La Medicina del Lavoro

Original article

Med. Lav. 2026; 117 (2): 18371
DOI: 10.23749/mdl.2026.18371

Artificial Intelligence in Occupational Health Surveillance: Evaluating AI-Assisted ILO Classification of Radiographs of Pneumoconioses

Antonio Baldassarre[1,2,*], Martina Padovan[3], Alessandro Palla[4], Augusto Quercia[5], Rita Leonori[6], Stefano Dugheri[7], Nicola Mucci[1,2], Veronica Traversini[1,2]

[1]Department of Experimental and Clinical Medicine, University of Florence, Florence, Italy
[2]Occupational Medicine, Careggi University Hospital, Florence, Italy
[3]Preventive Medicine, Tuscany North-West Health Local Unit, Italy
[4]Intel Corporation, Santa Clara, USA
[5]Past Chief Workplace Prevention and Safety Unit, Viterbo Health Local Unit, Italy
[6]Workplace Prevention and Safety Unit, Viterbo Health Local Unit, Italy
[7]Department of Life Science, Health, and Health Professions, Link Campus University, 00165 Rome, Italy

Keywords: ILO International Classification of Radiographs of Pneumoconioses; Occupational Medicine; Artificial intelligence

Abstract

Background: *Pneumoconioses remain an important occupational health issue, particularly in low- and middle-income countries. The International Labour Organization (ILO) Classification standardizes chest radiograph interpretation but requires trained readers and is affected by inter-reader variability. This study evaluated whether generative multimodal artificial intelligence (AI) models can approximate ILO-based diagnostic reasoning.* **Methods:** *Eighty-two chest radiographs from the official NIOSH B Readers syllabus were analysed using four AI systems (GPT-4o, GPT-5, MedGemma-4B, MedGemma-27B). Each image was evaluated with a standardized prompt based on the 2022 revised ILO guidelines using deterministic settings. Model outputs were mapped to ILO codes and compared with the official answer keys of the ILO Standard Radiograph Set used for B Reader training and examination. Performance metrics included balanced accuracy, sensitivity, specificity, precision, and Matthews correlation coefficient (MCC). Bootstrap 95% confidence intervals, McNemar's test, and Cohen's κ assessed performance variability and agreement.* **Results:** *All four AI models showed moderate diagnostic performance, with balanced accuracy ranging from 60.8% to 70.3%. Sensitivity remained limited (35.5%–54.9%), while specificity was consistently high (84.6%–86.2%). MedGemma-27B performed best for small opacities, GPT-5 for pleural abnormalities and for technical quality. Large opacities and rare findings were systematically under-detected. Statistical comparisons showed significant differences between models, although agreement patterns were broadly similar.* **Conclusion:** *All AI models partially followed structured ILO radiographic criteria but did not achieve expert-level performance, confirming that they cannot replace certified B Readers. Larger, real-world datasets are needed to assess their potential clinical utility as supportive tools in occupational health surveillance programs.*

1. Introduction

Pneumoconioses refers to a group of occupational interstitial lung diseases caused by prolonged inhalation of mineral dusts such as silica, coal, and asbestos [1]. Despite advances in safety standards and dust-exposure control, these diseases remain a major global public health concern. According to

Received 10.12.2025 – Accepted 08.04.2026
* Corresponding Author: Antonio Baldassare; E-mail: antonio.baldassarre@unifi.it

the Global Burden of Disease 2023 data, approximately 18700 deaths were attributed to pneumoconioses worldwide in 2023, representing an 18% increase in absolute mortality since 2000, although the age-standardised death rate declined by 37.6% over the same period [2]. This discrepancy highlights the persistent risk of occupational exposure, particularly in the mining, construction, and manufacturing sectors in low- and middle-income countries, where effective dust mitigation and surveillance systems remain limited [3]. Given the complex radiographic presentation of pneumoconiosis [4] and the need for diagnostic consistency across clinical and surveillance settings, the International Labour Organization (ILO) developed the International Classification of Radiographs of Pneumoconioses to standardize the description and grading of chest radiographic abnormalities caused by occupational dust exposure.

This system provides a structured framework for evaluating image quality, parenchymal opacities (small and large), pleural abnormalities, and other findings, ensuring reproducibility and uniform interpretation across readers and countries [5]. In the United States, the National Institute for Occupational Safety and Health (NIOSH) adopted this classification as the reference standard for its B Reader Program, which certifies physicians who demonstrate proficiency in the ILO system through structured training and testing. This certification ensures high intra- and inter-reader reliability and remains the cornerstone of pneumoconioses surveillance and compensation programs worldwide [6, 7]. This methodological rigor reflects a broader scientific pursuit: the standardization of observation and reasoning as a foundation of reproducible knowledge.

