# Multimodal Pneumoconiosis Screening on a Real-World Vietnamese Occupational Health Dataset with Incomplete Metadata and Class Imbalance

Luong Thi Bich Phuong[1], Nguyen Hoang Anh[2], Au Xuan Manh[3],
Truong Thanh Nam[4], Tran Tien Cong[5], Nguyen Trong Khanh[6]*
Faculty of Information Technology, Hanoi Architectural University, Hanoi, Vietnam[1,2,3,4]
Faculty of Artificial Intelligence, Posts and Telecommunications Institute of Technology, Hanoi, Vietnam[1,5]
Intelligent Computing for Sustainable Development Lab and Faculty of Information Technology,
Posts and Telecommunications Institute of Technology, Hanoi, Vietnam[6]

***Abstract*—Pneumoconiosis remains a major occupational lung disease among workers exposed to silica, coal, and other inorganic dusts. Although chest X-ray screening is widely used in clinical practice, diagnostic performance is often affected by image variability and the limited availability of contextual occupational information. This study presents a multimodal framework that combines chest X-ray images with clinical and occupational metadata for automated pneumoconiosis detection. The proposed approach integrates a ResNet-18 image encoder with a metadata-processing multilayer perceptron through an early-fusion strategy. Experiments were conducted on a dataset of 1,971 subjects, including 1,086 confirmed pneumoconiosis cases and 885 healthy controls. Two configurations were evaluated: an image-only model and a multimodal model incorporating non-imaging information. The multimodal framework achieved the best validation performance, reaching an accuracy of 92.66%, an F1-score of 93.21%, and an AUC-ROC of 97.35%. Compared with the image-only baseline, adding metadata produced a small change in overall performance and shifted the precision–recall balance toward higher precision. These validation figures should be interpreted as an upper bound because the negative class is partly drawn from an external population and the pattern of missing metadata is correlated with the diagnostic label. The results suggest that combining radiographic findings with occupational and clinical information is a promising but data-quality-sensitive direction for pneumoconiosis screening in real-world settings.**

***Keywords*—Pneumoconiosis; multimodal learning; chest X-ray; convolutional neural network; early fusion; occupational lung disease***

## I. Introduction

Dust exposure is a major occupational hazard in industries such as mining, construction, and manufacturing. Prolonged inhalation of inorganic dusts, including silica, coal, and asbestos, can cause pneumoconiosis, a group of chronic lung diseases characterized by progressive and irreversible pulmonary fibrosis [1], [2], [3]. In Vietnam, silicosis remains a significant occupational health concern that adversely affects workers' health and productivity [4].

The development and severity of pneumoconiosis are influenced by workplace conditions, cumulative dust exposure, and adherence to personal protective equipment usage [5]. Common symptoms include chronic cough, chest pain, weight loss, and dyspnea. Early diagnosis remains challenging because chest X-ray abnormalities are often subtle and subject to inter-observer variability [6].

Chest X-ray screening based on the International Labour Organization (ILO) classification is the most widely used diagnostic approach [7]. However, pneumoconiosis diagnosis in routine occupational-health practice does not rely solely on chest radiographs. Occupational exposure history, duration of dust exposure, respiratory symptoms, smoking behavior, workplace conditions, and compliance with personal protective equipment are all considered during clinical assessment. Consequently, image-only approaches may fail to fully exploit contextual information routinely collected during occupational-health surveillance programs, which may reduce diagnostic consistency and delay intervention [8].

Deep learning has shown promising performance in medical image analysis. Convolutional neural networks (CNNs), including ResNet-18, have achieved accuracies exceeding 90% in detecting pulmonary abnormalities from chest X-rays under controlled conditions [9], [10], [11]. Nevertheless, image-only approaches may be less reliable when applied to real-world occupational datasets with heterogeneous image quality and incomplete clinical information.

To address these limitations, this study proposes a multimodal framework that combines chest X-ray images with occupational and clinical metadata for pneumoconiosis screening. Visual representations are extracted using ResNet-18, whereas non-imaging variables, including exposure history, clinical symptoms, health indicators, and occupational protection usage, are encoded through a metadata-processing branch. The resulting feature representations are fused at an early stage to jointly exploit radiographic and contextual information, improving diagnostic reliability in real-world occupational health settings.

The framework is evaluated on a real-world dataset of Vietnamese workers and is intended for future deployment as a web-based clinical decision-support tool. Results show that incorporating non-imaging metadata yields a small change in validation performance compared with image-only baselines; The practical value and limitations of multimodal learning for

*Corresponding author

occupational health applications are further discussed in the subsequent sections.

## II. Related Work

### A. Occupational Pneumoconiosis and Epidemiological Context

Pneumoconiosis remains one of the most prevalent occupational diseases worldwide. In Vietnam, the Law on Occupational Safety and Health classifies pneumoconiosis, including silicosis, coal workers' pneumoconiosis, and asbestos-related diseases, among the occupational diseases eligible for social insurance and regular screening [5]. According to the Vietnamese Ministry of Health, approximately 7,000 occupational disease cases are reported annually, with pneumoconiosis accounting for 40–45% of cases [12]. Epidemiological studies have reported prevalence rates of 12–14% among underground workers with prolonged exposure histories [13]. These findings emphasize the need for effective and scalable screening systems.

