet-V2) for classifying chest X-ray images as normal or TB. Using a dataset of 4,200 images (700 TB-positive, 3500 normal), VGG16 achieved the best performance, reaching 99.4% accuracy, while requiring fewer computational resources. The study concluded that simpler architectures like VGG16 provide an optimal balance between diagnostic accuracy and computational efficiency for TB detection in chest X-ray images, emphasizing task-specific model selection for clinical applications.

In addition, ***Sharma et al.*** [10] proposed a deep-learning framework for TB detection in chest X-ray images that integrates lung segmentation, classification, and visual explainability. A UNet model was first trained on 704 chest X-rays to segment lung regions. The trained UNet was then applied to 1400 TB and normal images from the NIAID TB portal dataset to extract lung areas. For classification, an Xception model was used to distinguish TB from normal cases based on the segmented lungs. Model interpretability was enhanced using Grad-CAM heatmaps to visualize TB-related abnormalities from a radiological perspective. The Xception classifier achieved 99.29% accuracy.

Moreover, ***Rahman et al.*** [11] proposed a comprehensive deep-learning framework for reliable TB detection from chest X-ray images by integrating image preprocessing, data augmentation, lung segmentation, and classification. A balanced dataset of 7000 chest X-rays (3500 TB and 3500 normal) was constructed from multiple public databases. Nine pre-trained CNN models

(ResNet18, ResNet50, ResNet101, ChexNet, InceptionV3, VGG19, DenseNet201, SqueezeNet, and MobileNet) were evaluated using transfer learning. Three experiments were conducted: lung segmentation using two U-Net models, classification using full X-ray images, and classification using segmented lung regions. ChexNet achieved the best performance on whole X-ray images, with an accuracy of 96.47%. However, classification using segmented lung images outperformed full-image classification, where DenseNet201 achieved 98.6% accuracy and strong performance across all metrics. Visualization techniques confirmed that CNNs primarily learned from lung regions, explaining the improved results.

Furthermore, ***Noviandy et al.*** [12] addressed the challenge of TB diagnosis in low-resource settings by evaluating lightweight and explainable deep learning models for chest X-ray analysis. Three lightweight models—ShuffleNetV2, SqueezeNet, and MobileNetV3—were assessed for binary TB classification using a local dataset of 3008 X-ray images. Transfer learning was employed to enhance performance, and Grad-CAM was used to provide visual explanations of model decisions. MobileNetV3 and ShuffleNetV2 achieved perfect classification results, each attaining 100% accuracy.

As well, ***Maheswari et al.*** [13] proposed a lightweight and interpretable shallow CNN for automated TB screening from chest X-ray images. In contrast to very deep CNN architectures, the study emphasized a simpler model to enhance diagnostic transparency and clinical applicability. The proposed shallow-CNN comprised four convolution–max pooling layers, with hyperparameters optimized using Bayesian optimization. The model achieved accuracy of 95%. To further enhance model explainability, CAM and LIME were employed and compared with a state-of-the-art pre-trained DenseNet model.

Lastly, ***Özkurt*** [14] investigated TB diagnosis using deep learning with a strong emphasis on explainable AI to improve trust and reliability in clinical applications. The dataset comprised chest X-ray images of both TB-positive and normal cases, including 700 publicly available TB images, an additional 2800 TB images accessible via the NIAID TB portal, and 3500 normal images. The proposed CNN architecture consisted of three convolutional layers was used for binary TB classification. Explainable AI techniques such as SHAP and LIME were applied to interpret model predictions and identify influential image features.

The previous studies consistently demonstrate that deep learning–based approaches can achieve high accuracy in TB detection from chest X-ray images. The findings highlight that incorporating transfer learning, and data-efficient architectures significantly improves diagnostic performance, while lightweight and shallow CNNs can rival deeper models with reduced computational cost. Moreover, explainable AI techniques such as Grad-CAM, CAM, LIME, and SHAP play a crucial role in enhancing model transparency and clinical trust by localizing TB-related abnormalities. Collectively, these works emphasize the importance of balancing accuracy, efficiency, and interpretability to enable reliable and scalable AI-assisted TB screening, particularly in resource-constrained healthcare settings.