Throughout history, humans have sought to digitize and standardize processes of reasoning and calculation. From Charles Babbage's differential engine [8, 9] and punched-card systems [10], early mechanical algorithms that inspired modern computing to contemporary data-driven models [11], the pursuit of reproducibility and efficiency has been central to scientific progress. In occupational medicine, the same principle of structured reasoning underpins the ILO classification, which can itself be regarded as a diagnostic algorithm, defining explicit decision steps for classifying radiographic findings related to dust exposure. Since the First International Conference of Experts on Pneumoconioses, held in 1930 in Johannesburg, the classification of radiographs of pneumoconioses has undergone successive revisions in 1938 (American), 1944 (Eck and Hanaut's), 1948 (Hasselt), 1949 (British) 1950 (Sydney), 1958 (Geneva), 1968, 1971, 1980, 2000, 2011, and most recently in 2022. The initial versions primarily addressed silicosis, but by 1958 the system had been broadened to include all types and profusions of linear markings (Geneva classification).

A major advancement occurred in 1968, when the ILO classification was harmonized with the International Union Against Cancer (UICC) system, thereby extending its scope to encompass all dust-induced pneumoconioses, including irregular opacities characteristic of asbestos-related disease. This classification system was developed to provide a simple, systematic, and reproducible method for codifying radiographic abnormalities associated with pneumoconioses, thereby enabling reliable international comparisons of data and supporting epidemiological studies and research, reflecting the ongoing evolution of scientific methodology and technology toward greater objectivity and precision [12, 5, 13]. Since 1950 ILO delivers Guidelines for the use of the ILO International Classification of Radiographs of Pneumoconioses in Occupational Safety and Health Series (No. 22).

The NIOSH began the B Reader program (named after the Black lung or Coal Workers' X-ray Surveillance Program) in 1974, with the intent to train and certify physicians in the ILO Classification system.

Within this historical trajectory, the emergence of artificial intelligence (AI) and large language models (LLMs) represents a natural continuation of the same goal: to enhance accuracy, reproducibility, and diagnostic reasoning through digital cognition [14, 15]. Building on the foundations of LLMs, which process and generate textual information, the multimodal design allows AI systems to analyze and reason across both textual and visual data, bridging structured diagnostic frameworks such as the ILO classification with image-based interpretation [16]. These generative systems, capable of integrating text

and image interpretation, could represent a valuable supportive tool in radiology within occupational medicine, potentially assisting occupational physicians in applying the classical rule-based framework of the ILO classification more consistently. Previous research has explored the use of machine learning and deep learning algorithms for the automated detection of pneumoconioses on chest radiographs, achieving promising diagnostic accuracy [17]. Most of these models have relied on convolutional neural networks (CNNs) trained on labelled datasets [18, 19, 20]. Although some studies have incorporated the ILO International Classification of Radiographs of Pneumoconioses as a reference standard to improve grading consistency [21, 22], few have fully aligned with its structured coding system [23]. More recently, studies have begun to evaluate also LLMs for general radiographic interpretation tasks [24, 25, 26, 27], but no published work to date has assessed multimodal generative models like ChatGPT or MedGemma using the ILO 2022 classification as a diagnostic framework. This pilot study aimed to comparatively evaluate the diagnostic accuracy of four generative multimodal AI systems (GPT-4o, GPT-5, MedGemma-4B, and MedGemma-27B) in classifying chest radiographs according to the 2022 revised ILO International Classification of Radiographs of Pneumoconioses. The study assessed whether these models could approximate B Reader level performance and identified the domains in which current generative AI systems perform best or remain limited.

2. Methods

2.1. Dataset and Reference Classification

A total of 82 chest radiographs were analysed, retrieved in DICOM (Digital Imaging and Communications in Medicine) format, the international standard to transmit, store, retrieve, print, process, and display medical imaging information. The sample size reflects the pilot nature of the study, which was designed to provide an initial proof-of-concept assessment rather than definitive statistical power. All images were obtained from the NIOSH B Readers syllabus, which includes the official ILO Standard Radiograph Set used for B Reader training and examination. The images were viewed using the NIOSH B Reader software, which enables standardized visualization for the detection of pneumoconioses. Each image had an established reference classification based on consensus readings by certified B Readers, following the ILO guidelines. This classification provides a standardized framework for describing and quantifying radiographic abnormalities caused by occupational dust exposure. Each radiograph is first assessed for technical quality (Grades 1–4), with notation of specific defects such as overexposure, underexposure, poor contrast, improper positioning, artifacts, scapular superimposition, or inadequate lung inflation. Once image quality is deemed adequate, the reader evaluates parenchymal and pleural abnormalities following the ILO's structured format. Parenchymal findings are documented on small opacities, which are classified by shape, size, profusion, and anatomical distribution. Shape and size follow the ILO's six-category coding system: rounded opacities (p, q, r) and irregular opacities (s, t, u), each defined by size ranges illustrated in the ILO standard radiographs (p/s ≤ 1.5 mm; q/t > 1.5–3 mm; r/u > 3–10 mm). The opacities are recorded across upper, middle, and lower lung zones on each side. Profusion is assigned using the twelve ordered ILO subcategories (0/– to 3/+), based on comparison with the official reference radiographs to ensure reproducibility. When present, large opacities are coded as A, B, or C, according to their maximal dimension and the extent of involvement. Pleural abnormalities are coded for pleural plaques, diffuse pleural thickening, and costophrenic angle obliteration. For plaques, the reader records their site (in profile, face-on, diaphragmatic, or chest wall), laterality, extent, thickness category (a, b, or c), and calcification. Diffuse pleural thickening is documented using parallel criteria. Additional standardized ILO symbols allow relevant associated findings such as coalescence (ax), emphysema (em), or evidence of prior tuberculosis (tb) to be noted. In addition to these morphological domains, the ILO reading form includes a structured field related to the clinical relevance of radiographic abnormalities. Specifically, Section 4C requires the reader to indicate whether the worker should be advised to