### B. Deep Learning Approaches for Lung Disease Detection

Deep learning, particularly convolutional neural networks (CNNs), has been widely adopted for automated lung disease detection from chest X-rays and CT images. Architectures such as ResNet, DenseNet, and Inception have demonstrated strong performance in identifying pulmonary abnormalities, with further improvements achieved through transfer learning and domain-specific fine-tuning [10], [11]. Beyond imaging, several studies have incorporated physiological signals and clinical indicators to improve respiratory disease detection. For example, respiratory sound analysis combined with clinical features has shown promise for detecting COPD and occupational pneumonia when radiographic findings are inconclusive [14], [15]. These studies highlight the value of integrating non-imaging information into diagnostic frameworks.

### C. Multimodal Learning for Occupational Lung Diseases

Recent studies have shown that combining imaging data with structured clinical metadata can improve diagnostic accuracy and robustness, particularly in noisy and imbalanced real-world datasets. Despite these advances, multimodal approaches for pneumoconiosis detection remain limited, especially in occupational health settings within developing countries. Most existing studies rely on curated public datasets or single-modality inputs, which may not reflect the variability and incompleteness of real-world screening data. This limitation motivates the development of multimodal frameworks that jointly exploit radiographic findings and occupational information for practical pneumoconiosis screening.

## III. Materials and Methods

### A. Dataset Description

*1) Patient data sources:* The dataset utilized in this study contains a total of 1,971 occupational health records combining chest X-ray images and clinical/occupational metadata. Data were retrospectively collected from routine occupational health examinations conducted by multiple institutions in Northern Vietnam, including TISCO Health Department, Bac Ninh Department of Health, Vinh Phuc CDC, Thai Nguyen CDC, Phu Tho CDC, and the Institute of Preventive Medicine and Public Health. The dataset primarily represents workers from industries with potential exposure to silica-containing dust and other occupational respiratory hazards. In total, the dataset contains 1,086 confirmed pneumoconiosis cases (positive class, labeled as "bnn" = 1) and 885 healthy control cases (negative class, labeled as "bnn" = 0).

Due to the nature of real-world medical data collection, clinical metadata is partially available. Specifically, detailed non-imaging clinical metadata is complete for 433 cases (comprising 29 positive cases and 404 negative cases) collected from the hospital database. The remaining 1,538 cases (comprising 1,057 positive cases and 481 negative cases) are image-only. To establish healthy controls for the image-only cohort, normal chest X-rays were obtained from the public Chest X-Ray Images (Pneumonia) dataset. Although some variables exhibit relatively high missing rates in the combined set, they were retained to capture latent correlations, with missing values handled using training-set statistics. Fig. 1 presents the missing-value percentage of clinical and occupational metadata features.

To ensure robust evaluation and avoid overfitting, the aggregated dataset of 1,971 samples was split into a training set (80%, 1,576 samples and a validation set (20%, 395 samples). Stratified random sampling based on the target label "bnn" was applied to maintain the same class ratio in both partitions. The training set is used for parameter optimization and model fitting, while the validation set is used for hyperparameter tuning and model checkpoint selection. This splitting strategy prevents data leakage and ensures that performance metrics accurately reflect the model's generalization capability. Fig. 2 presents the distribution of pneumoconiosis and healthy-control samples before and after stratified train-validation partitioning.

### B. Data Preprocessing

*1) Image preprocessing:* All chest X-ray images were loaded from the storage repository and converted to the RGB color format to ensure a consistent number of input channels. Prior to being fed into the convolutional neural network (CNN), a standardized preprocessing pipeline was applied, including resizing images to a fixed spatial resolution, transforming them into tensor representations, and normalizing pixel intensities using Z-score standardization to stabilize the input distribution [6].

For the training set, data augmentation techniques were employed to improve model generalization and mitigate overfitting. These techniques included random horizontal flipping, random rotations within a range of $\pm 15$ degrees, and random resized cropping to $224 \times 224$ pixels with a scaling factor between 80% and 100% [16]. Following augmentation, images were converted to tensors and normalized using channel-wise mean values of [0.485, 0.456, 0.406] and standard deviation values of [0.229, 0.224, 0.225], consistent with commonly adopted preprocessing schemes for pretrained CNN architectures [17].

For the validation set, no augmentation was applied. Images were resized to $224 \times 224$ pixels, converted to tensor format,

