Vol. 8, No. 3, June 2026

Journal of Information Systems and Informatics

Published By Asosiasi Doktor Sistem Informasi Indonesia

Lung X-ray Image Classification for Distinguishing Tuberculosis and Pneumonia Using Pretrained CNN Feature Extractors and Supervised Classifiers

Ardian Mohib[1]*, Imam Yuadi[2], Ira Puspitasari[3], Yusi Dyah Patriani[4]

[1]Master's Program Human Resource Development-Data Analytics, Postgraduate School, [2]Department of Information and Library Science, Faculty of Social and Political Sciences, [3]Information System Study Program, Universitas Airlangga, Surabaya, Indonesia

[4]Department of Internal Medicine, Faculty of Medicine, Diponegoro University, Semarang, Indonesia.

Received:
October 4, 2025
Revised:
April 19, 2026
Accepted:
May 30, 2026
Published:
June 22, 2026

Corresponding Author:
Author Name*:
Ardian Mohib
Email*:
ardian.mohib-2025@pasca.unair.ac.id

DOI:
10.63158/journalisi.v8i3.1595

Abstract. Tuberculosis (TB) and pneumonia (PNA) are infectious lung diseases with overlapping chest X-ray (CXR) manifestations, making automated differential classification clinically important and methodologically challenging. This study proposes a supervised CXR classification workflow to distinguish TB from PNA using pretrained convolutional neural network (CNN) feature extractors and supervised classifiers. A publicly available de-identified dataset comprising 390 TB and 390 PNA images was used. Images were screened to exclude duplicates, corrupted files, non-CXR images, unclear labels, and identifiable cases. Preprocessing included format standardization, resizing according to CNN input requirements, and normalization. To reduce augmentation-based leakage risk, no heavy pre-validation augmentation was applied. Image embeddings were extracted using VGG-16, Inception V3, and VGG-19, then classified using Logistic Regression, Support Vector Machine, and Neural Network models. Performance was evaluated using stratified 5-fold cross-validation with AUC, accuracy, F1-score, precision, recall, MCC, and confusion matrix analysis. The Inception V3–Logistic Regression combination achieved the best performance, with AUC of 0.999, accuracy of 0.992, F1-score of 0.992, and MCC of 0.985.

Keywords: tuberculosis; pneumonia; chest radiography; transfer learning; CNN feature extraction; supervised classification

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

1. INTRODUCTION

Tuberculosis (TB) and pneumonia (PNA) remain major infectious lung diseases with substantial clinical and public health significance [1], [2]. TB is a chronic infectious disease that requires accurate diagnosis, prolonged treatment, and infection-control measures, whereas pneumonia is an acute lower respiratory tract infection that can progress rapidly if not diagnosed and treated appropriately [3], [4]. Although these two diseases differ in etiology, treatment strategy, and clinical management, both may present with overlapping respiratory symptoms and abnormal radiographic findings on chest X-ray images [2], [5]. This overlap can complicate early differentiation, particularly in settings where radiological expertise is limited or diagnostic workload is high [2].

Chest X-ray (CXR) imaging is widely used as an initial diagnostic modality for pulmonary disease assessment because it is relatively accessible, cost-effective, and clinically informative [6], [7]. However, distinguishing TB from PNA using CXR images remains challenging because both diseases can produce abnormal lung patterns, including infiltrates, opacities, consolidations, or other parenchymal changes [2], [5]. Unlike normal-versus-abnormal classification, TB-versus-PNA classification requires differentiation between two pathological conditions with potentially overlapping radiographic manifestations [8]. Therefore, automated image classification may provide useful decision-support information, although it should not be interpreted as a replacement for clinical judgment [9].

Recent developments in artificial intelligence, especially convolutional neural networks (CNNs) and transfer learning, have contributed to the growth of computer-aided diagnosis systems for chest radiography [8], [10]. Previous studies have applied deep learning to detect tuberculosis [6], [11], pneumonia [12], [13], COVID-19 [14], [15], and other thoracic abnormalities from CXR images [16], [17]. The availability of public CXR datasets has also supported reproducible research on pulmonary disease classification, including Kermany et al.'s pneumonia chest X-ray dataset and the public tuberculosis chest X-ray datasets reported by Jaeger et al. [18], [19]. These datasets provide de-identified images that can be used to evaluate image-based classification workflows without direct access to identifiable hospital records [18], [19].

Despite these advances, many prior CXR classification studies have focused on normal-versus-diseased classification [10], single-disease detection [20], or COVID-19-related classification tasks [14], [15]. Although such studies are clinically valuable, they do not fully address the narrower differential classification problem of distinguishing TB from PNA, where both classes represent abnormal infectious lung conditions [2]. This distinction is clinically relevant because TB and PNA require different treatment pathways, follow-up strategies, and infection-control considerations [3], [4]. Therefore, a focused classification workflow for TB-versus-PNA differentiation remains important [2].

Another methodological issue concerns the use of deep learning under moderate-sized public datasets [21]. End-to-end CNN training generally requires large and well-curated datasets, which are not always available in medical imaging [21]. In this context, pretrained CNN feature extraction offers a practical alternative by using established CNN architectures to generate image embeddings, which can then be classified using supervised learning algorithms [22], [23]. This approach reduces the need for training a deep CNN from scratch while enabling systematic comparison of feature extractors and classifiers under the same validation setting [24], [25].

Based on this gap, this study develops and evaluates a supervised CXR image classification workflow for distinguishing TB from PNA using pretrained CNN feature extractors and supervised classifiers. The novelty of this study is its focused evaluation of TB-versus-PNA differential classification using publicly available de-identified CXR datasets, rather than general normal-versus-abnormal classification. Image embeddings were generated using VGG-16, VGG-19, and Inception V3, and were then classified using Logistic Regression, Support Vector Machine, and Neural Network classifiers within the same stratified cross-validation setting. This study provides a reproducible comparison of pretrained CNN-derived embeddings for distinguishing two abnormal infectious lung conditions with potentially overlapping radiographic features.

