

Multiclass Chest X-Ray Classification Using Transfer Learning with Resnet50: Performance Evaluation Across Varying Train-Test Configurations

Joshua Kolapo Awe^{1,*}, Samson Afolabi Arekete¹, Olufade Falade Williams Onifade²,
Olufunmilayo Olopade^{3,4}, Benjamin Segun Aribisala⁵

¹ Department of Computer Science,
Redeemer's University, Ede, Nigeria

² Department of Computer Science,
University of Ibadan, Ibadan, Nigeria

³ Centre for Clinical Cancer Genetics and Global Health,
University of Chicago, Chicago, USA

⁴ Department of Medicine, University of Chicago, Chicago, USA

⁵ Department of Computer Science, Lagos State University, Ojo, Lagos, Nigeria

* Corresponding Author Email: awejoshuakolapo1@gmail.com
<https://doi.org/10.56201/ijcsmt.vol.12.no9.2026.pg57.64>

Abstract

Chest X-ray (CXR) imaging remains the most widely used diagnostic modality for chest X-ray abnormalities globally, yet its interpretation is heavily dependent on radiologist availability and is prone to inter-observer variability, particularly in low- and middle-income countries where specialist shortages delay diagnosis. While deep learning models have achieved radiologist-level accuracy in classifying chest X-ray abnormalities, many studies report performance on a single, fixed train-test split, leaving open the question of how the proportion of data allocated to training versus testing affects performance under severe class imbalance. This study developed and evaluated a multiclass chest X-ray classification model using transfer learning with ResNet50, and systematically compared its performance across four train-test split configurations. A master dataset of 93,853 chest X-ray images was assembled from four public sources (Chest X-ray Dataset for Tuberculosis Segmentation, Chest X-Ray Images (Pneumonia), NIH Chest X-rays, COVID-19 Radiography Database) and harmonised into eight diagnostic classes: COVID-19, Pneumonia, Tuberculosis, Pleural Effusion, Cardiomegaly, Atelectasis, Pneumothorax, and Normal. Following image preprocessing and class-weighted loss adjustment for dataset imbalance, a ResNet50-based convolutional neural network was trained using transfer learning. Four train-test split configurations (20/80, 40/60, 50/50, and 70/30) were compared, and the 20/80 split yielded the best overall performance, with 71.37% accuracy, 83.59% weighted precision, 71.37% weighted recall, and 75.72% weighted F1-score, alongside a macro-average ROC-AUC of 0.9020. The model performed strongly on Pneumonia and COVID-19 (F1-scores of 93% and 89%, respectively) but showed reduced precision for minority classes such as Atelectasis, Cardiomegaly, and Pneumothorax, largely attributable to systematic misclassification of Normal images rather than confusion among disease classes. These findings show that increasing the proportion of training data does not automatically improve generalisation under severe class imbalance, and highlight class imbalance as a key target for future model refinement.

Keywords: *Chest X-ray, Multiclass Classification, Deep Learning, Transfer Learning, ResNet50, Class Imbalance.*

Words count :289

Introduction

Chest X-ray (CXR) imaging is the most widely used medical imaging modality worldwide, valued for its low cost, speed of acquisition, and low radiation exposure, and it supports the diagnosis of cardiopulmonary conditions including pneumonia, tuberculosis, cardiomegaly, pleural effusion, and pneumothorax (Rajpurkar et al., 2020; Santosh, 2022). In many low- and middle-income countries (LMICs), CXR is often the only imaging modality accessible to clinicians managing patients with respiratory or cardiac symptoms, owing to limited access to CT or MRI (Lundervold & Lundervold, 2019). Interpretation, however, remains heavily dependent on radiologist availability and is prone to inter-observer variability, and workforce shortages in many settings delay diagnosis (Brady, 2017; Recht et al., 2020). Deep learning models, particularly convolutional neural networks (CNNs) such as ResNet50, have achieved classification accuracy on chest radiographs comparable to that of radiologists (Rajpurkar et al., 2020; Chen et al., 2020) and have performed strongly on tasks such as pneumonia and COVID-19 detection (Gidde et al., 2021; Blake et al., 2023). However, many studies report performance on a single, fixed train-test split, leaving open the question of how the proportion of data allocated to training versus testing affects classification performance and generalisation, particularly under the severe class imbalance that characterises real-world, multi-source chest X-ray datasets. This study addresses this gap by developing and evaluating a multiclass deep learning-based chest X-ray classification model and systematically comparing its performance across varying train-test split configurations. COVID-19 and Tuberculosis were treated as the primary target conditions because of their global disease burden and well-documented detectability on CXR; the remaining six categories (Pneumonia, Pleural Effusion, Cardiomegaly, Atelectasis, Pneumothorax, and Normal) were included because their radiographic appearances may overlap with or mimic the primary conditions, thereby providing a more clinically realistic and challenging multiclass classification setting.

