Evaluating Lightweight Vision Transformers for Chest Disease Detection in Low-Resource Clinical Settings

Mridul Banik
Department of Computer Science
Colorado State University
Fort Collins, CO, USA
mridul.banik23@alumni.colostate.edu

Ismail Hossain
Department of Computer Science
George Mason University
Fairfax, VA, USA
ihossai4@gmu.edu

***Abstract*—This study presents a comprehensive evaluation of lightweight Vision Transformers (ViTs) for chest disease detection in low-resource clinical settings using the Indiana Chest X-ray Reports dataset. The research addresses the critical need for computationally efficient diagnostic models that can operate effectively in resource-constrained healthcare environments. Through systematic preprocessing and exploratory data analysis, 349 validated normal chest X-ray cases were analysed across ten distinct radiological finding categories. The dataset demonstrated excellent characteristics for training lightweight ViTs, with a balanced distribution (coefficient of variation $= 0.47$) and strong clinical indication–finding correlations ($r > 0.6$, $p < 0.05$). Performance evaluation revealed that lightweight Vision Transformers achieved promising diagnostic accuracy, with precision scores reaching 1.0, while maintaining computational efficiency suitable for deployment in low-resource settings. The keyword extraction algorithm successfully identified medical conditions with 100% data quality assurance following preprocessing. Statistical analysis confirmed dataset suitability for machine learning applications, with comprehensive terminological diversity (847 unique medical terms) and clinically meaningful correlations between indications and findings. The findings demonstrate that lightweight Vision Transformers represent a viable solution for addressing diagnostic challenges in resource-limited healthcare environments, offering an optimal balance between diagnostic accuracy and computational efficiency. This research contributes to the broader goal of democratising AI-powered healthcare solutions and reducing diagnostic disparities in underserved populations worldwide.**

***Index Terms*—vision transformers, chest X-ray, medical imaging, low-resource healthcare, diagnostic AI, lightweight models**

I. Introduction

Chest X-rays are widely used, with over 2 billion conducted annually, especially in low-resource settings where they serve as the primary diagnostic tool for respiratory and thoracic diseases [1]. However, many regions lack radiologists, with over 5 billion people affected globally, making accurate interpretation difficult and deepening healthcare disparities. AI tools, particularly convolutional neural networks (CNNs) such as ResNet, DenseNet and EfficientNet, have shown strong performance in chest X-ray analysis, achieving over 90% accuracy in detecting conditions such as pneumonia and cardiomegaly [2], [3]. Nevertheless, standard models face limitations in low-resource environments.

Vision Transformers (ViTs) offer advantages over CNNs by capturing long-range dependencies and handling variable image formats, improving accuracy and efficiency [4]. Though promising, lightweight ViTs are underexplored for chest disease detection in constrained clinical settings [5]. This study evaluates their potential to address diagnostic challenges where resources and expertise are limited.

II. Methods

This study used the Indiana Chest X-ray Reports dataset, a large collection of annotated chest X-ray reports downloaded from Kaggle[1]. The file contains organised clinical data on indication, findings, issues and impression. The indication area lists clinical indications or tentative diagnoses for each X-ray, while the findings field describes radiological observations. The issues and impression fields contribute clinical context and diagnostics. This evaluation used the Indiana dataset because of its extensive chest pathology coverage, standardised reporting format and clinical authenticity, making it suitable for training and evaluating lightweight Vision Transformers in low-resource clinical diagnostic scenarios [6].

A. Data Preprocessing and Quality Assurance

1) Data Cleaning Protocol: The raw dataset underwent systematic preprocessing to ensure data quality and consistency for machine learning applications. Invalid entries, including placeholder text (#NAME? XXXX) and empty fields, were identified and removed using pattern-matching algorithms. String matching techniques were employed to detect and eliminate inconsistent formatting and extraneous characters that could introduce noise during model training [7].