2. METHODS

This study proposes a supervised chest X-ray (CXR) image classification workflow for distinguishing tuberculosis (TB) from pneumonia (PNA) using pretrained CNN-based feature extractors and supervised learning classifiers. As shown in Figure 1, the proposed

workflow includes six main stages: public dataset acquisition, data screening and label verification, image preprocessing, image embedding or feature extraction, supervised classification, and validation-based result analysis.

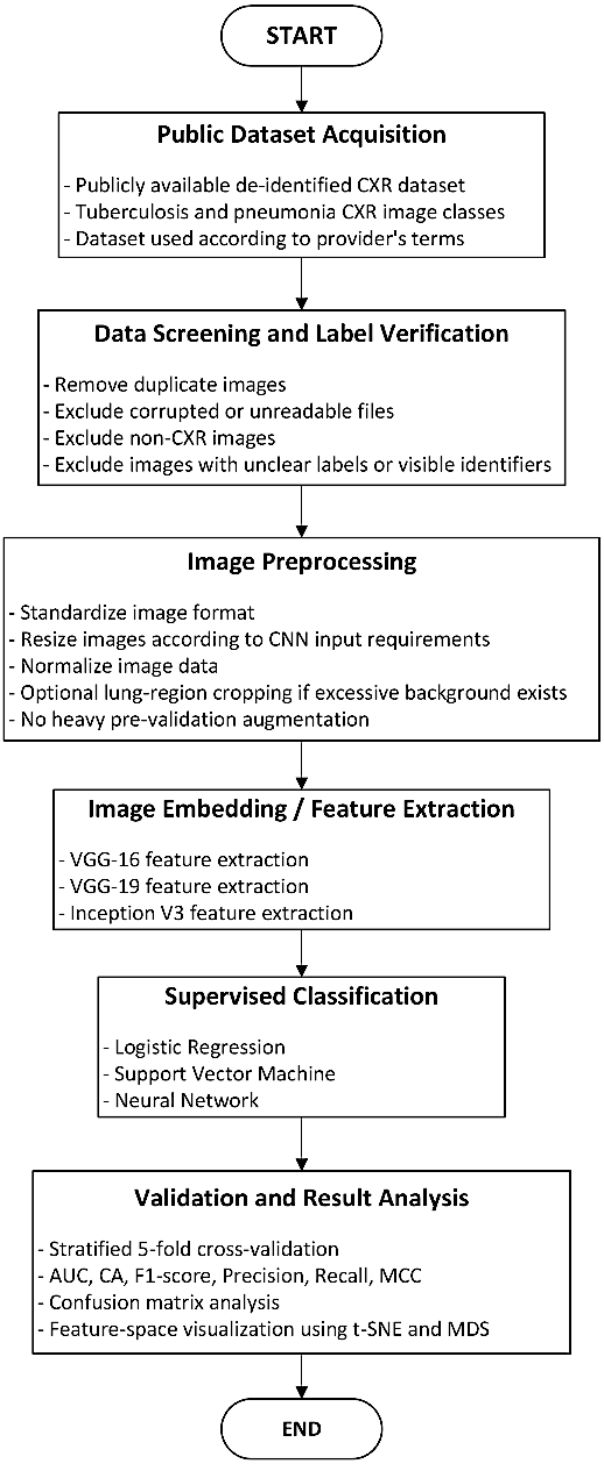

Figure 1. Flowchart of the proposed research workflow.

Unlike an end-to-end CNN training approach, this study used pretrained CNN architectures only as feature extractors [22]. The extracted image embeddings were subsequently classified using supervised machine-learning models [26]. This design was selected to evaluate whether pretrained CNN-derived representations could provide discriminative features for TB-versus-PNA classification while maintaining a reproducible and computationally efficient workflow [25].

2.1. Public Dataset Acquisition and Ethical/Data-Use Consideration

This study used de-identified CXR datasets that are publicly accessible [18], [19]. Pneumonia CXR images were obtained from the publicly available dataset by Kermany et al. [18], whereas tuberculosis CXR images were obtained from the public CXR datasets described by Jaeger et al. [19]. From these datasets, 390 pneumonia images and 390 tuberculosis images were selected, resulting in a balanced dataset of 780 CXR images. To improve the clarity of the dataset composition, preprocessing procedures, and validation strategy, Table 1 provides a compact summary of the dataset used in this study.

Table 1. Summary of dataset, preprocessing, and validation strategy

Dataset source	Class	Number of images	Preprocessing	Validation strategy
Kermany et al. public CXR dataset [18]	Pneumonia	390	Image screening, format standardization, resizing according to CNN input requirements, normalization, and lung-region cropping when necessary	Image-level stratified 5-fold cross-validation
Jaeger et al. public CXR dataset [19]	Tuberculosis	390	Image screening, format standardization, resizing according to CNN input requirements,	Image-level stratified 5-fold cross-validation

Dataset source	Class	Number of images	Preprocessing	Validation strategy
			normalization, and lung-region cropping when necessary	
Combined dataset	TB vs PNA	780	Exclusion of duplicate, corrupted, unreadable, non-CXR, unclear-label, or identifiable images before analysis	Image-level stratified 5-fold cross-validation

Since the datasets were publicly accessible and de-identified, patient-level identifiers were not provided. Therefore, although duplicate, corrupted, unreadable, non-CXR, unclear-label, and identifiable images were screened and excluded before analysis, the possibility of repeated patient cases within the original public dataset sources could not be fully verified. For this reason, patient-level independence across cross-validation folds could not be fully confirmed. The validation strategy in this study should therefore be interpreted as image-level stratified cross-validation rather than patient-level validation.