consult a physician because of findings noted on the radiograph. In this study, this item was analysed as the "clinical decision" category, reflecting the integrated judgement that combines technical, parenchymal, and pleural assessments to support occupational health management. All radiographs were anonymized and used exclusively for research and educational purposes in compliance with NIOSH policies. For each radiograph, a reading sheet consistent with the ILO classification scheme was available and had been previously completed by certified human readers. To facilitate direct comparison between the human reference classifications (provided in the official answer keys) and the AI-generated outputs, all results were transcribed into a structured Excel database. Each row corresponded to one radiograph and contained the reference values. This structure allowed a systematic case-by-case evaluation of concordance between human and AI readings across all classification domains.

2.2. Generative AI Models

Four LLMs were evaluated. One of the most widely known examples of LLMs is ChatGPT, developed by OpenAI and first introduced in 2022. Unlike traditional AI systems, ChatGPT is based on deep learning transformer architectures [28] trained on vast text corpora, enabling the model to analyse, synthesize, and generate natural language. Over time, OpenAI has continuously improved its models [29]. The GPT-4 series, released in March 2023, introduced multimodal capabilities allowing simultaneous interpretation of text and images. The enhanced GPT-4o ("omni"), officially released on May 13, 2024, unified text, vision, and audio processing within a single model architecture, greatly improving reasoning speed and visual comprehension (https://openai.com/index/hello-gpt-4o/). In August 7, 2025, OpenAI announced GPT-5, its latest multimodal model, expanding contextual depth, visual alignment, and multi-turn reasoning performance (https://openai.com/index/introducing-gpt-5/). For this study, the GPT-4o and GPT-5 multimodal models were tested, both capable of processing textual and visual inputs within a unified reasoning framework. Also, two variants of the MedGemma model were employed: MedGemma-4B and MedGemma-27B. MedGemma is an open, multimodal foundation model designed for health-related AI applications and released by Google DeepMind in 2024. The 4B model is a multimodal system (~4 billion parameters) that combines text and image understanding through a SigLIP (Sigmoid Loss for Language–Image Pretraining)-based visual encoder pretrained on medical images (chest radiography, dermatology, pathology). The 27B model (~27 billion parameters) is a larger-scale architecture optimized for medical reasoning; in this study, we specifically used its multimodal version, capable of processing both textual and visual inputs. Both models are openly available for research through the official DeepMind repository (https://deepmind.google/models/gemma/medgemma/). All LLMs evaluations were conducted on September 22, 2025.

2.3. Input Data and Prompting Procedure

All models (GPT-4o, GPT-5, MedGemma-4B and MedGemma-27B) were applied without additional fine-tuning, using a standardized prompt and textual context from the ILO classification guidelines to ensure consistent task interpretation [30]. The primary prompt was formulated as follows:

> *"Assume the role of an occupational physician specialized in chest radiography. You are trained according to the 2022 revised edition of the ILO International Classification of Radiographs of Pneumoconioses. Based on this classification, analyse and report the chest X-ray image provided. Structure your report strictly following the ILO 2022 criteria".*

This prompt was then integrated with the full 2022 revised ILO International Classification guidelines. For all model runs, the following system parameters were used to ensure consistency and reproducibility of outputs. The temperature was set to 0.0, making every response fully deterministic, so identical cases yield identical results regardless of when or how often the prompt is run. The maximum token limit was 4096, which allowed for detailed and complete classification reports. A greedy

decoding strategy (argmax) was applied, and top-p (nucleus sampling) was not utilized due to the zero-temperature setting. These settings ensured that the output for each prompt was stable and not subject to random sampling variation.