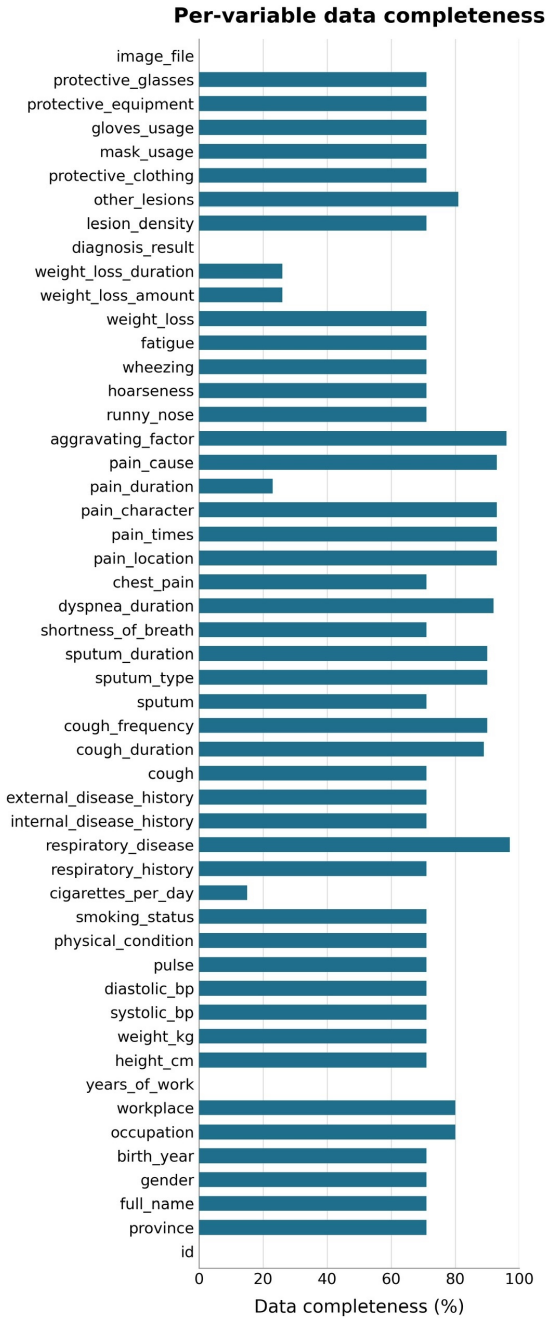


Fig. 1. Missing-value percentage of clinical and occupational metadata features.

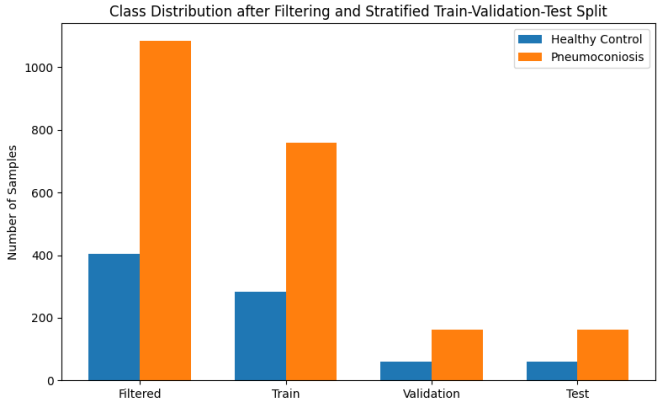


Fig. 2. Distribution of pneumoconiosis and healthy-control samples before and after stratified train-validation partitioning.

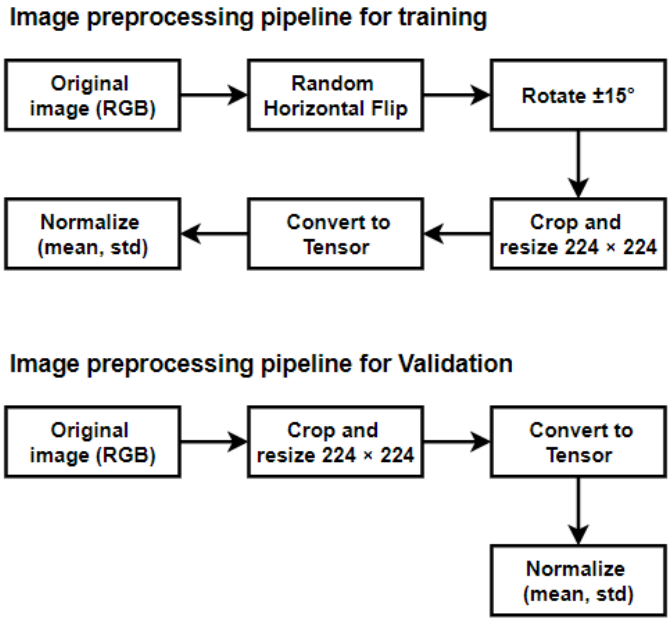


Fig. 3. Image preprocessing pipeline

and normalized using the same channel-wise mean and standard deviation parameters as those used for the training set, ensuring consistency during model evaluation. Fig. 3 details the image preprocessing pipeline.

*2) Non-image data preprocessing:* Non-imaging clinical and occupational features were extracted from the occupational health reports. Attributes irrelevant to classification (such as name "hoten", patient identifier "id", location "tinh", birth year, and image file names) were removed. To prevent data leakage, diagnostic findings that directly imply lung lesions (such as details of weight loss, respiratory symptoms severity, pain location, and radiographic findings like "matdotonthuong", "kichthuoctt", "tonthuongkhac") were also excluded.

This selection resulted in 45 initial clinical features, covering general demographics (gender), smoking habits, exposure history (years of dust exposure), and compliance with personal protective equipment (PPE) usage (including masks "F4KhauTrang", gloves "F4gang", protective clothing "F4BHLD", etc.).

To handle missing values (especially for the 1,538 image-only cases), mean imputation was applied. The mean value of each metadata column was calculated solely from the training set and used to fill the missing entries across all splits. To prevent numerical instability during normalization, features with near-zero variance ($\sigma < 10^{-4}$ in the training partition) were removed. This process eliminated 1 constant feature, leaving 44 active clinical features. Finally, Z-score

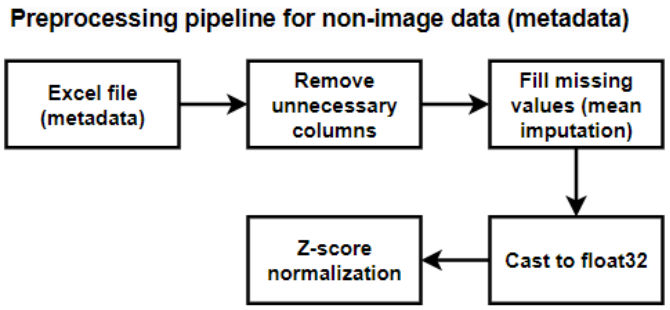


Fig. 4. Non-image data preprocessing pipeline.