The use of publicly available datasets was intended to improve reproducibility, transparency, and data-use clarity [18], [19]. Since the images were obtained from de-identified public repositories, this study did not involve direct patient contact, clinical intervention, or access to identifiable clinical records [18], [19]. The datasets were used according to the terms specified by the dataset providers.

2.2. Data Screening and Label Verification

Before analysis, all images were screened to ensure data quality and label consistency. The images were organized into two diagnostic classes: TB and PNA, and the original class labels provided by the dataset sources were retained. The screening procedure included removing duplicate images and excluding corrupted or unreadable files, non-CXR images, images with unclear labels, and images containing visible patient-identifiable information.

This step was performed to ensure that only relevant, readable, and properly labeled CXR images were included in the classification workflow. The use of a balanced public de-identified dataset reduced dependence on extensive image augmentation and improved the reproducibility and transparency of the classification workflow. This dataset design also provided clearer ethical and data-use conditions because the images were obtained from publicly available de-identified repositories [18], [19].

2.3. Image Preprocessing

Image preprocessing was conducted to standardize the input images before CNN-based feature extraction. The preprocessing stage included image format standardization, resizing according to CNN input requirements, normalization, and lung-region cropping when necessary. Lung-region cropping was performed as a region-of-interest (ROI) preprocessing step to expose the lung fields and reduce excessive non-thoracic background. This procedure was not intended as automated pixel-level lung segmentation, but rather as a standardized cropping process to focus the subsequent feature extraction on the clinically relevant thoracic area [27]. Figure 2 shows representative examples of original CXR images and their corresponding cropped lung-region images. The same cropping procedure was applied to all remaining TB and PNA images.

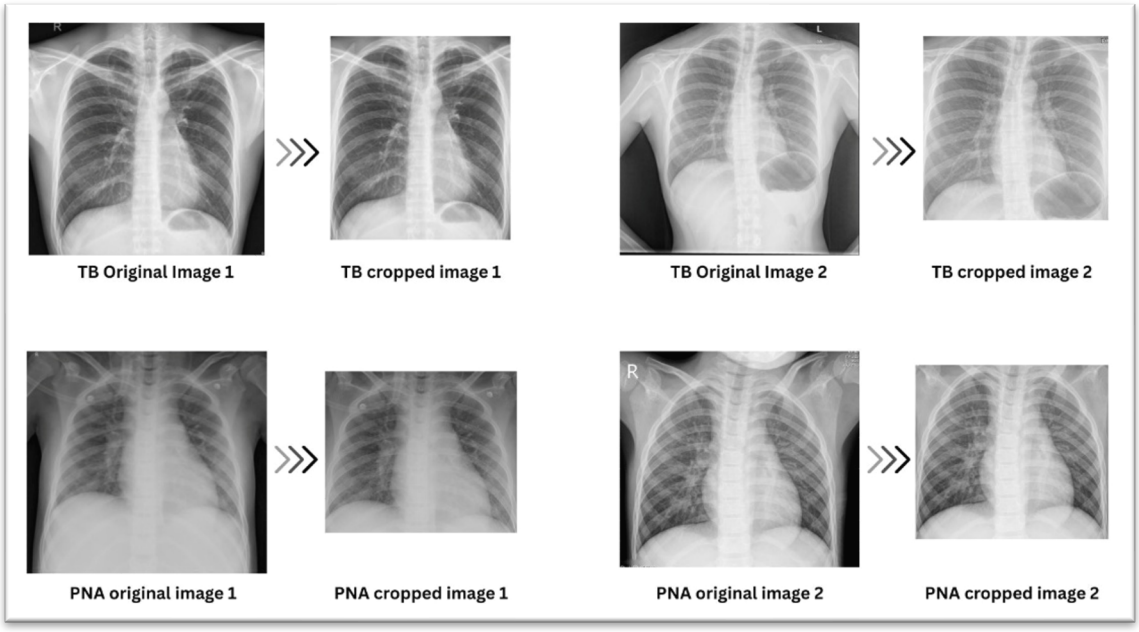

Figure 2. Representative examples of original and cropped CXR images for TB and PNA cases.

No heavy pre-validation augmentation was applied in this study. This decision was made because the dataset already contained a sufficient number of public CXR images for model evaluation [21]. Avoiding extensive augmentation before validation also minimized the risk of augmentation-based leakage, in which artificial variants of the same image could appear across different training and testing folds.

2.4. Image Embedding / Feature Extraction

Image embedding was performed using pretrained CNN models in Orange Data Mining 3.40.0. Pretrained CNN-based feature extraction was selected because transfer learning has been widely used in CXR classification tasks, particularly when training an end-to-end CNN model from scratch is not the main objective or when dataset size is moderate [23], [26]. This study evaluated three pretrained CNN architectures as feature extractors: VGG-16, VGG-19, and Inception V3 [22], [23]. The same input dataset was independently processed using each feature extractor. Each pretrained CNN transformed the CXR images into numerical embedding representations [22]. These embeddings were then used as input features for supervised classification. Using an identical dataset for all feature extractors enabled a fair comparison among VGG-16, VGG-19, and Inception V3 under the same classification and validation settings.

2.5. Supervised Classification

The CNN-derived embeddings were classified using three supervised classifiers: Logistic Regression, Support Vector Machine, and Neural Network. This feature-extraction-based strategy is consistent with previous CXR classification studies that used pretrained CNN representations or transfer learning to support downstream classification tasks [22], [26]. All classifiers were implemented in Orange Data Mining 3.40.0. Unless otherwise specified, the default parameter settings in Orange were used. These classifiers were selected to compare linear, margin-based, and neural-network-based supervised learning approaches using the same pretrained CNN-derived feature representations.