The prompt was crafted to maximize consistency and clarity across all model outputs, reflecting practical priorities in occupational radiology. By requiring outputs in a strict JSON schema mapped to the ILO coding system, the structure and format of each model's answer align directly with official criteria, eliminating ambiguity and simplifying downstream analysis. Assigning an explicit role (certified B Reader) instructs the model to adopt the mindset and approach of a trained occupational physician, which guides its interpretation toward domain-specific standards and reasoning.

Inclusion of the full ILO guidelines in the prompt removes the need for the model to rely on memory or incomplete information, ensuring that every classification is made with direct reference to the current standard. This comprehensive context supports both accuracy and fairness in evaluation, allowing each model to perform as if it were consulting the primary source material, just as a human expert would.

2.4. AI-Based Image Analysis

For each image, a new independent conversation was initiated with the LLM. This design simulated the workflow of an occupational physician performing a B Reader evaluation, allowing the model to generate a case-specific report based solely on the information contained in the image and the official ILO framework. By isolating each case in a distinct conversational environment, the study minimized inter-case contamination and ensured that the model's reasoning and descriptive output reflected a *de-novo* evaluation for every radiograph. As outlined above, each classification produced by the models was stored in a structured JSON output for subsequent analysis.

2.5. Evaluation and Comparison

The AI-based classifications were compared with the reference (human) classifications across all 82 images. For each model and each image, the predicted outputs were aligned with the corresponding ILO 2022 reference codes. The following performance metrics were calculated [31, 32, 33]:

- Sensitivity (Recall): proportion of pathological cases correctly identified by the model.
- Specificity: proportion of normal cases correctly identified as negative.
- Precision (Positive Predictive Value, PPV): probability that a case predicted as pathological is truly positive.
- Balanced Accuracy: mean of sensitivity and specificity, giving equal weight to both classes and minimizing class imbalance bias.
- Matthews Correlation Coefficient (MCC): overall correlation between predicted and true classifications, ranging from –1 (inverse correlation) to +1 (perfect agreement).

Additionally, category-specific analyses were conducted according to the ILO coding structure (technical quality, small and large parenchymal opacities, pleural abnormalities, and other findings) to identify the areas of highest and lowest consistency across models. To examine the effect of class imbalance on model behaviour, we also computed the Class Imbalance Ratio for each ILO category, defined as the ratio between the number of negative and positive cases predicted by the model. This metric provides an indication of whether the model tends to over-predict normal findings or pathological findings. Lower ratios indicate a greater tendency to detect positive cases, which is clinically relevant for screening applications. This analysis was exploratory and descriptive, and no inferential statistical testing was performed. All metrics were computed using full numerical precision.

2.6. Statistical Analysis

To assess performance differences among models, we conducted a comparative statistical analysis across all predictions. McNemar's test was applied for pairwise comparisons of classification errors between models, with a χ^2 statistic and a significance threshold set at $p<0.05$. In addition, bootstrap

resampling with 1,000 iterations was used to estimate 95% confidence intervals for balanced accuracy, sensitivity, specificity, PPV and MCC, providing a robust assessment of performance variability. Inter-model agreement was further evaluated using Cohen's kappa (κ), interpreted according to standard qualitative thresholds (from poor to almost perfect agreement). All analyses were performed on paired predictions derived from the same radiographic dataset, ensuring a valid and direct comparison across AI systems. All performance metrics and all statistical analysis were computed using Python (SciPy v1.11).

3. Results

3.1. Overall Model Performance

The comparative performance of the four generative AI models is summarized in Table 1. Among them, MedGemma-27B achieved the highest overall balanced accuracy (70.28%), followed by GPT-4o (69.43%), GPT-5 (65.45%), and MedGemma-4B (60.84%). MedGemma-27B also obtained the highest sensitivity (54.91%) and precision (PPV) (48.34%), while MedGemma-4B recorded the best specificity (86.22%) but the lowest sensitivity (35.45%), indicating a more conservative classification pattern. The MCC ranged from 0.206 (MedGemma-4B) to 0.387 (MedGemma-27B), suggesting moderate diagnostic correlation with the human reference standard.

3.2. Category-Specific Analysis

The analysis by ILO category revealed marked heterogeneity in diagnostic performance across the different models. When assessing technical quality, GPT-4o achieved the highest balanced accuracy (62.0%), slightly outperforming GPT-5 (60.8%) and MedGemma-27B (59.6%), indicating a relatively better ability to recognize image adequacy and exposure-related artifacts. For small parenchymal opacities, MedGemma-27B demonstrated the most consistent performance, with a balanced accuracy of 59.0% and a sensitivity of 80.2%, suggesting a greater capacity to detect subtle parenchymal changes compared with the other models. All models struggled with large opacity classification. GPT-4o, GPT-5, and MedGemma-27B showed balanced accuracies of 50% due to complete failure to detect any large opacities. MedGemma-4B showed improved performance (75% balanced accuracy), but its sensitivity of 50% still indicates limited clinical reliability. In contrast, pleural abnormalities were identified with comparatively higher accuracy, particularly by GPT-5, which achieved the best-balanced accuracy (71.8%). This finding is of clinical

Table 1. Overall diagnostic performance of four generative AI models for ILO pneumoconioses classification. Balanced Accuracy, Sensitivity (Recall), Specificity, Precision (PPV), Matthews Correlation Coefficient (MCC) are reported for each model, along with their corresponding 95% confidence intervals. All metrics are reported as percentages, except the Matthews Correlation Coefficient (MCC), which is reported on its natural scale (−1 to +1), with confidence intervals expressed on the same scale.