This study utilizes an expanded dataset of 7208 chest X-ray images, created by merging two separate sources, with data augmentation applied to enhance model performance. In addition to improving diagnostic accuracy, an interpretability method—integrated gradients—is implemented to provide clear insights into the model's decision-making. The study aims to achieve precise TB detection while emphasizing the importance of explanation techniques in

building clinician trust and understanding, thereby enhancing the transparency, reliability, and effectiveness of AI-driven TB screening tools.

3. Methodology

This study employed a deep learning approach to identify TB in chest X-ray scans, focusing on making the results more understandable through integrated gradient technique. The process involved categorizing the X-ray images into two groups: normal or TB. The key stages of this approach are illustrated in Figure 1.

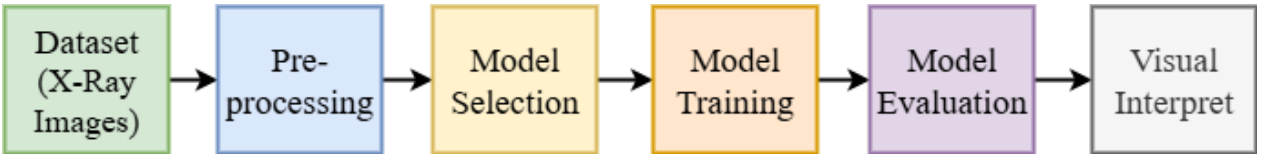


Figure 1. Method steps

3.1. Dataset and preprocessing

The dataset used in this study is a combination of two sources. The first dataset, obtained from Mendeley [15], contains 2494 chest X-ray images of TB patients and 514 normal chest X-ray images. The second dataset was compiled by [13] and consists of a well-organized collection of chest X-ray images—700 TB cases and 3500 normal cases—which is publicly available on Kaggle [16]. These two datasets were merged to form the final dataset used in this study, resulting in a total of 7208 images, comprising 3194 TB and 4014 normal images. Figure 2 illustrates sample images from the combined dataset, showing examples of both TB and normal cases.

All images were resized to a standardized resolution of 224 × 224 pixels and normalized. Additionally, 20% of the dataset was set aside for testing purposes. To enhance model generalization and reduce overfitting, augmentation techniques were applied during training. Random horizontal flip was used to simulate variations in image orientation by flipping images horizontally. Random rotation with a maximum angle of ±10 degrees. Additionally, contrast adjust employed introducing slight changes in image appearance that reflect real-world variability in X-ray imaging conditions. Together, these augmentations increase dataset diversity and improve the model's robustness in clinical scenarios.