2.6. Validation and Result Analysis

Model evaluation was conducted using stratified 5-fold cross-validation through the Test & Score module in Orange Data Mining 3.40.0. Stratification was applied to preserve the class distribution of TB and PNA images across folds. During each validation round, four folds were assigned for training, while the remaining fold was used for testing. This

procedure continued until all folds had served once as the test fold. The evaluation metrics included Area Under the Receiver Operating Characteristic Curve (AUC), Classification Accuracy (CA), F1-score, Precision, Recall, and Matthews Correlation Coefficient (MCC). These metrics were selected because they are commonly used in CXR-based classification studies to evaluate discrimination ability, prediction correctness, class-level balance, and robustness of binary classification performance [3], [7].

A confusion matrix was used to examine class-level prediction outcomes between TB and PNA. This analysis was used to identify true positive, true negative, false positive, and false negative predictions. In addition, feature-space visualization using t-distributed stochastic neighbor embedding (t-SNE) and multidimensional scaling (MDS) was applied as a supplementary analysis to examine the separability of CNN-derived embeddings, consistent with visualization-based interpretation in previous CXR studies [26], [27]. These visualization techniques were not treated as primary classification metrics, but as supporting tools to interpret the distribution of feature representations.

3. RESULTS AND DISCUSSION

3.1. Performance Evaluation

Table 2 presents the performance results of three pretrained CNN feature extractors, namely VGG-16, Inception V3, and VGG-19, when paired with three supervised classifiers: Logistic Regression (LR), Neural Network (NN), and Support Vector Machine (SVM). The models were tested using stratified 5-fold cross-validation. The evaluation included Area Under the Receiver Operating Characteristic Curve (AUC), Classification Accuracy (CA), F1-score, Precision, Recall, and Matthews Correlation Coefficient (MCC). These measures were used to examine discrimination capability, prediction accuracy, class-level balance, and robustness in binary TB-versus-PNA classification.

Table 2. Logistic Regression (LR), Neural Network, and Support Vector Machine (SVM) Results

Classifier	Feature Extractor	AUC	CA	F1	Prec	Recall	MCC
LR	Inception V3	0.999	0.992	0.992	0.992	0.992	0.985
	VGG-16	0.997	0.976	0.976	0.976	0.976	0.951

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

Classifier	Feature Extractor	AUC	CA	F1	Prec	Recall	MCC
NN	VGG-19	0.998	0.986	0.986	0.986	0.986	0.972
	Inception V3	0.997	0.990	0.990	0.990	0.990	0.980
	VGG-16	0.998	0.978	0.978	0.978	0.978	0.956
	VGG-19	0.997	0.985	0.985	0.985	0.985	0.969
SVM	Inception V3	0.998	0.987	0.987	0.987	0.987	0.974
	VGG-16	0.998	0.977	0.977	0.977	0.977	0.954
	VGG-19	0.996	0.974	0.974	0.975	0.974	0.949

The classification results in Table 2 show that all combinations of pretrained CNN feature extractors and supervised classifiers achieved high performance in distinguishing TB from PNA CXR images. Among all evaluated combinations, the Logistic Regression model using Inception V3 embeddings produced the strongest overall result, with an AUC of 0.999, CA of 0.992, F1-score of 0.992, Precision of 0.992, Recall of 0.992, and MCC of 0.985. The next highest result was obtained by the Inception V3 and Neural Network combination, with an AUC of 0.997, CA of 0.990, F1-score of 0.990, Precision of 0.990, Recall of 0.990, and MCC of 0.980. The Inception V3 and SVM combination also showed high performance, with an AUC of 0.998, CA of 0.987, F1-score of 0.987, Precision of 0.987, Recall of 0.987, and MCC of 0.974. These results suggest that Inception V3 generated the most informative embeddings among the evaluated CNN feature extractors. VGG-19 also demonstrated strong performance. Within the VGG-19 group, Logistic Regression produced the best result, with an AUC of 0.998, CA of 0.986, F1-score of 0.986, Precision of 0.986, Recall of 0.986, and MCC of 0.972. Neural Network with VGG-19 achieved slightly lower but still high performance, while SVM with VGG-19 showed the lowest performance within the VGG-19 group.

For VGG-16, Neural Network achieved the highest result, with an AUC of 0.998, CA of 0.978, F1-score of 0.978, Precision of 0.978, Recall of 0.978, and MCC of 0.956. Logistic Regression and SVM with VGG-16 also produced high classification scores, although their performance was lower than the best Inception V3 and VGG-19 combinations. Overall, the results indicate that pretrained CNN-derived embeddings can provide highly informative feature representations for distinguishing TB and PNA CXR images. However, the high values across all models should be interpreted carefully. Although the dataset did not

rely on heavy pre-validation augmentation, the evaluation was still conducted at the image level using public datasets. Therefore, these findings should be considered strong evidence within the evaluated dataset, but not yet definitive evidence of clinical generalizability. Confusion matrix analysis was performed to examine class-level prediction patterns for each combination of feature extractor and classifier. The confusion matrices are shown in Figure 3–11. In these matrices, each row indicates the actual class, whereas each column indicates the predicted class.