Metric	GPT-4o	GPT-5	MedGemma-27B	MedGemma-4B
Balanced Accuracy (95% CI)	69.43% (67.24-71.81)	65.45% (63.29-67.80)	70.28% (68.16-72.69)	60.84 (58.60-63.97)
Sensitivity (Recall) (95% CI)	54.08% (50.00-58.39)	46.31% (42.56-50.40)	54.91% (50.79-59.37)	35.45% (30.81-44.17)
Specificity (95% CI)	84.78% (83.84-85.76)	84.59% (83.62-85.58)	85.66% (84.96-86.40)	86.22% (82.22-87.76)
Precision PPV (95% CI)	46.28% (44.10-48.50)	41.95% (39.69-44.12)	48.34% (45.90-50.71)	31.43% (28.98-34.52)
MCC (95% CI)	0.367 (0.330-0.410)	0.298 (0.257-0.337)	0.387 (0.349-0.426)	0.206 (0.168-0.251)

relevance, as pleural irregularities often represent early or coexisting manifestations of asbestos-related disease. Performance for other abnormalities was uniformly poor across all models, with sensitivity values below 11%, indicating that atypical or less frequent radiological signs are still poorly recognized by current generative AI. Finally, the clinical decision category, which integrates multiple features into an overall interpretative judgment, showed only moderate consistency across systems, with balanced accuracies ranging from 50% to 55%. GPT-5 shows the best performance in this category with 54.83% balanced accuracy. However, sensitivity is concerning (21.43%), indicating potential missed cases. No model clearly outperformed the others in this integrative diagnostic domain, suggesting that while AI systems can approximate human scoring in individual parameters, the holistic synthesis required for final classification remains challenging.

3.3. Class Balance and Overall Reliability

Technical quality yields uniformly low sensitivity and precision, while MedGemma-27B achieves the best recall for small opacities, though with only moderate precision. Performance collapses for large opacities, with all models showing perfect sensitivity but no specificity. Pleural abnormalities represent the only domain with a more favorable precision–recall balance, with GPT-5 and MedGemma-27B performing best.

3.4. Statistical Validation of Model Performance

Pairwise McNemar's tests showed statistically significant differences across all model comparisons (all $p < 0.001$), confirming non-equivalent error distributions. GPT-4o significantly outperformed GPT-5 and MedGemma-4B, while MedGemma-27B significantly outperformed GPT-5 and MedGemma-4B. Bootstrap analysis (1,000 resamples) demonstrated robust stability of performance estimates across all models, with narrow confidence intervals indicating limited variability in resampled performance metrics. Among the evaluated systems, MedGemma-27B achieved the highest balanced accuracy (70.28%, 95% CI: 68.16–72.69) and MCC (0.387, 95% CI: 0.349–0.426), suggesting a more favourable trade-off between sensitivity and specificity. GPT-4o ranked second overall, exhibiting slightly lower sensitivity but comparable specificity, while GPT-5 showed moderate degradation across all metrics, particularly in recall (46.31%, 95% CI: 42.56–50.40). In contrast, MedGemma-4B displayed the lowest discriminative capacity, consistent with its reduced sensitivity and MCC. Cohen's κ agreement scores confirmed high consistency in model decision patterns, with almost-perfect agreement between GPT-4o and MedGemma-27B (κ = 0.9191), and substantial agreement between all remaining pairs (κ = 0.6659–0.8691). These findings show that although MedGemma-27B and GPT-4o were statistically superior, all models tended to produce similar judgment trends, with disagreement mostly concentrated in borderline ILO categories. Statistical analysis is shown in Table 2.

4. Discussion

Crucially, these results reaffirm that the ultimate diagnostic responsibility and clinical judgment must remain with the occupational physician. AI systems should be viewed strictly as supportive tools

Table 2. Statistical comparison between generative AI models on ILO pneumoconioses classification performance.

Model Comparison	McNemar χ^2	p-value	Cohen's κ
GPT-4o vs GPT-5	93.582	<0.001	0.8691
GPT-4o vs MedGemma-27B	11.753	<0.001	0.9191
GPT-4o vs MedGemma-4B	245.263	<0.001	0.7002
GPT-5 vs MedGemma-27B	67.189	<0.001	0.8484
GPT-5 vs MedGemma-4B	33.914	<0.001	0.7101
MedGemma-27B vs MedGemma-4B	141.103	<0.001	0.6659

designed to assist, rather than replace, human expertise in occupational health practice. While models were prompted to follow the structured logic of the ILO classification, we did not evaluate internal reasoning pathways; no claims are made regarding their ability to reproduce human cognitive mechanisms.