The specific objectives of the study were to: (i) develop a deep learning model capable of classifying chest X-ray images into the eight target diagnostic categories; (ii) compare model performance across four train-test split configurations (20/80, 40/60, 50/50, and 70/30) to evaluate the effect of training data proportion on classification performance under class imbalance; and (iii) evaluate the performance of the resulting model using standard classification and discrimination metrics.

Methodology

This study adopted a retrospective experimental design using secondary, publicly available chest X-ray data; no primary or hospital-based data collection was involved. The overall pipeline comprised dataset assembly and harmonisation, image preprocessing and augmentation, deep learning-based classification via transfer learning, and multi-metric performance evaluation across four train-test split configurations.

Data Sources and Description

A master dataset of 93,853 chest X-ray images was assembled from four public sources — the COVID-19 Radiography Database, NIH ChestX-ray14, the Chest X-ray Images (Pneumonia)

dataset, and Chest X-ray datasets for Tuberculosis segmentation — and harmonised into eight standardised diagnostic classes: Normal, COVID-19, Pneumonia, Tuberculosis, Pleural Effusion, Cardiomegaly, Atelectasis, and Pneumothorax. Because the source datasets used different labelling conventions, a cross-dataset label harmonisation step mapped semantically equivalent labels onto this common eight-class taxonomy. The combined dataset was substantially imbalanced: The Normal class contained 72,495 images, while Tuberculosis — one of the two primary target conditions — contained only 345 images. The remaining classes comprised Pneumonia (5,940), Atelectasis (4,215), Pleural Effusion (3,955), COVID-19 (3,616), Pneumothorax (2,194), and Cardiomegaly (1,093) images. This imbalance motivated the use of class-weighted loss during training.

Data Preprocessing and Augmentation

Preprocessing comprised data cleaning, CLAHE-based contrast enhancement, resizing to 224×224 pixels, conversion to a 3-channel grayscale representation, and normalisation using ImageNet mean and standard deviation values. Data augmentation (flipping, flopping, and rotation) was applied to the training set to increase image diversity and improve generalisation.

Model Architecture and Feature Extraction

A ResNet50 convolutional neural network, pretrained on ImageNet, was used as the feature-extraction backbone. Early convolutional layers were kept frozen to retain general visual representations, while the final residual block (layer4) was fine-tuned on the chest X-ray data. Global Average Pooling reduced the resulting feature maps to a compact feature vector, which was passed to a fully connected classification layer and a class-weighted Softmax output layer covering the eight diagnostic classes. Because the classifier uses a Softmax output, the model produces a single primary prediction — the highest-probability class for each image — rather than simultaneous multi-label identification of coexisting conditions. This is a deliberate scope decision discussed further in the Discussion section.

Evaluation Metrics and Experimental Design

Model performance was assessed using Accuracy, weighted Precision, weighted Recall, weighted F1-score, Specificity, Balanced Accuracy, Confusion Matrix analysis, and one-vs-rest ROC-AUC, considered at both the overall and class levels. Four train-test split configurations — 20/80, 40/60, 50/50, and 70/30 (train/test) — were compared to select the best-performing configuration for detailed evaluation.

Results

The results reported below summarise the performance of the four train-test split configurations and the detailed performance of the selected model.

Data Split Comparison

Table 1 compares the four split configurations. The 20/80 (train/test) split achieved the best overall balance across accuracy, recall, and F1-score (71.37% accuracy, 71.37% weighted recall, 75.72% weighted F1-score), although the 70/30 split achieved marginally higher weighted precision (85.84%). More training data did not translate into better generalisation in this study — the 70/30 configuration, despite using more than three times as much training data as the 20/80

configuration, did not outperform it on accuracy, recall, or F1-score. The 20/80 configuration was therefore selected as the final model for detailed evaluation.

Table 1: Performance comparison of the evaluated train-test split configurations

Split (Train/Test)	Val. Accuracy	Weighted Precision	Weighted Recall	Weighted F1-score
20/80	71.37%	83.59%	71.37%	75.72%
40/60	69.89%	84.63%	69.89%	74.94%
50/50	61.89%	85.57%	61.89%	68.59%
70/30	66.54%	85.84%	66.54%	72.45%

Performance of the Selected (20/80) Model

On the full held-out test set of 75,083 images, the selected 20/80 model achieved an overall accuracy of 71.37%, weighted precision of 83.59%, weighted recall of 71.37%, weighted F1-score of 75.72%, balanced accuracy of 65.48%, and a macro-average ROC-AUC of 0.9020 (Table 2). The relatively high macro-average AUC alongside a moderate overall accuracy indicates that the model retains strong ability to discriminate between the eight classes across decision thresholds, even under substantial class imbalance.