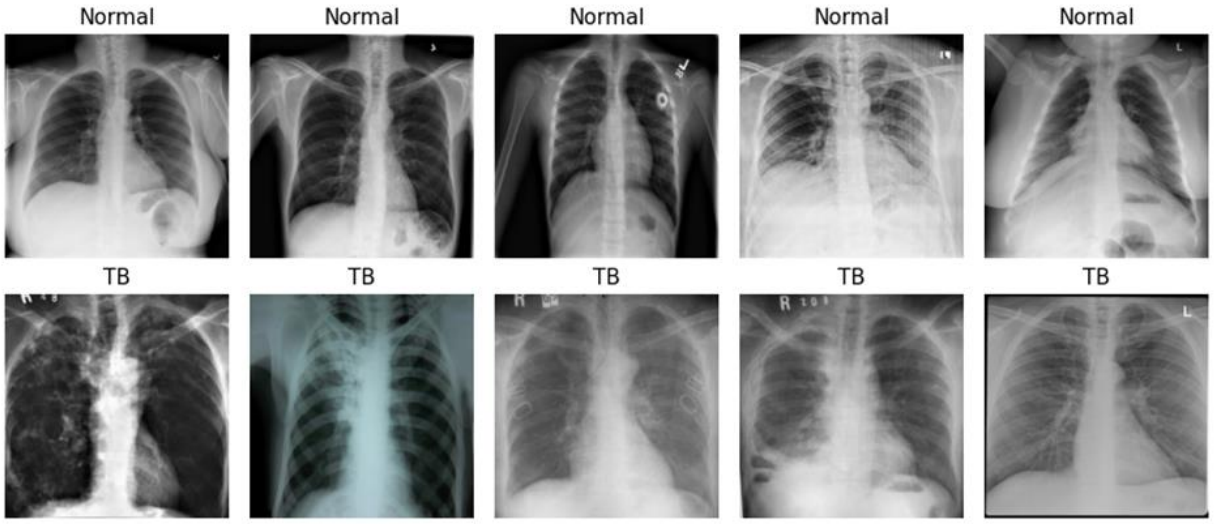


Figure 2. Sample from dataset

3.2. Model training and evaluation

A pre-trained CNN based on the VGG architecture—specifically VGG16—was employed as the backbone for feature extraction. The model was fine-tuned on the TB dataset, with its final fully connected layers modified to perform binary classification between TB and normal images.

VGG16 [17] is a deep CNN architecture consists of 16 weight layers, including 13 convolutional layers and 3 fully connected layers, followed by a softmax output layer. VGG16 is known for its simplicity and uniform architecture, using small 3×3 convolutional filters and 2×2 max-pooling layers throughout the network. VGG16 has been widely used in image classification and transfer learning tasks due to its strong performance and ease of implementation. Pre-trained versions of the model, especially those trained on the ImageNet dataset, are commonly used as feature extractors in various computer vision applications.

The model was trained using the Cross-Entropy loss function and the Adam optimizer with a learning rate of 0.0001. Training was conducted with a batch size of 16, and early stopping was employed to prevent overfitting, halting the process after 15 epochs. The model's performance was assessed on a separate, unseen test set using standard evaluation metrics to provide a comprehensive understanding of its diagnostic effectiveness. Additionally, a confusion matrix was generated to further analyze the classification results.

3.3. Explainability and visualization

To enhance interpretability, integrated gradients was used to visualize the regions in the chest X-rays that influenced the model’s predictions.

Integrated Gradients [18] is an attribution method used to explain the predictions of deep learning models. It works by quantifying the contribution of each input feature (such as pixels in an image) to the model's output. The method involves computing the integral of the model’s gradients as the input varies along a straight path from a baseline (often a black image or zero input) to the actual input. This provides a more reliable and theoretically grounded explanation compared to simple gradient-based methods, which can be noisy. The outputs were plotted in three-panel layouts: the original X-ray, the attribution heatmap, and an overlay of both. These

visualizations provided intuitive insights into model behavior and supported transparent decision-making.

4. Results and Discussions

The performance of the model during training is illustrated in Figure 3, which displays the training and validation accuracy and loss curves across epochs.