Recent studies have explored automated detection of pneumoconiosis using deep learning applied to chest radiographs. Conventional CNN-based approaches have reported strong performance, typically achieving accuracy between ~90% and 98% or AUC values above 0.90 in binary detection tasks and simplified multi-class staging [19,20,34]. Advanced attention-based architectures have further improved feature extraction from lesion-specific regions[35]. Earlier work relying on handcrafted features demonstrated feasibility but was limited by small datasets and non-standard classification schemes [36]. Hybrid pipelines combining lung segmentation with classical machine-learning classifiers showed promising performance when incorporating ILO guidance, although they did not reproduce full ILO scoring [22]. In parallel, early initiatives using LLM-based strategies for pneumoconioses imaging suggest emerging opportunities for multimodal AI in occupational radiology [24], while other studies emphasize the need for robust, scalable AI solutions, especially in low-resource settings [37].

In contrast to these studies, our work is, to our knowledge, the first to systematically evaluate generative multimodal AI systems in a zero-shot setting using the complete revised 2022 ILO classification. This approach reflects realistic clinical deployment scenarios, in which pre-trained models are used without re-training. Certified B Readers remain the gold standard for radiographic diagnosis of pneumoconioses according to ILO guidelines. The NIOSH certification examination using a digital set (revised 2022) requires candidates to classify 72 chest radiographs within four hours covering the full range of technical quality and pneumoconiotic findings, with performance compared against expert reference standards. Despite formal certification, substantial intra- and inter-reader variability has been documented among B Readers [38]. Such variability reflects the perceptual and cognitive demands of applying the full ILO rubric, particularly in borderline profusion grades, subtle pleural abnormalities, and distinction between dust-related and non-occupational changes. As highlighted by the Italian experience with ILO certification courses [39] and consistent with NIOSH data, the mean passing rate for initial B Reader certification between 1987 and 2018 was only about 40% [7]. This confirms the intrinsic perceptual and cognitive complexity of radiographic diagnosis according to the ILO classification system. In this context, the balanced accuracy observed for the AI models in this study, although obtained on a smaller dataset of 82 images, may approximate the performance typically achieved by less experienced human readers. Variability across investigated models may also parallel the inter-operator variability observed among human readers. By potentially reducing inter-reader variability and serving as a 'second reader' or prescreening tool, such AI applications could ultimately enhance the protection of worker health through more consistent and timely disease detection in occupational surveillance programs [7].

4.1. Limitations and Future Perspectives

This pilot study has several limitations that should be acknowledged. The relatively small dataset and the reliance on standardized training images may limit the generalizability of results. Differences in image quality as well as variability in resolution and inspiratory phase, could have influenced both human and AI interpretations. From an algorithmic standpoint, generative AI systems may exhibit hallucinations or reasoning biases depending on their training data which likely constrained diagnostic precision. Methodologically, each image was analysed through a standardized prompt and the 2022 ILO guidelines, simulating a structured B Reader workflow. Although this ensured consistency, future research could test alternative paradigms such as interactive or iterative readings that simulate real-time reasoning. Furthermore, future evaluations should aim to reproduce the operational conditions of the official NIOSH B Reader certification examination, which historically required candidates to classify

approximately 72 chest radiographs within four hours [7]. Replicating these temporal and quantitative parameters would allow a more realistic benchmarking of multimodal AI systems against human performance, assessing not only diagnostic accuracy but also efficiency and cognitive consistency under standardized testing constraints. Expanding the evaluation to additional multimodal architectures beyond GPT and MedGemma could also clarify whether diagnostic reliability depends more on model design or prompt structure. Statistical power was limited by the small sample size, which may have reduced the ability to detect subtle between-model differences. Additionally, no human B Reader re-assessment of the reference labels was performed; the study relied on the original NIOSH-certified readings as the diagnostic gold standard. While this approach reflects real-world reference conditions, subsequent work should include dual-reader adjudication to evaluate model-human agreement in parallel with human-human reproducibility. Further research should include larger and more heterogeneous datasets, explore domain-specific fine-tuning, and assess intra- and inter-reader variability to benchmark AI reproducibility against human performance. In this context, multimodal generative AI could become a valuable tool for supporting training, pre-screening, and comparative research, contributing to greater consistency and harmonization in the radiological diagnosis of pneumoconioses.