Table 2: Overall performance of the selected 20/80 model

Metric	Value
Accuracy	71.37%
Weighted Precision	83.59%
Weighted Recall	71.37%
Weighted F1-score	75.72%
Balanced Accuracy	65.48%
Macro-average AUC	0.9020
Weighted-average AUC	0.8790

Class-Level Performance and Misclassification Patterns

Class-level performance (Table 3) varied considerably. The model performed strongly on Pneumonia (precision 92%, recall 93%, F1-score 93%) and COVID-19 (precision 82%, recall 98%, F1-score 89%). Performance was markedly weaker for Atelectasis, Cardiomegaly, Pleural Effusion, and Pneumothorax, all with F1-scores between 23% and 35%. Specificity, however, remained high across all classes (0.82–1.00), indicating that the lower precision and F1-scores for these classes were not driven by widespread confusion between disease classes themselves.

Confusion matrix analysis showed that this pattern was driven primarily by a substantial number of true Normal images being misclassified into the minority disease classes — for example, 6,910 Normal images were predicted as Atelectasis and 4,870 as Pleural Effusion — rather than by confusion among the minority classes with each other. Tuberculosis, the smallest class in the dataset, achieved the best separation of the disease classes, with a recall of 88% despite comprising only 345 training images overall.

Table 3: Class-level performance of the selected 20/80 model

Class	Precision	Recall	F1-Score	Specificity
Atelectasis	0.16	0.44	0.23	0.89
COVID-19	0.82	0.98	0.89	0.99
Cardiomegaly	0.18	0.31	0.23	0.98
Normal	0.93	0.72	0.81	0.82
Pleural Effusion	0.25	0.60	0.35	0.92
Pneumonia	0.92	0.93	0.93	0.99
Pneumothorax	0.21	0.37	0.27	0.97
Tuberculosis	0.55	0.88	0.67	1.00

ROC-AUC Analysis

One-vs-rest ROC-AUC analysis (Table 4) confirmed excellent class separability for COVID-19 (AUC = 0.9989) and Tuberculosis (AUC = 0.9984), and good separability for Pneumonia (AUC = 0.9778). Atelectasis (AUC = 0.7845) and Pneumothorax (AUC = 0.8363) showed the lowest, though still reasonable, AUC values. The gap between comparatively high AUC values and low precision at the default decision threshold for several minority classes suggests that classification performance for these classes could potentially be improved through threshold adjustment, without necessarily requiring model retraining.

Table 4: One-vs-rest ROC-AUC scores for the selected 20/80 model

Class	AUC
Atelectasis	0.7845
COVID-19	0.9989
Cardiomegaly	0.8735
Normal	0.8714
Pleural Effusion	0.8752
Pneumonia	0.9778
Pneumothorax	0.8363
Tuberculosis	0.9984

Discussion

The results demonstrate that a ResNet50-based transfer learning model can achieve strong multiclass discrimination on chest X-ray images despite substantial class imbalance in the training data. The strong performance on Pneumonia and COVID-19 (F1-scores of 93% and 89%) is consistent with these classes having larger training samples and, plausibly, more visually distinctive radiographic features. By contrast, the weaker precision observed for Atelectasis, Cardiomegaly, Pleural Effusion, and Pneumothorax was shown by the confusion matrix to stem largely from Normal images being misrouted into these categories, rather than from confusion among the minority classes themselves — indicating that the model struggled to establish an

appropriate decision boundary against the numerically dominant Normal class rather than failing to learn the minority conditions' features outright. This is supported by the comparatively strong macro-average AUC (0.9020) for these same classes, which shows that useful discriminative information remains available even where precision at the default threshold is low.

The comparison across split configurations reinforces that more training data does not automatically yield better generalisation: the 70/30 split, despite the largest training set, was outperformed on accuracy, recall, and F1-score by the 20/80 split. This finding, together with the class-level results, underscores the importance of evaluating imbalanced medical imaging models using multiple complementary metrics rather than accuracy alone.

It is important to interpret these results within the correct scope. ResNet50 does not identify named clinical findings such as consolidation or infiltrates; it learns image representations that are useful for distinguishing between the predefined categories. Similarly, Softmax probability reflects classification confidence, not clinical urgency. The model is intended to support, not replace, clinician judgement, and was evaluated retrospectively using public datasets rather than in live clinical deployment. Several limitations should be acknowledged. The dataset is highly imbalanced (72,495 Normal images versus 345 Tuberculosis images), which was the principal driver of the weaker minority-class precision observed. The four source datasets differ in acquisition equipment, patient populations, and labelling conventions, which may introduce heterogeneity not fully captured by harmonisation. The Softmax-based architecture produces a single primary diagnosis per image and does not model coexisting conditions, even though some of the included categories may clinically coexist with COVID-19 or Tuberculosis. Finally, the model has not yet been externally validated on data from an independent institution.