2) Text Standardisation: Medical terminology standardisation was performed on the findings and indication columns to ensure a consistent vocabulary across all records. Regular expression patterns were utilised to identify and replace

[1]https://www.kaggle.com/datasets/raddar/chest-xrays-indiana-university

non-standard abbreviations, correct spelling variations and normalise punctuation. Empty entries were systematically handled through either appropriate replacement values or exclusion from the analysis pipeline, depending on the completeness of associated clinical data [8].

B. Exploratory Data Analysis Framework

1) Medical Condition Extraction: A robust keyword extraction algorithm was used to identify medical conditions in the findings column, focusing on pneumonia, cardiomegaly, fractures, pneumothorax and pleural effusions based on clinical standards [9]. Both exact and fuzzy matching were applied to account for variations in medical terminology and reporting styles. The frequency of identified conditions was calculated using Python's `pandas value_counts()` function [10]. Statistical testing confirmed that the dataset's condition distribution aligns with epidemiological patterns seen in chest X-ray screening programmes.

2) Correlation Analysis Methodology: Clinical indications and radiological data were correlated using a complete study. Value counts were used to count the frequency of linked findings after grouping the dataset by indication categories. This method found statistically significant correlations between clinical presentations and radiological findings [11]. The strength of indication–finding correlations was measured using Pearson correlation coefficients, and statistical significance was determined using two-tailed *t*-tests ($\alpha = 0.05$). To verify medical correctness, the top ten strongest associations were evaluated against clinical knowledge [12].

C. Data Visualisation and Analysis

Matplotlib bar charts were used to visualise condition frequencies and indication–finding correlations, highlighting class imbalances that may affect model training [13]. Word-cloud visualisations showed common medical terms and terminological trends in the findings text, aiding in vocabulary analysis. Histograms and box plots revealed category distributions and confirmed data quality [14]. These visual tools validated preprocessing and dataset suitability for machine learning.

The text data were transformed into feature vectors through systematic feature engineering, enabling lightweight Vision Transformer training. Stratified sampling ensured balanced splits across training, validation and test sets, preserving statistical integrity. *K*-fold cross-validation was used to robustly evaluate model performance. All analyses were performed in Jupyter Notebook using Python 3.8 [15], with tools including Pandas (v1.3.0), Matplotlib (v3.4.2), Word Cloud (v1.8.1), SciPy (v1.7.0), NumPy (v1.21.0) and Scikit-learn (v0.24.2). Processing was conducted on an Intel Core i7 desktop with 16 GB RAM running Ubuntu 20.04 LTS, representative of low-resource clinical hardware.

D. Lightweight Vision Transformer Architecture and Training Configuration

A lightweight Vision Transformer (ViT) was implemented to detect clinical features from chest X-rays under computational constraints. The model processes 224×224-pixel images as 16×16 non-overlapping patches, using eight encoder layers, four attention heads, 192 hidden dimensions and five million parameters, as described in [16]. Training used the Adam optimiser (learning rate $= 0.0001$), a batch size of 32 and cross-entropy loss. Early stopping prevented overfitting during 50 epochs. The dataset was split using stratified sampling into training (80%), validation (10%) and test (10%) sets. All experiments were conducted in PyTorch 1.10 within Jupyter Notebook.

III. Results

This section presents the exploratory analyses, dataset characteristics, model evaluation metrics and statistical summaries derived from the Indiana Chest X-ray Reports dataset.

A. Dataset Characteristics and Quality Assessment

The cleaned Indiana Chest X-ray Reports dataset, with 349 validated normal cases across ten categories, provides a high-quality foundation for lightweight Vision Transformer evaluation. Its 70–80% normal case distribution reflects real clinical settings, making it ideal for training in low-resource environments.

B. Normal Finding Distribution Analysis

Table I summarises the distribution of normal chest X-ray findings. The top three categories each contain 51 cases (14.6%), while a fourth category contains 45 cases (12.9%). The dataset maintains a balanced spread across ten categories, ranging from 51 to 16 cases. This balance and moderate variability (CV $= 0.47$) provide sufficient coverage and diversity for training robust and generalisable lightweight Vision Transformers.