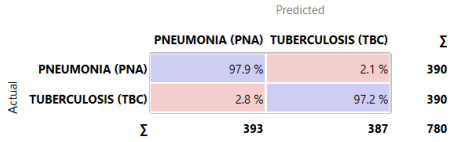

Figure 3. Logistic Regression (LR) using VGG-16

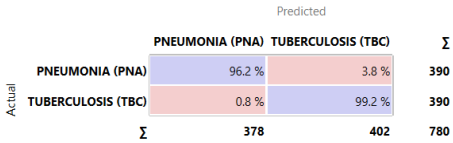

Figure 4. Support Vector Machine (SVM) usingVGG-16

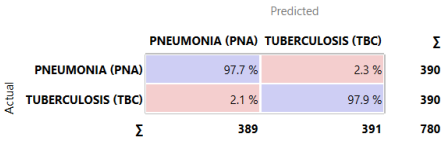

Figure 5. Neural Network (NN) using VGG-16

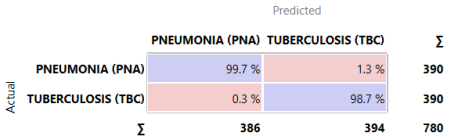

Figure 6. Logistic Regression (LR) using Inception V3

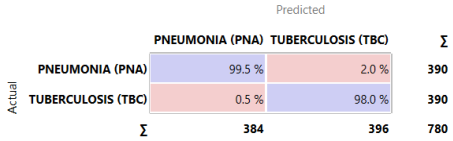

Figure 7. Support Vector Machine (SVM) using Inception V3

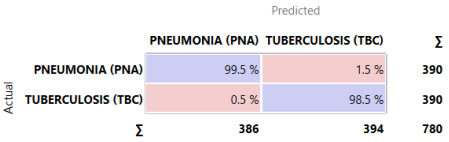

Figure 8. Neural Network (NN) using Inception V3

For VGG-16, the confusion matrices presented in Figure 3–5 indicate that most PNA and TB images were correctly classified by all three classifiers. Logistic Regression with VGG-16 showed balanced performance between the two classes, with low misclassification rates for both PNA and TB. SVM with VGG-16 showed a slightly higher proportion of PNA images predicted as TB, whereas Neural Network with VGG-16 produced a more balanced error distribution. This pattern is consistent with Table 2, where Neural Network achieved the highest CA, F1-score, and MCC among the VGG-16 classifiers.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

For Inception V3, the confusion matrices presented in Figure 6–8 demonstrate the strongest class-level prediction performance. Inception V3 with Logistic Regression produced the lowest overall misclassification pattern, which is consistent with its highest CA, F1-score, and MCC in Table 2. The confusion matrix confirms that the high performance of this model was not driven by correct prediction of only one class, but reflected strong and balanced discrimination between both PNA and TB classes. Inception V3 with Neural Network and SVM also produced very low error patterns, supporting the interpretation that Inception V3 generated highly discriminative embeddings.

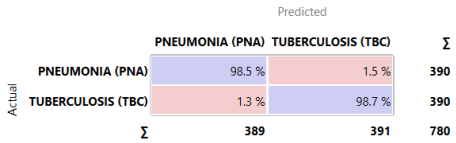

Figure 9. Logistic Regression (LR) using VGG-19

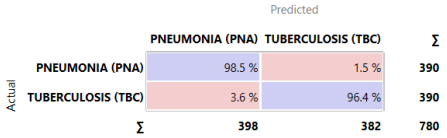

Figure 10. Support Vector Machine (SVM) using VGG-19

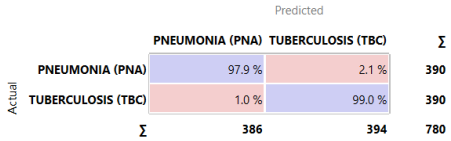

Figure 11. Neural Network (NN) using VGG-19

For VGG-19, the confusion matrices in Figure 9–11 also show strong classification performance. Logistic Regression with VGG-19 demonstrated balanced prediction between PNA and TB, consistent with its strong MCC value. SVM with VGG-19 showed a relatively higher proportion of TB images predicted as PNA compared with the other VGG-19 classifiers, which is consistent with the lower MCC reported in Table 2. Neural Network with VGG-19 showed high TB recognition but a slightly higher PNA-to-TB error compared with Logistic Regression.

Taken together, the confusion matrices support the quantitative results in Table 2. The best-performing models show both high overall performance and balanced class-level prediction. The remaining false positive and false negative cases likely reflect CXR images with overlapping radiographic features between TB and PNA, which is clinically plausible because both diseases may present with infiltrates, opacities, and abnormal lung textures [2], [5].