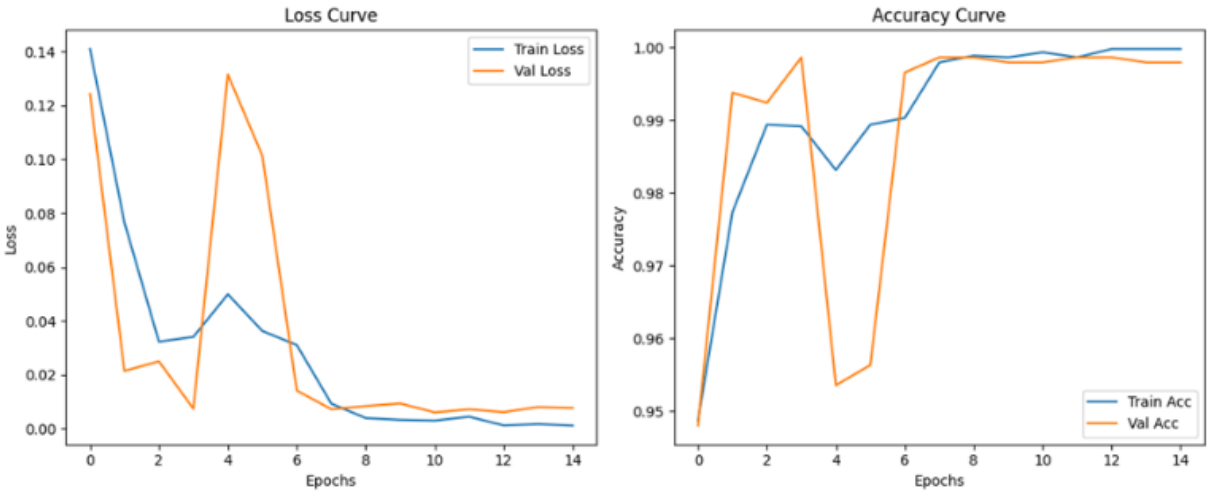


Figure 3. Training curves

During the early stages of training, specifically between epochs 3 and 6, the model exhibits signs of overfitting, as indicated by an increase in validation loss and a decrease in validation accuracy, despite continued improvement in training performance. This suggests that the model began memorizing the training data rather than learning patterns that generalize well to unseen data. However, after epoch 6 or 7, both the training and validation loss curves steadily decline and converge to low values, while accuracy for both sets rises to approximately 99%. This convergence indicates that the model was able to recover from the initial overfitting phase and ultimately generalize well. The final state—characterized by low validation loss and high validation accuracy—demonstrates effective training and strong overall performance. The model was evaluated on an unseen test set of chest X-ray images. The performance results are summarized in Table 1, which presents the classification metrics of the trained model across the two classes (Normal and TB).

Table 1 Classification performance on the test set

Accuracy	Class	Precision	Recall	F1-score	Support
0.9993	Normal	0.9988	1.0000	0.9994	803
	TB	1.0000	0.9984	0.9992	639

The model achieved an overall accuracy of 99.93%, indicating exceptional performance on the test set. For the Normal class, the model attained a precision of 0.9988 and a recall of 1.0000, resulting in an F1-score of 0.9994, with all 803 Normal cases correctly identified. For the TB class, the model achieved a perfect precision of 1.0000 and a recall of 0.9984, leading to an F1-score of 0.9992, based on 639 samples. These metrics suggest a near-perfect balance between precision and recall for both classes, with minimal false positives and false negatives. The

consistently high scores across all metrics demonstrate the model's strong capability to distinguish between TB and normal chest X-ray images, making it a reliable tool for automated TB screening.

Figure 4 presents a confusion matrix, which quantifies the model's classification accuracy. It clearly shows a high number of true positives and true negatives. Specifically, 803 normal cases were correctly identified as normal (true negative), and 638 TB cases were correctly identified as TB (true positive). Crucially, there are zero false positives (no normal cases misclassified as TB) and only 1 false negative (one TB case misclassified as normal). This nearly perfect confusion matrix signifies exceptional accuracy and precision in distinguishing between normal and TB affected lungs.

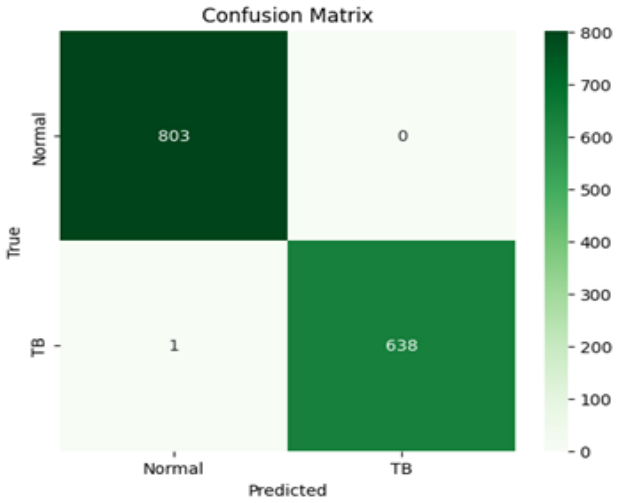


Figure 4. Confusion matrix

Figure 5 illustrates the model's predictions on sample chest X-ray images, showing strong performance in detecting TB cases, all of which were correctly classified with 100% confidence. Among the normal cases, four were correctly identified, while one was misclassified as TB with full confidence—indicating a false positive. This suggests the model is highly sensitive to TB features but may occasionally over-predict TB in normal images. Overall, the model demonstrates excellent accuracy, particularly in identifying TB, though refinement is needed to reduce false positives in normal cases.