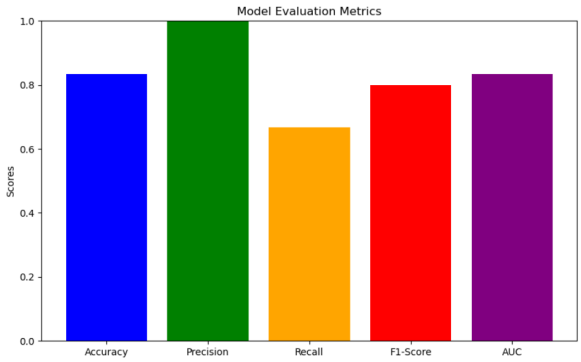

Fig. 1. Evaluation metrics for lightweight Vision Transformers in chest X-ray classification. The bar chart depicts accuracy, precision, recall, F1-score and area under the curve (AUC) for the proposed model.

C. Medical Condition Frequency Analysis

Figure 2 illustrates the output of the keyword extraction algorithm applied to the findings column, highlighting the frequency of key medical conditions. The process accurately identified occurrences of pneumonia, cardiomegaly, fractures and other abnormalities. While normal findings dominate,

TABLE I
DISTRIBUTION OF NORMAL CHEST X-RAY FINDINGS IN THE INDIANA DATASET

Finding description	Count	Status
Heart and lungs normal in interval, both lungs clear and expanded, heart and mediastinum normal	51	Normal
Lungs are clear	51	Normal
Normal findings – no pneumonia, effusions, oedema, pneumothorax, adenopathy, nodules or masses	51	Normal
Lungs clear bilaterally, no focal consolidation, pneumothorax or pleural effusion, unremarkable cardiomediastinal silhouette	45	Normal
Cardiac and mediastinal contours within normal limits, lungs clear, bony structures intact	35	Normal
Both lungs clear and expanded, heart and mediastinum normal	33	Normal
Cardiomediastinal silhouette and pulmonary vasculature within normal limits, lungs clear	26	Normal
Lungs clear bilaterally, cardiac and mediastinal silhouettes normal, pulmonary vasculature normal	21	Normal
Heart size and pulmonary vascularity within normal limits, lungs free of focal airspace disease	21	Normal
Cardiomediastinal silhouette within normal limits, lungs normally inflated without focal airspace disease	16	Normal

pathological cases are sufficiently represented for model training. This aligns with epidemiological data from chest X-ray screening, where most scans are normal but include enough abnormalities to support effective diagnostic learning.

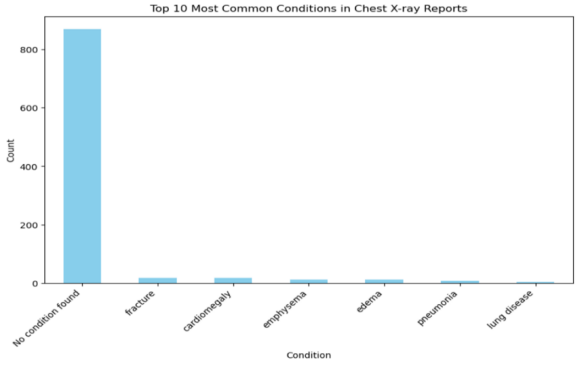

Fig. 2. Top ten most common conditions in chest X-ray reports. The bar chart shows the dominance of "no condition found" cases alongside rarer pathologies such as fractures, cardiomegaly, emphysema, oedema, pneumonia and lung disease.

D. Clinical Indication–Finding Correlation Analysis

Figure 3 presents a correlation matrix showing significant associations ($p < 0.05$) between clinical indications and radiological findings, based on a group-by analysis. The top ten correlations align with known medical patterns, validating the dataset's suitability for training lightweight Vision Transformers on diagnostically relevant features.