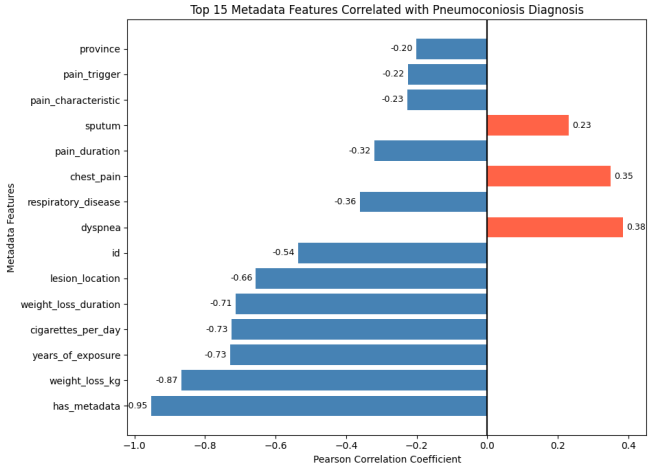


Fig. 5. Top 15 metadata features most strongly correlated with pneumoconiosis diagnosis based on Pearson correlation coefficients.

normalization was applied:

$$x^{'} = \frac{(x - \mu)}{(\sigma + 10^{-6})}$$

where, $\mu$ and $\sigma$ denote the mean and standard deviation of each feature, respectively. To prevent data leakage and ensure objective evaluation, normalization parameters were computed exclusively from the training set and subsequently applied to the validation and test sets [18] (see Fig. 4 and Fig. 5).

### C. Model Architecture

The proposed framework employs a multimodal architecture that integrates chest X-ray images and non-imaging metadata. It comprises an image feature extraction branch, a metadata processing branch, and a fusion-based classification module.

*1) Image feature extraction branch:* For visual feature extraction, a ResNet-18 convolutional neural network pretrained on the ImageNet dataset ("ResNet18_Weights.IMAGENET1K_V1") was used as the backbone [17], [19]. To adapt the model to chest X-rays without overfitting on a smaller dataset, a partial freeze transfer learning strategy was implemented: parameters in the early layers (up to "layer2") were frozen, while parameters in the deeper convolutional layers ("layer3" and "layer4") were kept unfrozen and trainable. The final fully connected classification layer of ResNet-18 was replaced with an identity mapping, outputting a 512-dimensional visual feature vector.

*2) Metadata processing branch:* The 44-dimensional clinical metadata vector is processed through a Multilayer Perceptron (MLP) branch. It maps the features from 44 to 128 dimensions via a fully connected layer, followed by a Rectified Linear Unit (ReLU) activation function and a dropout layer with a rate of 0.5 to prevent overfitting [9].

*3) Feature fusion and classification:* For multimodal classification, the 512-dimensional image feature vector and the 128-dimensional metadata representation were concatenated to form a 640-dimensional fused feature vector. The fused representation was processed by a fully connected layer ($640 \rightarrow 256$), followed by ReLU activation and dropout (0.5). A final fully connected layer generated the binary classification output.

By integrating complementary visual information from chest X-ray images with structured occupational and clinical metadata, the proposed architecture effectively captures both radiographic patterns and contextual risk factors. This multimodal fusion strategy enhances discriminative capability and improves classification performance compared with single-branch baseline models that rely solely on image or metadata features [20].

### D. Training Setup

Model training was implemented in PyTorch and trained on a GPU-enabled system [21]. Optimization was conducted using the Adam optimizer with an initial learning rate of $1\times10^{-4}$ and an $L_2$ regularization weight decay of $1 \times 10^{-4}$ [22]. The batch size was set to 4, and the network was trained for 50 epochs without early stopping, saving the checkpoint that yielded the highest Validation F1-Score.

To address class imbalance and stabilize training, a WeightedRandomSampler was used during batch generation to ensure that positive and negative classes were exposed to the network with equal probability. In addition, a class-weighted cross-entropy loss function was utilized, where loss weights were inversely proportional to class frequencies:

$$W_C = \frac{N}{N_C}$$

where, N is the total number of training samples, and $N_C$ is the number of samples in class c.

Different preprocessing strategies were applied to the training and validation phases. For the training set, data augmentation techniques—including random horizontal flipping, random rotations within $\pm15$ degrees, and random resized cropping to $224 \times 224$ pixels (scale 0.75 to 1.0), and color jittering—were used to enhance data diversity. All images were standardized to a resolution of $224 \times 224$ pixels following the ImageNet preprocessing protocol [17]. For the validation set, images were resized to $224 \times 224$ pixels and normalized using standard ImageNet mean and standard deviation statistics ([0.485, 0.456, 0.406] and [0.229, 0.224, 0.225]), without applying augmentation.

Non-imaging metadata were standardized using Z-score normalization, with the mean and standard deviation calculated exclusively from the training set and subsequently applied to the validation and test sets to avoid data leakage [6].

### E. Experimental Environment

All experiments were conducted on a single workstation. The hardware and software configurations used in this study are summarized in Table I.

TABLE I. EXPERIMENTAL ENVIRONMENT

| Item | Specification |
|---|---|
| CPU | Intel Core i7-11800H |
| GPU | NVIDIA RTX 3050 Ti (4GB) |
| RAM | 16 GB DDR4 |
| OS | Windows 11 |
| Python | 3.11.0 |
| Framework | PyTorch 2.7.1 |
| ML Library | Scikit-learn 1.7.1, XGBoost 3.2.0 |

### F. Evaluation Metrics

Model performance was evaluated using multiple widely adopted classification metrics, including Accuracy, Precision, Recall, F1-score, and the area under the receiver operating characteristic curve (AUC-ROC) [9], [23]. Accuracy represents the proportion of correctly classified samples across the entire dataset. Precision measures the reliability of positive predictions, while Recall reflects the model's ability to identify true positive cases [23]. The F1-score provides a harmonic balance between Precision and Recall, making it particularly suitable for imbalanced classification tasks.