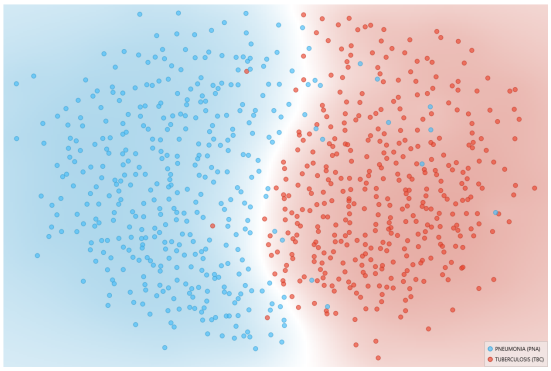

Figure 12. MDS visualization using Inception V3

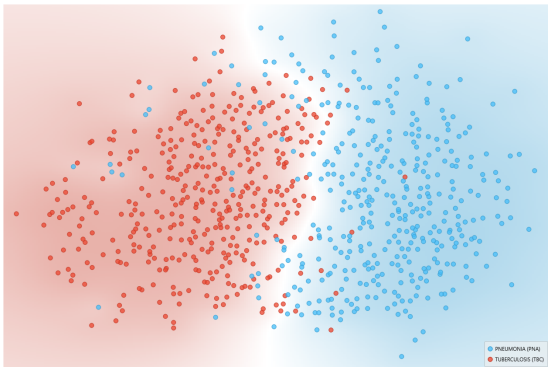

Figure 13. MDS visualization using VGG-16

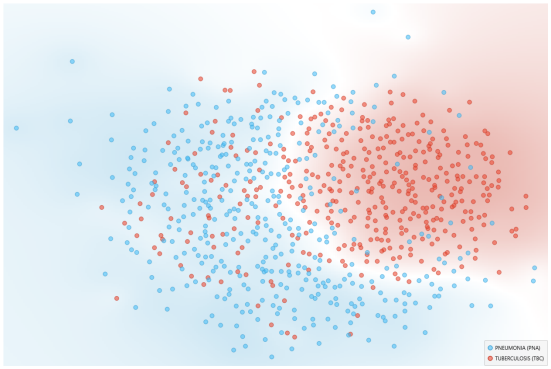

Figure 14. MDS visualization using VGG-19

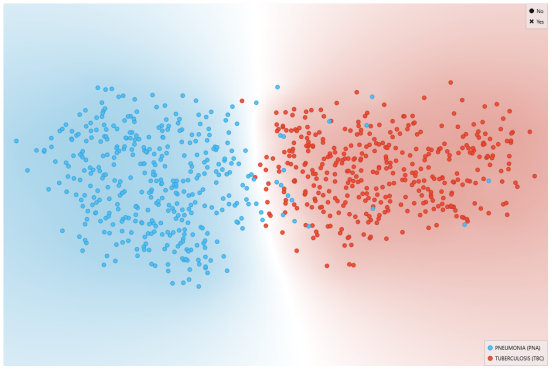

Figure 15. t-SNE visualization using Inception V3

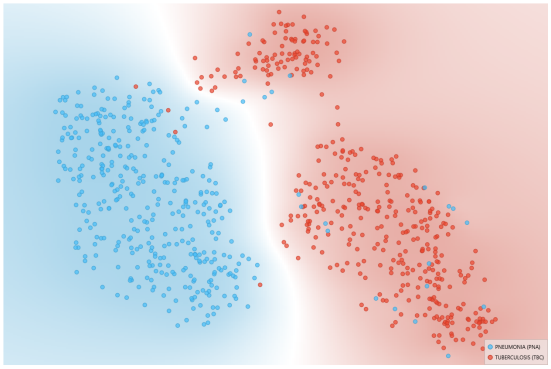

Figure 16. t-SNE visualization using VGG-16

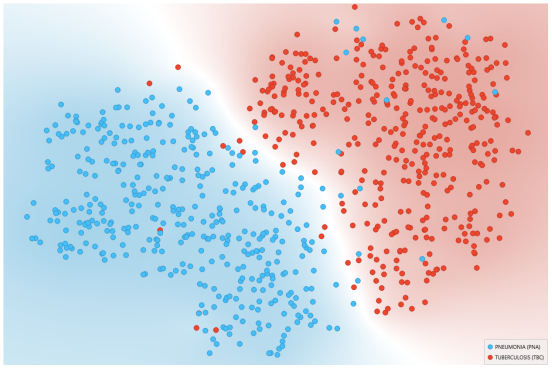

Figure 17. t-SNE visualization using VGG-19

Feature-space visualization was performed using MDS and t-SNE to explore the separability of CNN-derived embeddings. The red points in Figure 12–17 represent TB images, while the blue points represent PNA images. These visualizations were used as supplementary interpretive tools and were not treated as primary classification metrics. Figure 13 shows the MDS result using VGG-16. The distribution indicates partial separation between TB and PNA, but the overlap area appears more visible than in the Inception V3 representation. This may help explain why VGG-16 produced slightly lower CA and MCC

compared with Inception V3. Figure 14 shows the MDS result using VGG-19. The visualization also demonstrates separation between TB and PNA, but with some mixed regions. This aligns with the strong but slightly lower performance of VGG-19 compared with Inception V3.

The t-SNE plots in Figure 15–17 provide additional visualization of local feature-space structure. Figure 15 shows that Inception V3 embeddings produced a clear separation tendency between TB and PNA clusters, consistent with the best classification result obtained by Inception V3 with Logistic Regression. Figure 16 shows that VGG-16 embeddings also formed distinguishable clusters, although some TB and PNA points were located near overlapping regions. Figure 17 shows that VGG-19 embeddings produced a generally separable distribution with some areas of class proximity. Overall, the MDS and t-SNE visualizations support the classification results by showing that CNN-derived embeddings tend to form distinguishable feature distributions between TB and PNA. However, the presence of overlap regions indicates that some images remain visually ambiguous in the embedding space. These overlap regions may correspond to the remaining misclassifications observed in the confusion matrices.

3.2. Discussion

The results suggest that pretrained CNN-derived embeddings combined with supervised classifiers can distinguish TB from PNA CXR images in the evaluated public dataset. Among all model combinations, the combination of Inception V3 and Logistic Regression produced the highest overall performance, with an AUC of 0.999, CA of 0.992, F1-score of 0.992, Precision of 0.992, Recall of 0.992, and MCC of 0.985. The high MCC value is important because MCC considers the balance among all four confusion-matrix outcomes, namely true positives, true negatives, false positives, and false negatives, making it appropriate for evaluating binary classification performance beyond accuracy alone [3], [7].

Compared with previous CXR classification studies, the present study focuses on a more specific differential classification task between TB and PNA, rather than general normal-versus-abnormal classification or single-disease detection. Prior studies have demonstrated the value of deep learning and transfer learning in tuberculosis detection, pneumonia identification, COVID-19 classification, and multicategory thoracic abnormality

detection from CXR images [6], [7], [10]-[17]. In contrast, this study evaluates whether pretrained CNN-derived embeddings can distinguish two abnormal infectious lung conditions with partially overlapping radiographic manifestations. This distinction is important because TB and PNA may both show abnormal lung patterns, making their differentiation more challenging than separating normal and abnormal CXR images.