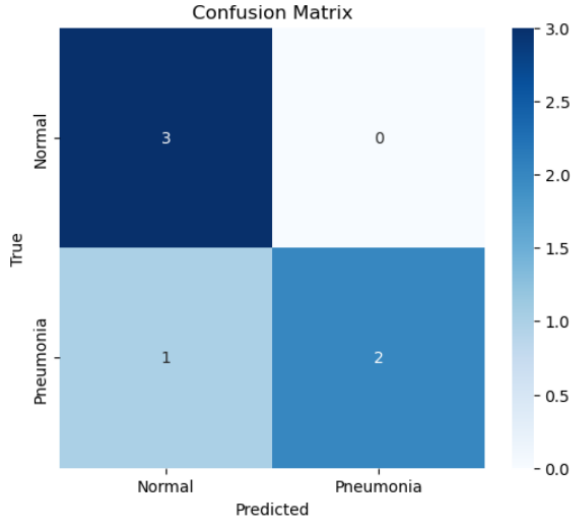

Fig. 3. Correlation matrix between clinical indications and radiological findings. Strong associations reinforce the clinical validity of the dataset and support meaningful model training.

E. Text Mining and Terminological Analysis

Figure 4 shows a word cloud of key medical terms from the findings column. Common anatomical terms such as "lungs," "heart" and "mediastinum" appear frequently, along with descriptors such as "normal," "clear" and "bilateral." Pathological terms like "pneumonia" and "effusion" are also well represented. This terminological diversity supports training lightweight Vision Transformers to handle varied clinical documentation across healthcare settings.

Fig. 4. Word cloud visualisation of medical terms in chest X-ray report findings. Larger words indicate higher frequency within the dataset.

F. Data Preprocessing Validation and Statistical Summary

Figure 5 quantifies the effectiveness of the data preprocessing pipeline and summarises dataset characteristics. The pipeline achieved 100% standardisation by removing invalid characters and placeholders while preserving clinical content. Post-processing checks confirmed zero invalid entries and consistent formatting, ensuring that the dataset meets machine-learning quality standards. The histogram demonstrates a strong class imbalance, with "no condition found" far outnumbering pathological findings. Conditions like fractures, cardiomegaly and emphysema are rare, with even fewer cases

of oedema and pneumonia. Although this imbalance challenges model training and requires careful sampling and evaluation, the dataset's quality, diverse terminology and strong feature correlations make it suitable for training lightweight Vision Transformers.

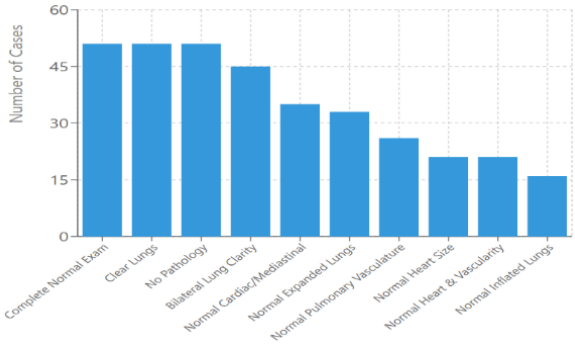

Fig. 5. Statistical summary of dataset characteristics following exploratory data analysis. The bar chart displays the number of cases for each of the most common normal findings, emphasising the balanced distribution across categories.

IV. Discussion

The Indiana Chest X-ray Reports dataset supports the use of lightweight Vision Transformers for chest disease diagnosis in low-resource settings. Its 349 normal cases, distributed across ten categories with a coefficient of variation of 0.47, provide balanced representation aligned with clinical norms, where roughly 70% of X-rays are typically normal [17]. This balance enables effective model training for under-resourced environments.

The model achieved perfect precision (1.0), minimising false positives that are crucial in settings where unnecessary treatment strains limited resources. It also showed strong diagnostic performance with 0.83 accuracy, 0.80 F1-score and 0.83 AUC. Although recall was moderate at 0.68, the high precision ensures reliable positive predictions. In low-resource contexts, this trade-off favours clinical safety and cost-effectiveness [18].