5. Conclusion

The ILO Classification of Pneumoconioses and the NIOSH B Reader Program represent decades of coordinated international efforts to standardize occupational lung disease surveillance. While the B Reader certification program has long served to enhance standardization and reduce inter-reader variability, its effectiveness remains incomplete - particularly in complex or borderline cases where subjective judgment plays a substantial role, underscoring the need for transparent and impartial diagnostic processes. In this context, AI integration holds promise not only for improving diagnostic accuracy but also for reducing variability and increasing efficiency across large-scale screening programs.

This study presents the first evaluation of generative AI applied to the 2022 ILO International Classification of Radiographs of Pneumoconioses. By testing four LLMs (GPT-4o, GPT-5, MedGemma-4B, and MedGemma-27B), the research explored the feasibility of using multimodal AI within a standardized diagnostic framework. Although current performance remains limited, the results suggest that these models may begin to approximate the structured reasoning process underlying the ILO system. These findings highlight both the potential and the current limitations of multimodal AI in radiology applied to occupational medicine, particularly in the diagnostic assessment of pneumoconioses and provide a methodological basis for future work aimed at improving accuracy, consistency, and interpretive transparency. As AI systems continue to evolve, the most promising future lies in a human-AI interface – where technology augments the expertise of B Readers, improving both efficiency and consistency while retaining essential human oversight and clinical responsibility. Achieving this vision will require sustained collaboration among AI developers, occupational health professionals, regulatory authorities, and worker communities to ensure that these powerful tools fulfill their ultimate purpose: safeguarding worker health through timely and accurate disease detection, firmly grounded in the principles of occupational medicine, without neglecting ethical aspects.

Funding: This research received no external funding.

Acknowledgments: Intel Corporation Mindshare Program 2024.

Declaration of Interest: The authors declare no conflict of interest.

Author Contribution Statement: A.B. designed the study and reviewed the manuscript, M.P. carried out statistical analysis, drafted and reviewed the manuscript, A.P. designed the AI system and carried out the experiment, A.Q. and R.L. carried out classification benchmarks, S.D. retrieved datasets, N.M. reviewed the manuscript, and V.T. reviewed the literature.