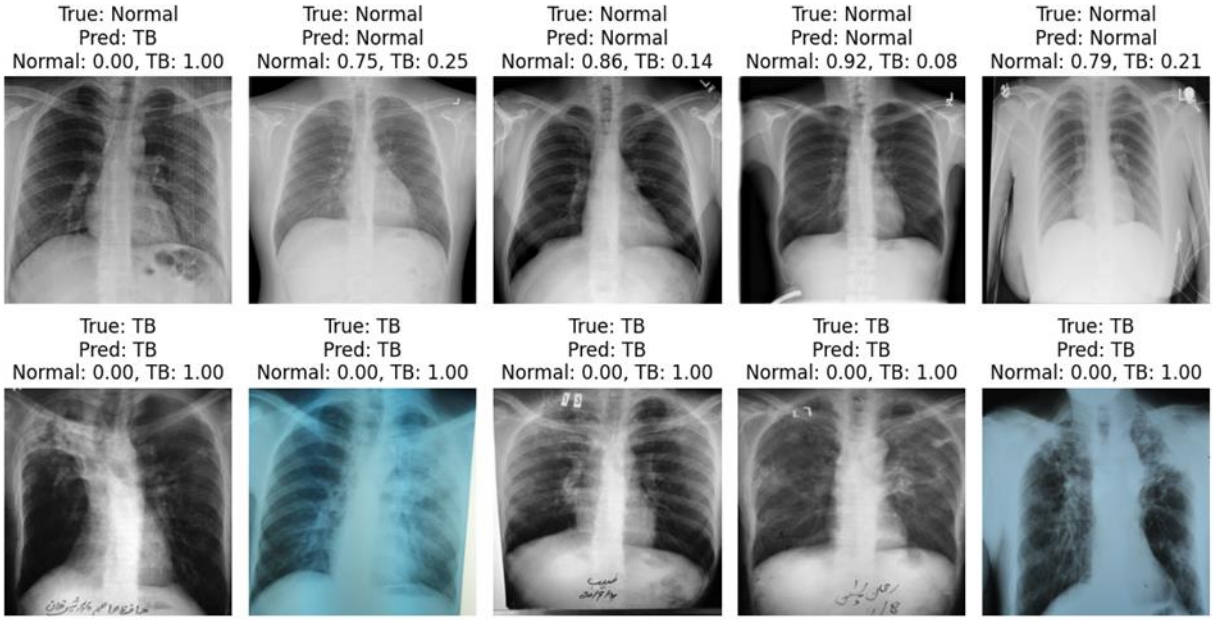


Figure 5. Sample of predictions

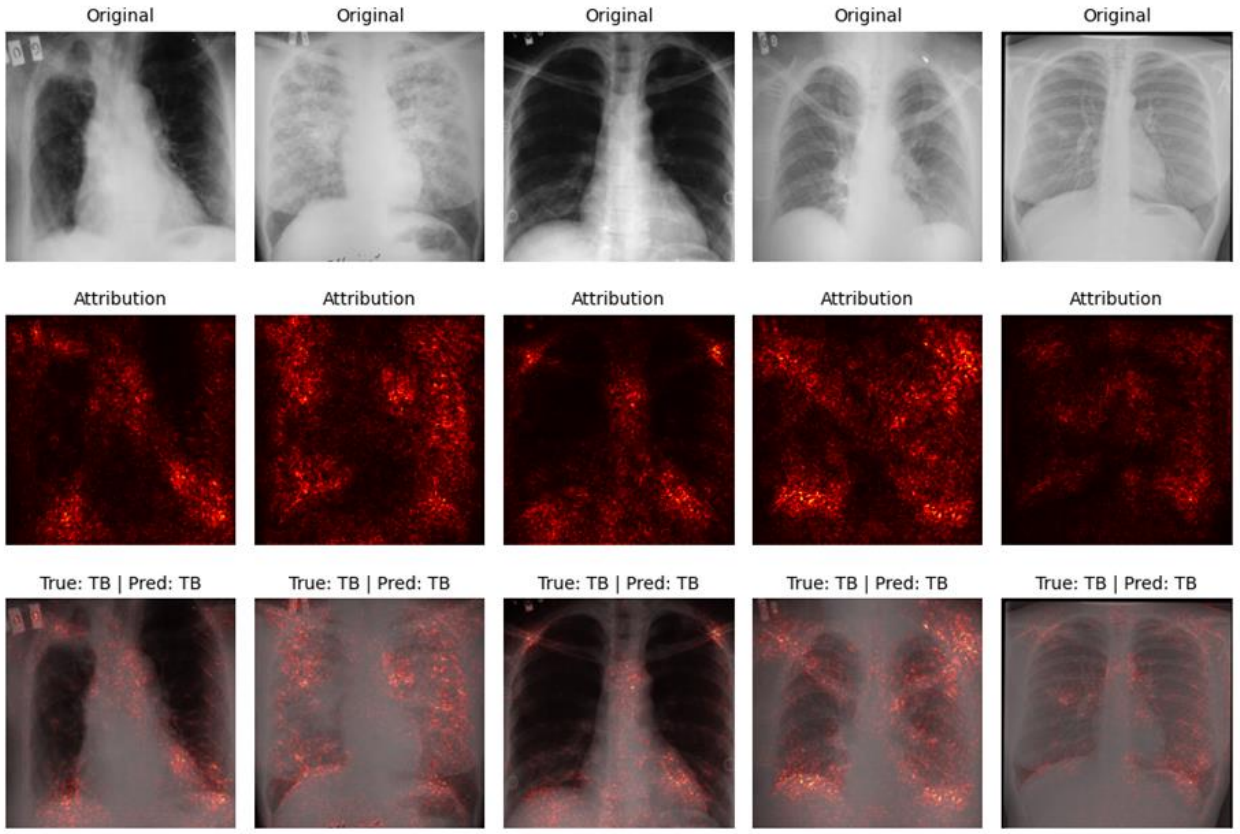


Figure 6. Visualizing model decisions in TB detection

Figure 6 showcases five sets of chest X-rays, each consisting of an original image, an attribution map, and a combined image displaying the true and predicted labels. All five examples consistently show true predictions, indicating that the model correctly identified TB in all these cases. The attribution maps, highlight the specific regions within the X-ray images that the

model focused on when making its TB prediction. These maps reveal areas of high importance, often corresponding to lung fields, where visual anomalies indicative of TB is likely present. The overlaid images in the bottom row further emphasize these critical regions by visually combining the original X-ray with the attribution heatmaps, providing a clear visual representation of what the model deemed most relevant for its accurate diagnosis of TB. This consistent and strong attribution to the lung areas across all five samples suggests that the model effectively learned to identify key pathological features associated with TB.

5. Conclusion

This study presented an interpretable deep learning approach for the automated detection of TB using chest X-ray images. By leveraging a CNN architecture alongside robust data augmentation techniques, the proposed system demonstrates strong performance in distinguishing between TB and normal cases. Beyond predictive accuracy, this work also emphasizes model transparency by integrating the integrated gradient technique to provide visual explanations for the network's decisions. This is especially critical in medical imaging applications, where clinical trust and accountability are essential. Integrated gradient visualizations enabled clear localization of abnormal regions typically associated with TB pathology, offering clinicians an interpretable layer that supports diagnostic confidence and validation. The model's strong performance and explainability highlight the practical potential of deep learning for TB screening, especially in settings with limited radiology resources. However, challenges like dataset variability and the need for broader validation remain. This study underscores the importance of combining accurate AI with explainable methods to improve early diagnosis and foster trust, advancing transparent and equitable AI-driven healthcare.