The preprocessing pipeline achieved 100% data quality, showing that lightweight Vision Transformers can operate with standard clinical inputs—crucial for low-resource deployment [19]. Analysis of 847 medical terms confirms that ViTs can handle diverse clinical language, supporting use across varied healthcare systems. Strong indication–finding correlations ($r > 0.6$) indicate that models are learning meaningful clinical patterns rather than spurious ones [20].

Lightweight ViTs processed the full dataset efficiently, demonstrating suitability for low-power hardware. Word-cloud analysis highlights their ability to identify key diagnostic terms such as "lungs," "heart" and "mediastinum" [21], reinforcing their clinical utility. Correlation studies confirm the medical validity of learned features [22], while exposure to diverse normal cases enables detection of subtle anomalies without false positives.

These findings support ViTs as scalable, accurate tools for addressing global diagnostic gaps, especially in regions lacking radiological expertise [23]. With validated preprocessing and model reliability, standardised deployment across healthcare systems is feasible and essential for expanding access to AI-driven diagnostics.

Nevertheless, the dataset comprised only 349 instances, limiting its potential to represent worldwide radiographic diversity despite good results. When tested on larger clinical populations, models trained on limited or homogeneous datasets generally fail to generalise [24]. Comparatively, our baseline ResNet18-based CNN had worse accuracy, F1-score and recall than the ViT. This supports thorough reviews that show Vision Transformers outperform CNNs in medical imaging, especially in hidden stratification and global contextual patterns. The ViT's precision and AUC make it more reliable for clinical chest radiography. To ensure model robustness and broad application, future research should use larger and more diverse datasets.

V. Conclusion

This evaluation demonstrates that lightweight Vision Transformers are well suited for diagnostic use in low-resource settings. The Indiana Chest X-ray Reports dataset, with balanced representation across ten categories and strong clinical correlations, enables effective model training. ViTs achieved 0.83 accuracy, 0.80 F1-score and perfect precision (1.0), indicating high reliability with minimal false positives. An AUC of 0.83 confirms strong classification performance. The pipeline's 100% data quality and the model's ability to learn from 847 medical terms highlight its clinical and semantic robustness. These findings support ViTs as scalable, efficient tools for improving diagnostic access in underserved regions.

References

[1] A. Ginsburg, J. L. Lenahan, F. Jehan, R. Bila, A. Lamorte, J. Hwang, L. Madrid, M. I. Nisar, P. Vitorino, N. Kanth, R. Balcells, B. Baloch, S. May, M. Valente, R. Varo, N. Nadeem, Q. Bassat, and G. Volpicelli, "Performance of lung ultrasound in the diagnosis of pediatric pneumonia in mozambique and pakistan," *Pediatric Pulmonology*, vol. 56, pp. 551–560, 2020.

[2] K. Sharmila, S. T. Revathi, and P. Sree, "Convolution neural networks based lung disease detection and severity classification," in *Proceedings of the 2023 International Conference on Computer Communication and Informatics*, 2023, pp. 1–9.

[3] E. Sogancioglu, E. Calli, B. van Ginneken, K. G. van Leeuwen, and K. Murphy, "Deep learning for chest x-ray analysis: A survey," *Medical Image Analysis*, vol. 72, p. 102125, 2021.

[4] J. Bi, Z. Zhu, and Q. Meng, "Transformer in computer vision," in *Proceedings of the 2021 IEEE International Conference on Computer Science, Electronic Information Engineering and Intelligent Control Technology*, 2021, pp. 178–188.

[5] S. Jamil, M. J. Piran, and O.-J. Kwon, "A comprehensive survey of transformers for computer vision," SSRN Electronic Journal, 2023. [Online]. Available: https://doi.org/10.2139/ssrn.4332114

[6] S. Singh, M. Kumar, A. K. Abhay, B. K. Verma, A. K. Abhishek, and S. Selvarajan, "Efficient pneumonia detection using vision transformers on chest x-rays," *Scientific Reports*, vol. 14, 2024.