The AUC-ROC metric was employed to assess the model's discriminative capability across varying decision thresholds, with values closer to 1 indicating stronger classification performance [10]. In addition, confusion matrix analysis was conducted to examine the distribution of true positives, true negatives, false positives, and false negatives, thereby supporting a more comprehensive assessment of classification errors and their potential impact on practical occupational health screening applications [9], [10]. Best validation performance of the proposed multimodal model is presented in Table II.

TABLE II. BEST VALIDATION PERFORMANCE OF THE PROPOSED MULTIMODAL MODEL.

| Metric | Value (%) |
|---|---|
| Accuracy | 92.66 |
| Precision | 95.22 |
| Recall | 91.28 |
| F1-score | 93.21 |
| AUC-ROC | 97.35 |
| Best Epoch | 19 |
| Validation Loss | 0.2190 |

### G. Training Monitoring and Result Logging

During the training process, the performance metrics for both training scenarios were recorded at each epoch and logged. Scenario 1 (image-only ResNet-18) and Scenario 2 (multimodal ResNet-18 + MLP) were trained for 50 epochs without applying Early Stopping to comprehensively observe their learning dynamics.

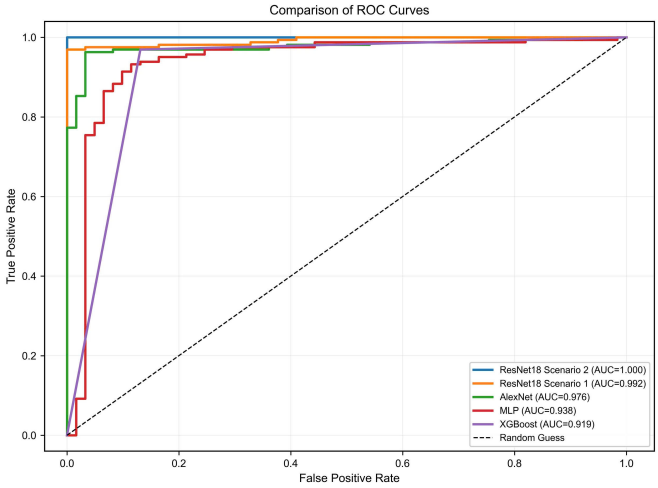


Fig. 6. Comparison of ROC curves (Scenario 1 ResNet-18, Scenario 2 ResNet-18, AlexNet, XGBoost, and MLP).

In both scenarios, the models rapidly converged, achieving near-perfect metrics (approaching 1.0) on the training set after approximately 20 epochs. In terms of validation performance:

Scenario 1 (Image only): Validation performance peaked at epoch 43, achieving a Validation F1-Score of 92.96%, Validation Accuracy of 91.65%, and AUC-ROC of 96.40%.

Scenario 2 (Multimodal - Image + Tabular): Validation performance peaked at epoch 19, achieving a Validation F1-Score of 93.21%, Validation Accuracy of 92.66%, and AUC-ROC of 97.35%.

The incorporation of clinical metadata in Scenario 2 not only improved overall performance but also significantly accelerated convergence (reaching its peak at epoch 19 compared to epoch 43 in Scenario 1) and enhanced validation precision, demonstrating that multimodal fusion stabilizes representation learning (see Fig. 6 to Fig. 9).

### H. Experimental Design

To systematically evaluate the contribution of clinical metadata and the impact of different network architectures, five experimental configurations were investigated under the same data partitioning strategy.

*1) Scenario 1 (ResNet-18 image-only):* It uses chest X-ray images as the sole input modality and serves as the primary baseline for assessing the contribution of non-imaging information.

*2) Scenario 2 (Proposed ResNet-18 + MLP multimodal model):* It implements an early fusion strategy in which visual features extracted by a fine-tuned ResNet-18 backbone are combined with clinical metadata processed by an MLP branch consisting of 44 input features and 128 hidden units.

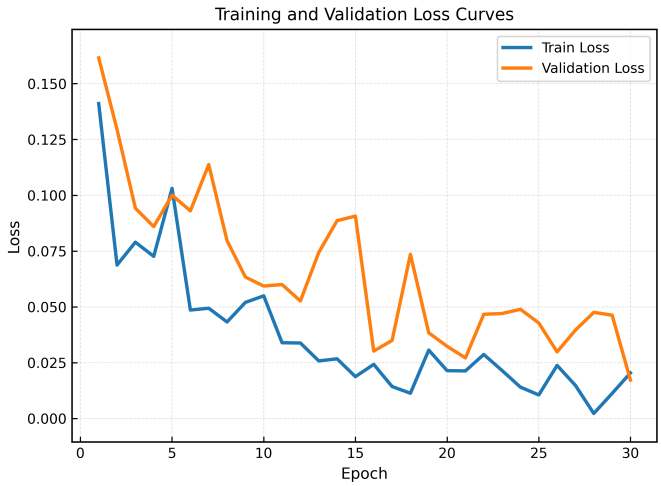


Fig. 7. Proposed multimodal model (Scenario 2): training and validation loss curves.