The strong performance of Logistic Regression should be interpreted in the context of CNN-based feature extraction. Although Logistic Regression is a linear classifier, the classification was not performed on raw CXR images. Instead, pretrained CNN models first transformed the images into numerical embeddings. When pretrained CNN-derived embeddings are highly discriminative, the resulting feature space may become more separable, allowing a linear classifier to perform competitively with more flexible models [22], [26]. Therefore, the stronger result obtained by Logistic Regression indicates that Inception V3 produced feature representations that were sufficiently informative for TB-versus-PNA separation.

The quantitative results in Table 2 are supported by the confusion matrices and feature-space visualizations. The confusion matrices show that the best-performing models produced only a small number of false positive and false negative predictions, indicating balanced class-level discrimination between TB and PNA. Meanwhile, MDS and t-SNE visualizations show that TB and PNA embeddings generally formed distinguishable distributions, particularly for Inception V3. The overlap regions observed in some MDS and t-SNE plots may correspond to visually ambiguous CXR images and help explain the remaining misclassifications in the confusion matrices. This interpretation is clinically plausible because TB and PNA may present with overlapping CXR abnormalities [4], [6].
Although the results are strong, they should be interpreted cautiously. The use of publicly available de-identified datasets and the absence of heavy pre-validation augmentation reduced the risk of augmentation-based leakage [18], [19]. However, the high performance should still be understood within the context of the evaluated public dataset and should not be interpreted as definitive evidence of clinical generalizability.

Several threats to validity need to be considered in interpreting the results of this study. First, the evaluation was performed using image-level stratified cross-validation because patient-level identifiers were not available in the public datasets. Therefore, patient-level

independence across folds could not be fully confirmed. Second, the use of TB and PNA images from different public dataset sources may introduce dataset-source bias, including differences in image acquisition, image resolution, preprocessing history, patient population, and labeling procedures. Third, this study did not use an independent external validation dataset. Therefore, although the proposed workflow achieved high performance within the evaluated dataset, further patient-level and multi-institutional external validation is needed before clinical application.

The findings support the feasibility of pretrained CNN feature extraction combined with supervised classification for TB-versus-PNA CXR classification. However, these results should be considered dataset-specific evidence rather than definitive clinical validation. Future studies should validate the proposed workflow using independent external datasets, patient-level separation, and multi-institutional data to confirm robustness and clinical generalizability.

4. CONCLUSION

This study developed and evaluated a supervised chest X-ray image classification workflow for distinguishing tuberculosis from pneumonia using pretrained CNN feature extractors and supervised classifiers. Using publicly available de-identified datasets consisting of 390 TB and 390 PNA images, the highest performance was obtained by the combination of Inception V3 and Logistic Regression, with an AUC of 0.999, classification accuracy of 0.992, F1-score of 0.992, precision of 0.992, recall of 0.992, and MCC of 0.985. These findings suggest that pretrained CNN-derived embeddings can provide discriminative representations for TB-versus-PNA classification within the evaluated public dataset. However, the results should be interpreted cautiously because the evaluation was conducted at the image level, without available patient-level identifiers, and no independent external validation dataset was used. Therefore, the proposed approach should be considered a promising preliminary classification workflow rather than a clinically ready diagnostic system. Further validation using larger, independent, multi-institutional datasets with patient-level separation is required to confirm robustness and clinical generalizability.

ACKNOWLEDGMENT

The authors acknowledge the availability of the public chest X-ray datasets used in this study and the support of the academic supervisors during the research process.

REFERENCES

[1] N. Oktavia, B. S. Miranda, and D. I. Swasono, "CNN-Based Classification of Infectious Lung Diseases using Thorax X-Ray Analysis," *Engineering and Technology Journal*, vol. 09, no. 10, Oct. 2024, doi: 10.47191/etj/v9i10.14.

[2] S. K. Mohapatra, M. Abebe, L. Mekuanint, S. Prasad, P. K. Bala, and S. K. Dhala, "Pneumonia and tuberculosis detection with chest x-ray images and medical records using deep learning techniques," *Review of Computer Engineering Research*, vol. 10, no. 4, pp. 136–149, Nov. 2023, doi: 10.18488/76.v10i4.3533.

[3] B. U. Maheswari *et al.*, "Explainable deep-neural-network supported scheme for tuberculosis detection from chest radiographs," *BMC Med. Imaging*, vol. 24, no. 1, p. 32, Feb. 2024, doi: 10.1186/s12880-024-01202-x.

[4] K. Guo, J. Cheng, K. Li, L. Wang, Y. Lv, and D. Cao, "Diagnosis and detection of pneumonia using weak-label based on X-ray images: a multi-center study," *BMC Med. Imaging*, vol. 23, no. 1, p. 209, Dec. 2023, doi: 10.1186/s12880-023-01174-4.

[5] L. Venkataramana, D. V. V. Prasad, S. Saraswathi, C. M. Mithumary, R. Karthikeyan, and N. Monika, "Classification of COVID-19 from tuberculosis and pneumonia using deep learning techniques," *Med. Biol. Eng. Comput.*, vol. 60, no. 9, pp. 2681–2691, Sep. 2022, doi: 10.1007/s11517-022-02632-x.

[6] T. Xu and Z. Yuan, "Convolution Neural Network With Coordinate Attention for the Automatic Detection of Pulmonary Tuberculosis Images on Chest X-Rays," *IEEE Access*, vol. 10, pp. 86710–86717, 2022, doi: 10.1109/ACCESS.2022.3199419.

[7] W. Khan, N. Zaki, and L. Ali, "Intelligent Pneumonia Identification From Chest X-Rays: A Systematic Literature Review," *IEEE Access*, vol. 9, pp. 51747–51771, 2021, doi: 10.1109/ACCESS.2021.3069937.