Future work should prioritise targeted strategies for class imbalance (e.g., focal loss, balanced sampling, or synthetic augmentation for minority classes), class-specific decision-threshold tuning informed by the ROC-AUC results, external validation on independent datasets, comparison against alternative architectures (e.g., EfficientNet, DenseNet, or Vision Transformers), probability calibration, and extension to multi-label classification (e.g., via a Sigmoid output layer) to capture coexisting conditions.

Conclusion

In conclusion, this study shows that a ResNet50-based transfer learning model can achieve strong multiclass chest X-ray classification performance, with an overall accuracy of 71.37% and a macro-average ROC-AUC of 0.9020 on a held-out set of 75,083 images using the 20/80 train-test split, and that increasing the proportion of training data does not automatically improve generalisation under severe class imbalance. The model showed strong performance on the two primary target conditions (COVID-19 and Tuberculosis) and on Pneumonia, and identifiable, threshold-addressable weaknesses on several minority classes driven predominantly by confusion with the Normal class. These findings identify class imbalance and train-test split configuration as key targets for future model refinement.

Data Availability Statement

The chest X-ray images used in this study were drawn entirely from four publicly available, de-identified sources: the COVID-19 Radiography Database, NIH Chest X-rays, the Chest X-ray Images (Pneumonia) dataset, and the Chest X-ray Datasets for Tuberculosis Segmentation. No new patient data were collected, and no personally identifiable information was accessed at any stage of the study. Use of each dataset complied with the respective data providers' terms of use

and licensing conditions for academic and research purposes. Code and additional experimental materials supporting the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest Statement

The authors declare no conflict of interest.

Acknowledgements

The DATA Science and Medical Image Analysis Training for Improved Health CARE Delivery in Nigeria (DATICAN) project is acknowledged for financial support provided through an MSc scholarship, under which this work was undertaken.

Supporting Information

Supporting information for this study, including sample source code excerpts, mathematical formulations of the performance evaluation metrics, detailed dataset source and class-distribution tables, and an extended ethical considerations and data usage statement, is available from the corresponding author upon reasonable request.

References

- Blake, S. R., Das, N., Tadeipalli, M., Reddy, B., Singh, A., Agrawal, R., Chatteraj, S., Shah, D., & Putha, P. (2023). Using artificial intelligence to stratify normal versus abnormal chest X-rays: External validation of a deep learning algorithm at East Kent Hospitals University NHS Foundation Trust. *Diagnostics*, 13(22), 3408. <https://doi.org/10.3390/diagnostics13223408>
- Brady, A. P. (2017). Error and discrepancy in radiology: Inevitable or avoidable? *Insights into Imaging*, 8(1), 171–182.
- Chen, K. C., Yu, H. R., Chen, W. S., Lin, W. C., Lee, Y. C., Chen, H. H., Jiang, J. H., Su, T. Y., Tsai, C. K., Tsai, T. A., Tsai, C. M., & Lu, H. H. S. (2020). Diagnosis of common pulmonary diseases in children by X-ray images and deep learning. *Scientific Reports*, 10, 17374. <https://doi.org/10.1038/s41598-020-73831-5>
- Gidde, P. S., Prakash, C. D., Saini, A., Yadav, M., Sethi, S., & Behera, S. K. (2021). Validation of expert system enhanced deep learning algorithm for automated screening for COVID-pneumonia on chest X-rays. *Scientific Reports*, 11, 23210.
- Lundervold, A. S., & Lundervold, A. (2019). An overview of deep learning in medical imaging focusing on MRI. *Zeitschrift für Medizinische Physik*, 29(2), 102–127.
- Rajpurkar, P., O'Connell, C., Schechter, A., Asnani, N., Li, J., Kiani, A., ... Lungren, M. P. (2020). CheXaid: Deep learning assistance for physician diagnosis of tuberculosis using chest X-rays in patients with HIV. *NPJ Digital Medicine*, 3(1), 115.
- Recht, M. P., Dewey, M., Dreyer, K., Langlotz, C., Niessen, W., Prainsack, B., & Smith, J. J. (2020). Integrating artificial intelligence into the clinical practice of radiology: Challenges and recommendations. *European Radiology*, 30(6), 3576–3584.
- Santosh, K. C. (2022). Advances in deep learning for tuberculosis screening using chest X-rays: The last 5 years review. *Journal of Medical Systems*, 46(11), 82.