References

[1] Gill, C. M., Dolan, L., Piggott, L. M., & McLaughlin, A. M. (2022). New developments in tuberculosis diagnosis and treatment. *Breathe*, *18*(1).

[2] Yayan, J., Franke, K. J., Berger, M., Windisch, W., & Rasche, K. (2024). Early detection of tuberculosis: a systematic review. Pneumonia, 16(1), 11.

[3] Bartolomeu-Gonçalves, G., Marinho, J., Fernandes, B. T., Almeida, F., Ferreira Correia, G., Madeira, I., ... & Yamada-Ogatta, S. F. (2024). *Tuberculosis diagnosis: Current, ongoing, and future approaches. Diseases, 12 (9), 202–202*.

[4] Swalehe, H. M., & Obeagu, E. I. (2024). Tuberculosis: Current diagnosis and management. *Elite Journal of Public Health*, *2*(1), 23-33.

[5] Shrivas, A., & Singh, S. (2025). Tuberculosis diagnosis and management: recent advances. *Journal of Global Infectious Diseases*, *17*(1), 3-9.

[6] Albuquerque, C., Henriques, R., & Castelli, M. (2025). Deep learning-based object detection algorithms in medical imaging: Systematic review. *Heliyon*, *11*(1).

[7] Puttagunta, M. K., & Ravi, S. (2021, February). Detection of Tuberculosis based on Deep Learning based methods. In *Journal of Physics: Conference Series* (Vol. 1767, No. 1, p. 012004). IOP Publishing.

[8] Kotei, E., & Thirunavukarasu, R. (2024). A comprehensive review on advancement in deep learning techniques for automatic detection of tuberculosis from chest X-ray images. *Archives of Computational Methods in Engineering*, *31*(1), 455-474.

[9] Mirugwe, A., Tamale, L., & Nyirenda, J. (2025). Improving Tuberculosis Detection in Chest X-Ray Images Through Transfer Learning and Deep Learning: Comparative Study of Convolutional Neural Network Architectures. JMIRx Med, 6, e66029.

[10] Sharma, V., Gupta, S. K., & Shukla, K. K. (2024). Deep learning models for tuberculosis detection and infected region visualization in chest X-ray images. *Intelligent Medicine*, *4*(2), 104-113.

[11] Rahman, T., Khandakar, A., Kadir, M. A., Islam, K. R., Islam, K. F., Mazhar, R., ... & Chowdhury, M. E. (2020). Reliable tuberculosis detection using chest X-ray with deep learning, segmentation and visualization. *Ieee Access*, *8*, 191586-191601.

[12] Noviandy, T. R., Idroes, G. M., Zulfikar, T., & Idroes, R. (2025). Explainable Deep Learning with Lightweight CNNs for Tuberculosis Classification. *Infolitika Journal of Data Science*, *3*(1), 21-30.

[13] Maheswari, B. U., Sam, D., Mittal, N., Sharma, A., Kaur, S., Askar, S. S., & Abouhawwash, M. (2024). Explainable deep-neural-network supported scheme for tuberculosis detection from chest radiographs. *BMC Medical Imaging*, *24*(1), 32.

[14] Özkurt, C. (2024). Improving tuberculosis diagnosis using explainable artificial intelligence in medical imaging. *Journal of Mathematical Sciences and Modelling*, *7*(1), 33-44.

[15] Kiran, S., & Jabeen, I. (2024). Dataset of tuberculosis chest x-rays images. *Mendeley Data*, *2*.

[16] Tuberculosis (TB) Chest X-ray Database, https://www.kaggle.com/datasets/tawsifurrahman/tuberculosis-tb-chest-xray-dataset

[17] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

[18] Sundararajan, M., Taly, A., & Yan, Q. (2017, July). Axiomatic attribution for deep networks. In *International conference on machine learning* (pp. 3319-3328). PMLR.