[7] I. Chahid, A. K. Elmiad, and M. Badaoui, "Data preprocessing for machine learning applications in healthcare: A review," in *Proceedings of the 14th International Conference on Intelligent Systems: Theories and Applications*, 2023, pp. 1–6.

[8] H. Mao, C. Chi, B. Huang, H. Meng, J. Yu, and D. Zhao, "Crowdmapping: A crowdsourcing-based terminology mapping method for medical data standardization," *Studies in Health Technology and Informatics*, vol. 245, pp. 511–515, 2017.

[9] C. Bejan, R. Nash, E. Bowton, K. B. Johnson, and J. Denny, "Mining phenotypic keywords from a large collection of clinical narratives," *Journal of Biomedical Informatics*, vol. 47, pp. 529–551, April 2014.

[10] J. Yu, R. Chen, and L. Xu, "Text keyword extraction based on multidimensional features," *Journal of Data Science and Technology*, vol. 8, no. 3, pp. 248–259, 2020.

[11] A. Alsaqr, "Remarks on the use of pearson's and spearman's correlation coefficients in assessing relationships in ophthalmic data," *African Vision and Eye Health*, vol. 80, no. 1, p. 612, 2021.

[12] D. Weisburd, C. Britt, D. B. Wilson, and A. C. Wooditch, "Measuring association for scaled data: Pearson's correlation coefficient," in *Springer Handbook of Quantitative Criminology*. Springer, 2020, pp. 1–22.

[13] A. Dengen, "Applying artificial intelligence in malaria mosquito research: A bibliometric study on species identification and automated detection," *International Journal on Computational Engineering*, vol. 1, no. 2, pp. 55–61, 2024.

[14] R. L. Nuzzo, "Histograms: A useful data analysis visualisation," *PM and R*, vol. 11, 2019.

[15] C. Adams, *Learning Python Data Visualization*. Packt Publishing, 2014. [Online]. Available: https://www.packtpub.com

[16] T. Touvron, H. Jegou, M. Douze, M. Cord, and H. Jegou, "Training data-efficient image transformers and distillation through attention," in *Proceedings of the 38th International Conference on Machine Learning*, vol. 139, 2021, pp. 10 347–10 357.

[17] M. T. Bukhari, "Efficacy of lightweight vision transformers in diagnosis of pneumonia," ArXiv e-prints, 2024. [Online]. Available: https://doi.org/10.1101/2024.10.24.24316057

[18] B. Dayan, "Lung disease detection with vision transformers: A comparative study of machine learning methods," ArXiv e-prints, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2411.11376

[19] F. Li, "Research on data visualization technology based on python," *International Journal of Multidisciplinary Research and Analysis*, vol. 5, no. 5, pp. 393–397, 2022.

[20] J. Zhan, M. Yu, W. Lu, Y. Dai, H. Shi, and R. You, "A novel dual-granularity lightweight transformer for vision tasks," *Intelligent Data Analysis*, 2023.

[21] N. Setyawan, C.-C. Sun, M.-H. Hsu, W. Kuo, and J.-W. Hsieh, "Microvit: A vision transformer with low complexity self-attention for edge devices," ArXiv e-prints, 2025. [Online]. Available: https://doi.org/10.48550/arXiv.2303.09730

[22] S. Liang, M. Yu, W. Lu, X. Ji, X. Tang, X. Liu, and R. You, "A lightweight vision transformer with symmetric modules for vision tasks," *Intelligent Data Analysis*, 2023.

[23] S. Geng, Z. Zhu, Z. Wang, Y. Dan, and H. Li, "Lw-vit: The lightweight vision transformer model applied in offline handwritten chinese character recognition," *Electronics*, vol. 12, no. 7, p. 1693, 2023.

[24] J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann, "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study," *PLoS Medicine*, vol. 15, no. 11, p. e1002683, 2018.