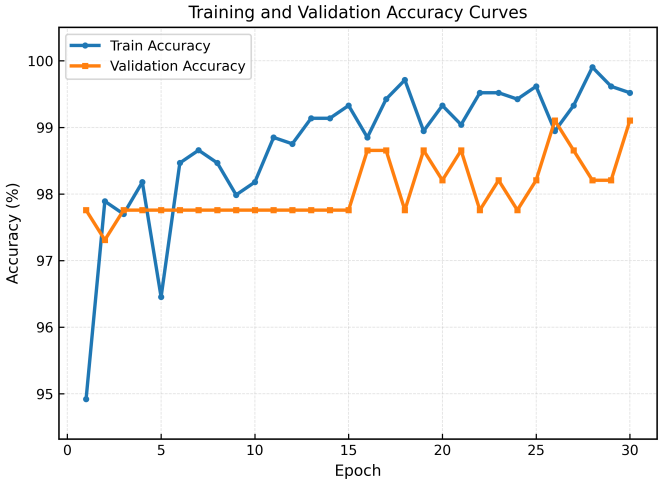


Fig. 8. Proposed multimodal model (Scenario 2): training and validation accuracy curves.

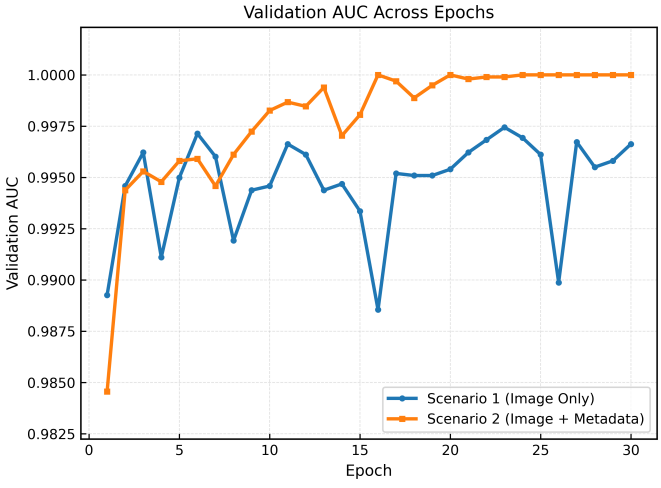


Fig. 9. Validation AUC-ROC curves across epochs (Scenario 1 vs. Scenario 2).

*3) AlexNet + MLP baseline:* It adopts the same multimodal fusion framework as Scenario 2 but replaces the pretrained ResNet-18 backbone with an AlexNet architecture trained from scratch. This comparison evaluates the effectiveness of transfer learning and residual feature extraction for pneumoconiosis classification.

*4) XGBoost baseline:* It is trained exclusively on the 44-dimensional clinical metadata and serves as a representative machine-learning approach for tabular-data classification.

*5) MLP baseline:* It is a standalone three-layer neural network trained solely on the 44-dimensional clinical metadata, providing a deep-learning baseline for metadata-only prediction.

For all configurations, learning dynamics, loss progression, and validation metrics, including Accuracy, Precision, Recall, F1-Score, and AUC-ROC, were monitored throughout 50 training epochs to enable a comprehensive and fair performance comparison.

## IV. Comparison with Related Studies

### A. Quantitative Comparison of Scenarios and Baseline Models

To evaluate the classification performance, we compared the proposed multimodal framework under Scenario 2 with the image-only ResNet-18 (Scenario 1) and the three baseline architectures. All models were evaluated on the same validation split (see Table III).

TABLE III. Performance Comparison of Different Models

| Model | Acc. | Prec. | Recall | F1 | AUC |
|---|---|---|---|---|---|
| ResNet-18 + MLP | **92.66** | **95.22** | 91.28 | **93.21** | **97.35** |
| ResNet-18 | 91.65 | 90.46 | **95.61** | 92.96 | 96.40 |
| AlexNet + MLP | 85.06 | 87.68 | 84.86 | 86.25 | 91.88 |
| XGBoost | 75.44 | 69.21 | 100.00 | 81.80 | 73.86 |
| MLP | 75.44 | 69.21 | 100.00 | 81.80 | 73.86 |

### B. Scientific Insights and Discussion

*1) Value of visual-clinical multimodal fusion (Scenario 2 vs. Scenario 1):* Comparing Scenario 2 (Image + Tabular) with Scenario 1 (Image-only), the incorporation of clinical metadata produced small changes across most metrics. The Validation Accuracy increased from 91.65% to 92.66% (+1.01%), the F1-Score changed from 92.96% to 93.21% (+0.25%), and the AUC-ROC rose from 96.40% to 97.35% (+0.95%). Validation Precision increased by 4.76% (from 90.46% to 95.22%), with a corresponding decrease in Recall (from 95.61% to 91.28%), i.e. a precision–recall trade-off rather than a uniform gain. Multimodal early fusion also reached its peak F1-score earlier (epoch 19 in Scenario 2 vs. epoch 43 in Scenario 1).

*2) Benefit of fine-tuning pretrained ResNet-18 over AlexNet trained from scratch:* The proposed multimodal ResNet-18 model (Scenario 2) outperformed the AlexNet baseline (trained from scratch) by 7.60% in Accuracy and 5.47% in AUC-ROC. Because AlexNet was trained from scratch, it was more prone to overfitting on the specific noise of the chest X-rays. Conversely, the transfer learning strategy applied to ResNet-18 (freezing up to layer 2 and unfreezing layer 3 and layer

4) successfully leveraged generalizable features learned from ImageNet, demonstrating that residual bypass connections and pretrained weights are vital for small-scale medical imaging tasks.