References

1. Hou X, Wei Z, Jiang X, et al. A comprehensive retrospect on the current perspectives and future prospects of pneumoconiosis. *Front Public Health*. 2025;12:1435840.
2. Naghavi M, Hmwe Hmwe K, Bhoomadevi A, et al. Global burden of 292 causes of death in 204 countries and territories and 660 subnational locations, 1990–2023: a systematic analysis for the Global Burden of Disease Study 2023. *The Lancet*. 2025, vol. 406.
3. Zhang JS, Xiong X, Ruan, et al. Global burden of pneumoconiosis from 1990 to 2021: a comprehensive analysis of incidence, mortality, and socio-demographic inequalities in 204 countries and territories. *Front Public Health*. 2025;13:1579851.
4. Matyga AW, Chelala L, Chung JH. Occupational Lung Diseases: Spectrum of Common Imaging Manifestations. *Korean J Radiol*. 2023;24:8.
5. Guidelines for the use of the ILO International Classification of Radiographs of Pneumoconioses Revised edition 2022. Available online at: https://www.ilo.org/resource/ilo-international-classification-radiographs-pneumoconioses-1 (Last Accessed 3-10-25).
6. Morgan RH. Proficiency examination of physicians for classifying pneumoconiosis chest films. *AJR Am J Roentgenol*. 1979;132(5).
7. Halldin, CN, Hale JM, Weissman, et al. The National Institute for Occupational Safety and Health B Reader Certification Program – An Update Report (1987 to 2018) and Future Directions. *J Occup Environ Med*. 2019;61(12):1045-1051.
8. Lewis MA. Charles Babbage: Reclaiming an operations management pioneer. *J Oper Manag*. 2007;25(2):248-259.
9. Grattan-Guinness I. Charles Babbage as an Algorithmic Thinker. *IEEE Ann Hist Comput*. 1992;14(3):34-48.
10. Strawn G. Masterminds of Punched Card Data Processing: Herman Hollerith and John Billings. *IT Prof*. 2023;25(6):90-93
11. Ziv L, Nakash M. Behind the Algorithm: International Insights into Data-Driven AI Model Development. *Mach Learn Knowl Extr*. 2025;7(4):122.
12. Y. Hosoda. ILO international classifications of radiographs of pneumoconioses – Past, present and future. *International Classification of HRCT for Occupational and Environmental Respiratory Diseases*, 2005.
13. Muszyńska-Graca M, Dąbkowska B, Brewczyński, PZ. Guidelines for the use of the International Classification of Radiographs of Pneumoconioses of the International Labour Office (ILO): Substantial changes in the currrent edition. *Med Pr*. 2016;67(6):833-837.
14. Buess L, Keicher M, Navab N, et al. From large language models to multimodal AI: A scoping review on the potential of generative AI in medicine. *Biomed Eng*. 2025;15:845–863.
15. Nam Y, Kim DY, Kyung S, et al. Multimodal Large Language Models in Medical Imaging: Current State and Future Directions. *Korean J Radiol*. 2025;26(10):900-923.
16. Soni N, Ora M, Agarwal A, et al. A Review of the Opportunities and Challenges with Large Language Models in Radiology: The Road Ahead. *AJNR Am. Neuroradiol*, 2025;46(7):1292-1299.
17. Sun W, Wu D, Luo Y, et al. A Fully Deep Learning Paradigm for Pneumoconiosis Staging on Chest Radiographs. *IEEE J Biomed Health Inform*. 2022;26(10):5154-5164.
18. Zheng R, Deng K, Jin H, et al. An improved CNN-based pneumoconiosis diagnosis method on X-ray chest film. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2019.
19. Zhang L, Rong R, Li Q, et al. A deep learning-based model for screening and staging pneumoconiosis. *Sci Rep*. 11,2201(2021).
20. Zhang Y, Zheng B, Zeng F, et al. Potential of digital chest radiography-based deep learning in screening and diagnosing pneumoconiosis: An observational study. *Medicine*. 2024;103(25):e38478.
21. Yang F, Tang ZR, Chen J, et al. Pneumoconiosis computer aided diagnosis system based on X-rays and deep learning. *BMC Med Imaging*. 2021;(1):2201.
22. Hanampa V, Astete J, Castaneda B , Romero S. Diagnosis of Pneumoconiosis with Machine Learning. in *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2024.
23. Li X, Liu CF, Guan L, et al. Deep Learning in Chest Radiography: Detection of Pneumoconiosis. *Biomed Enviro Sci*. 2021;34(10):842-845.
24. Song M, Wang J, Yu Z, et al. PneumoLLM: Harnessing the power of large language model for pneumoconiosis diagnosis. *Med Image Anal*. 2024;97:103248.
25. Tian D, Jiang S, Zhang L, Lu X, Xu Y. The role of large language models in medical image processing: a narrative review. *Quant Imaging Med Surg*. 2024,14(1), 1108-1121.
26. Akinci D'Antonoli T, Stanzione A, Bluethgen C, et al. Large language models in radiology: fundamentals, applications, ethical considerations, risks, and future directions. *Diagn Interv Radiol*. 2024,30(2):80-90.
27. Lanzafame LRM, Gulli C, Mazziotti S, et al. Chatbots in Radiology: Current Applications, Limitations and Future Directions of ChatGPT in Medical Imaging. *Diagnostics*. 2025;15(13):1635.
28. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need, *Advances in Neural Information Processing Systems*, 2017.
29. OpenAI, Achiam J, Adler S, et al. GPT-4 Technical Report. 2024 Available online at: http://arxiv.org/abs/2303.08774 (Last Accessed on 20-10-25).
30. Vilakati S. Prompt engineering for accurate statistical reasoning with large language models in medical research. *Front Artif Intell*. 2025;8:1658316.

31. Devnath L, Fan Z, Luo S, et al. Detection and Visualisation of Pneumoconiosis Using an Ensemble of Multi-Dimensional Deep Features Learned from Chest X-rays. *Int J Environ Res Public Health*. 2022;19(18):11193.
32. Holzmann H, Klar B. Robust performance metrics for imbalanced classification problems. *arXiv preprint*. 2024,2404.07661.
33. Brodersen KH, Ong CS, Stephan KE, Buhmann JM. The Balanced Accuracy and Its Posterior Distribution. *20th International Conference on Pattern Recognition*, 2010, pp. 3121-3124.
34. Li X, Xu M, Yan Z, et al. Deep convolutional network-based chest radiographs screening model for pneumoconiosis. *Front Med.* 2024:11:1290729.
35. Alam MS, Wang D, Sowmya A. DLA-Net: dual lesion attention network for classification of pneumoconiosis using chest X-ray images. *Sci Rep*, 2024;14:11616.
36. Okumura E, Kawashita I, Ishida T. Computerized Classification of Pneumoconiosis on Digital Chest Radiography Artificial Neural Network with Three Stages. *J Digit Imaging*. 2017;30(4):413-426.
37. Akhter Y, Ranjan R, Singh R, et al. On AI-Assisted Pneumoconiosis Detection from Chest X-rays. *IJCAI International Joint Conference on Artificial Intelligence*, 2023.
38. Halldin CN, Blackley DJ, Petsonk EL, Laney AS. Pneumoconioses radiographs in a large population of U.S. coal workers: Variability in a reader and B Reader classifications by using the international labour office classification. *Radiology*. 2017;284(3):870-876.
39. Leonori R, Cardona E, Napoli G. Risultati di un'esperienza di formazione sulle linee guida ilo per le pneumoconiosi, *Giornale Italiano Di Medicina Del Lavoro Ed Ergonomia*, 2025, vol. XLVII.

Mattioli 1885