[8] S.-L. Yi, S.-L. Qin, F.-R. She, and T.-W. Wang, "RED-CNN: The Multi-Classification Network for Pulmonary Diseases," *Electronics (Basel).*, vol. 11, no. 18, p. 2896, Sep. 2022, doi: 10.3390/electronics11182896.

[9] Y. Xie *et al.*, "Computer-Aided System for the Detection of Multicategory Pulmonary Tuberculosis in Radiographs," *J. Healthc. Eng.*, vol. 2020, pp. 1–12, Aug. 2020, doi: 10.1155/2020/9205082.

[10] Y.-X. Tang *et al.*, "Automated abnormality classification of chest radiographs using deep convolutional neural networks," *NPJ Digit. Med.*, vol. 3, no. 1, p. 70, May 2020, doi: 10.1038/s41746-020-0273-z.

[11] E. Showkatian, M. Salehi, H. Ghaffari, R. Reiazi, and N. Sadighi, "Deep learning-based automatic detection of tuberculosis disease in chest X-ray images," *Pol. J. Radiol.*, vol. 87, pp. 118–124, Feb. 2022, doi: 10.5114/pjr.2022.113435.

[12] R. Kundu, R. Das, Z. W. Geem, G.-T. Han, and R. Sarkar, "Pneumonia detection in chest X-ray images using an ensemble of deep learning models," *PLoS One*, vol. 16, no. 9, p. e0256630, Sep. 2021, doi: 10.1371/journal.pone.0256630.

[13] T. H. Mandeel, S. M. Awad, and S. Naji, "Pneumonia binary classification using multi-scale feature classification network on chest x-ray images," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 11, no. 4, p. 1469, Dec. 2022, doi: 10.11591/ijai.v11.i4.pp1469-1477.

[14] A. I. Khan, J. L. Shah, and M. M. Bhat, "CoroNet: A deep neural network for detection and diagnosis of COVID-19 from chest x-ray images," *Comput. Methods Programs Biomed.*, vol. 196, p. 105581, Nov. 2020, doi: 10.1016/j.cmpb.2020.105581.

[15] A. Abbas, M. M. Abdelsamea, and M. M. Gaber, "Classification of COVID-19 in chest X-ray images using DeTraC deep convolutional neural network," *Applied Intelligence*, vol. 51, no. 2, pp. 854–864, Feb. 2021, doi: 10.1007/s10489-020-01829-7.

[16] A. Harshavardhan, S. Cheerla, A. Parkavi, S. A. Latha Mary, K. Qureshi, and H. R. Mhaske, "Deep learning modified neural networks with chicken swarm optimization-based lungs disease detection and severity classification," *J. Electron. Imaging*, vol. 32, no. 06, May 2023, doi: 10.1117/1.JEI.32.6.062603.

[17] D. M. Ibrahim, N. M. Elshennawy, and A. M. Sarhan, "Deep-chest: Multi-classification deep learning model for diagnosing COVID-19, pneumonia, and lung cancer chest diseases," *Comput. Biol. Med.*, vol. 132, p. 104348, May 2021, doi: 10.1016/j.compbiomed.2021.104348.

[18] D. Z. K. G. M. Kermany, "Large Dataset of Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images," Mendeley Data.

[19] 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.," *Quant.*

Imaging Med. Surg., vol. 4, no. 6, pp. 475–7, Dec. 2014, doi: 10.3978/j.issn.2223-4292.2014.11.20.

[20] S. Urooj, S. Suchitra, L. Krishnasamy, N. Sharma, and N. Pathak, "Stochastic Learning-Based Artificial Neural Network Model for an Automatic Tuberculosis Detection System Using Chest X-Ray Images," *IEEE Access*, vol. 10, pp. 103632–103643, 2022, doi: 10.1109/ACCESS.2022.3208882.

[21] W. Zhang, H. Wang, Z. Lai, and C. Hou, "Constrained Contrastive Representation: Classification On Chest X-Rays With Limited Data," in *2021 IEEE International Conference on Multimedia and Expo (ICME)*, IEEE, Jul. 2021, pp. 1–6. doi: 10.1109/ICME51207.2021.9428273.

[22] N. Habib, Md. M. Hasan, Md. M. Reza, and M. M. Rahman, "Ensemble of CheXNet and VGG-19 Feature Extractor with Random Forest Classifier for Pediatric Pneumonia Detection," *SN Comput. Sci.*, vol. 1, no. 6, p. 359, Nov. 2020, doi: 10.1007/s42979-020-00373-y.

[23] T. Rahman *et al.*, "Transfer Learning with Deep Convolutional Neural Network (CNN) for Pneumonia Detection Using Chest X-ray," *Applied Sciences*, vol. 10, no. 9, p. 3233, May 2020, doi: 10.3390/app10093233.

[24] O. A. Fagbuagun, O. Nwankwo, S. A. Akinpelu, and O. Folorunsho, "Model development for pneumonia detection from chest radiograph using transfer learning," *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 20, no. 3, p. 544, Jun. 2022, doi: 10.12928/telkomnika.v20i3.23296.

[25] R. Wajgi *et al.*, "Optimized tuberculosis classification system for chest X-ray images: Fusing hyperparameter tuning with transfer learning approaches," *Engineering Reports*, vol. 6, no. 11, Nov. 2024, doi: 10.1002/eng2.12906.

[26] J. Luján-García, C. Yáñez-Márquez, Y. Villuendas-Rey, and O. Camacho-Nieto, "A Transfer Learning Method for Pneumonia Classification and Visualization," *Applied Sciences*, vol. 10, no. 8, p. 2908, Apr. 2020, doi: 10.3390/app10082908.

[27] 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, doi: 10.1109/ACCESS.2020.3031384.