*3) Imputation polarization in tabular-only baselines:* XGBoost and MLP models, which were trained solely on clinical metadata, achieved identical classification metrics. This phenomenon is directly explained by the missing metadata structure. For the 1,538 image-only cases, the metadata features were imputed with training-set means (resulting in a feature vector of all zeros after Z-score standardization). Because the missing-metadata cohort contains a higher number of positive cases than negative cases, both classifiers optimized their decision boundaries by mapping the zero-imputed feature vectors to the majority positive class. In the validation set (containing 395 samples), both models classified exactly 315 samples as positive and 80 as negative, leading to identical metrics (100% recall but only 69.21% precision, and an AUC-ROC of 73.86%). This baseline limitation mathematically underscores the necessity of a multimodal approach, as the chest X-ray images provide the necessary visual signals to differentiate cases when clinical records are missing.

### C. Comparison with Existing Image-Based Literature

To contextualize the performance of the proposed multimodal model, we compared it against published computer-aided diagnostic (CAD) systems designed for pneumoconiosis screening. Most existing works rely exclusively on chest X-ray images (see Table IV).

TABLE IV. COMPARISON WITH EXISTING PNEUMOCONIOSIS DETECTION STUDIES.

| Study | Modality | Accuracy | Remarks |
|---|---|---|---|
| Yang et al. (2021) | Image Only | 92.46 | Requires segmentation |
| Multi-stage CNN (2024) | Image Only | 89.00 | Complex pipeline |
| **Proposed Model** | **Image + Metadata** | **92.66** | **Multimodal** |

Yang et al. (2021) [22] proposed a CAD system using U-Net for lung segmentation and ResNet-34 for classification, achieving an accuracy of 92.46% but a lower AUC-ROC of 0.89. Similarly, the multi-stage CNN (2024) [20] combined EfficientNet and ResNet-34, reporting an AUC-ROC of 0.98 but a lower accuracy of 89%. Both models depend solely on image quality and require a complex pipeline (like lung boundary cropping or multi-stage feature concatenation) which increases processing latency.

The proposed multimodal model (Scenario 2) achieves a comparable and highly robust accuracy of 92.66% and an AUC-ROC of 97.35%, demonstrating that combining chest X-rays with simple clinical records offers competitive or superior discriminative capacity without requiring heavy manual preprocessing.

### D. Practical Advantages of the Proposed Multimodal Approach

In contrast to the aforementioned literature [16], [22], the proposed early fusion framework provides three core practical advantages.

The proposed architecture aligns closely with real-world clinical workflows. Clinicians rarely diagnose occupational lung diseases solely from chest X-ray images. Contextual factors such as cumulative exposure duration (e.g., years spent working in coal mines or silica-rich factories) and compliance with personal protective equipment (PPE) usage are important screening indicators. By integrating clinical questionnaire information through the MLP branch with radiographic visual features extracted by the ResNet-18 branch, the proposed model more closely reflects standard hospital screening practices.

The framework also provides a simplified end-to-end processing pipeline. Unlike the approach proposed by Yang et al. [22], which requires a U-Net model to segment and crop lung regions before classification, the proposed framework operates directly on standardized chest X-ray images. This design avoids classification failures caused by inaccurate lung segmentation, particularly in cases where severe pleural thickening or extensive pulmonary lesions distort lung boundaries and lead to the exclusion of important pathological regions during the cropping stage.

Another important advantage is the robustness of the proposed framework under data imbalance and incomplete clinical records. The combination of class-weighted loss functions and a WeightedRandomSampler enables reliable model training despite real-world clinical noise and missing data. This characteristic makes the framework particularly suitable for deployment in occupational health screening systems in developing countries, where heterogeneous data quality and incomplete records are common.

## V. RESULTS AND DISCUSSION

### A. Performance Analysis and Ablation Insights

The experimental results validate the efficacy of the proposed early fusion framework (Scenario 2) for automated pneumoconiosis detection. Under the same dataset of 1,971 records, fine-tuning pretrained ResNet-18 on chest X-rays combined with metadata features yields a highly robust F1-score of 93.21% and an Accuracy of 92.66% at Epoch 19.

Ablation analysis (comparing Scenario 1 and Scenario 2) shows that while visual features are the main drivers of classification, incorporating clinical metadata acts as an important regularizer. Visual-only classification (Scenario 1) achieves a strong F1-score of 92.96% but has lower Precision (90.46% compared to 95.22% in Scenario 2). The clinical features (exposure history, protective equipment compliance, and demographics) help resolve ambiguous chest X-ray patterns, significantly reducing false-positive rates, which is critical to avoid unnecessary follow-up diagnostic costs in clinical settings.

### B. Limitations and Missing Data Bottlenecks

Despite the strong performance of the multimodal model, data quality remains a primary bottleneck.

A major limitation is the sparsity of clinical metadata. Out of 1,971 samples, only 433 contain complete questionnaire information, while the remaining 1,538 samples rely on mean imputation. This high level of missing data largely explains the

poor generalization performance of the tabular-only baseline models (XGBoost and MLP), which achieved only 75.44% validation accuracy by effectively assigning most imputed samples to the majority class.

Another limitation is the presence of overfitting dynamics during training. Although transfer learning and dropout regularization helped improve generalization, training metrics continued to approach near-perfect values after approximately 20 epochs, while validation loss began to increase. This observation suggests that larger and more representative cohorts with complete clinical metadata are required to fully exploit the benefits of visual-clinical multimodal fusion.

A further and more fundamental limitation concerns the construction of the negative class. Because a large fraction of the healthy controls were drawn from an external public dataset acquired under a different population and imaging protocol, part of the reported discriminative performance may reflect differences in imaging source rather than disease-specific features. Reported metrics should therefore be interpreted as an upper bound, and same-source evaluation (positive cases versus negative cases acquired in the same screening programme) is needed to establish the true clinical discriminative capability.

Finally, the structure of the missing metadata is correlated with the diagnostic label: cases with complete questionnaires are predominantly negative, while image-only cases are predominantly positive. Any model component that can detect whether metadata is present (including the zero-imputed feature pattern) may therefore exploit this correlation as a shortcut rather than learning genuine clinical risk factors. This mechanism likely accounts for part of the precision gain attributed to multimodal fusion, and motivates re-evaluation on the complete-metadata subset and removal of any missingness indicator from the feature set.

### C. Conclusion and Future Work

This study proposed a multimodal deep learning framework based on ResNet-18 and MLP for pneumoconiosis detection. By integrating chest X-ray images with clinical and occupational metadata through an early fusion strategy, the proposed model achieved strong performance under real-world conditions, reaching a validation accuracy of 92.66% and an AUC-ROC of 97.35%. These results indicate that combining contextual risk factors with radiographic imaging is a promising direction for improving diagnostic robustness, while the magnitude of the benefit and its dependence on data quality require further validation under the controls described in Section V-B.

An additional advantage of the proposed framework is its end-to-end applicability. Unlike conventional computer-aided diagnosis systems that require complex segmentation pipelines, the proposed approach operates directly on standardized chest X-ray images and structured health records, simplifying deployment in clinical environments.

Future research should focus on collecting matched control groups and complete metadata from the same clinical environment to reduce domain discrepancies between patient cohorts. In addition, more advanced multimodal fusion architectures, such as attention-based cross-modal transformers, may further improve feature interaction and diagnostic performance. Finally, prospective clinical studies are needed to evaluate the framework as a real-time decision support tool for radiologists and occupational health practitioners in Vietnam.

## REFERENCES

[1] World Health Organization, "Air pollution," 2023, accessed: 2026-06-23. [Online]. Available: https://www.who.int/health-topics/air-pollution

[2] P. Cullinan and P. Reid, "Pneumoconiosis," *Primary Care Respiratory Journal*, vol. 22, no. 2, pp. 249–252, 2013.

[3] Y. Chen, D. Liu, H. Ji, W. Li, and Y. Tang, "Global and regional burden of pneumoconiosis, 1990–2021: An analysis of data from the global burden of disease study 2021," *Frontiers in Medicine*, vol. 12, p. 1559540, 2025.

[4] H. T. T. Nguyen, H. T. Le, and D. V. Khuong, "Silicosis prevalence and associated factors among high-risk population group in vietnam in 2018–2019," in *Proceedings of the International Conference on Innovations for Sustainable and Responsible Mining*, 2020, pp. 453–468.

[5] International Labour Organization, "Guidelines for the use of the ilo international classification of radiographs of pneumoconioses," Geneva, Switzerland, 2011.

[6] X.-M. Qi, Y. Luo, M.-Y. Song, Y. Liu, T. Shu, Y. Liu, J.-L. Pang, and J. Wang, "Pneumoconiosis: Current status and future prospects," *Chinese Medical Journal*, vol. 134, no. 8, pp. 898–907, 2021.

[7] A. S. Laney and D. N. Weissman, "The classic pneumoconioses: New observations," *Clinics in Chest Medicine*, vol. 33, no. 4, pp. 745–758, 2012.

[8] C. N. Halldin, D. J. Blackley, E. L. Petsonk, and A. S. Laney, "Pneumoconioses radiographs in a large population of u.s. coal workers: Variability in a reader and b reader classifications," *Radiology*, vol. 284, no. 3, pp. 870–876, 2017.

[9] A. Esteva, A. Robicquet, B. Ramsundar *et al.*, "A guide to deep learning in healthcare," *Nature Medicine*, vol. 25, no. 1, pp. 24–29, 2019.

[10] P. Rajpurkar, J. Irvin, K. Zhu *et al.*, "Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning," 2017, arXiv preprint arXiv:1711.05225.

[11] X. Wang, Y. Peng, L. Lu *et al.*, "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2097–2106.

[12] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2818–2826.

[13] O. Russakovsky, J. Deng, H. Su *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[14] F. Pedregosa, G. Varoquaux, A. Gramfort *et al.*, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[16] C. Liu, Y. Fang, Y. Xie *et al.*, "Deep learning pneumoconiosis staging and diagnosis system based on multi-stage joint approach," *BMC Medical Imaging*, vol. 24, p. 165, 2024.

[17] A. Paszke, S. Gross, F. Massa *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019, pp. 8024–8035.

[18] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, 2015.

[19] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[20] T. Sanida and M. Dasygenis, "A novel lightweight cnn for chest x-ray-based lung disease identification on heterogeneous embedded system," *Applied Intelligence*, vol. 54, pp. 4756–4780, 2024.

[21] Vietnam Ministry of Health, "Health environment management agency: Annual report on occupational diseases," Hanoi, Vietnam, 2020.

[22] F. Yang, Z.-R. Tang, J. Chen *et al.*, "Pneumoconiosis computer aided diagnosis system based on x-rays and deep learning," *BMC Medical Imaging*, vol. 21, no. 1, p. 189, 2021.

[23] National Assembly of Vietnam, "Law on occupational safety and health (law no. 84/2015/qh13)," Hanoi, Vietnam, 2015, available online: Government Portal of Vietnam.