



Automated engineered-stone silicosis screening and staging using Deep Learning with X-rays

Blanca Priego-Torres ^{a,b} , ^{*}, Daniel Sanchez-Morillo ^{a,b} , Ebrahim Khalili ^{a,b}, Miguel Ángel Conde-Sánchez ^e, Andrés García-Gómez ^d, Antonio León-Jiménez ^{b,c}

^a Bioengineering, Automation and Robotics Research Group, Department of Automation Engineering, Electronics and Computer Architecture and Networks, School of Engineering, University of Cadiz, Puerto Real, 11519, Cádiz, Spain

^b Biomedical Research and Innovation Institute of Cadiz (INIBICA), Puerta del Mar University Hospital, Cádiz, 11009, Spain

^c Pulmonology Department, Puerta del Mar University Hospital, Cádiz, 11009, Spain

^d Radiology Department, Puerta del Mar University Hospital, Cádiz, 11009, Spain

^e Radiology Department, Puerto Real University Hospital, Puerto Real, 11510, Cádiz, Spain

ARTICLE INFO

Keywords:

Silicosis
Deep learning
Chest X-ray
Progressive massive fibrosis (PMF)
Simple silicosis (SS)
Engineered stone

ABSTRACT

Silicosis, a debilitating occupational lung disease caused by inhaling crystalline silica, continues to be a significant global health issue, especially with the increasing use of engineered stone (ES) surfaces containing high silica content. Traditional diagnostic methods, dependent on radiological interpretation, have low sensitivity, especially, in the early stages of the disease, and present variability between evaluators. This study explores the efficacy of deep learning techniques in automating the screening and staging of silicosis using chest X-ray images.

Utilizing a comprehensive dataset, obtained from the medical records of a cohort of workers exposed to artificial quartz conglomerates, we implemented a preprocessing stage for rib-cage segmentation, followed by classification using state-of-the-art deep learning models. The segmentation model exhibited high precision, ensuring accurate identification of thoracic structures. In the screening phase, our models achieved near-perfect accuracy, with ROC AUC values reaching 1.0, effectively distinguishing between healthy individuals and those with silicosis.

The models demonstrated remarkable precision in the staging of the disease. Nevertheless, differentiating between simple silicosis and progressive massive fibrosis, the evolved and complicated form of the disease, presented certain difficulties, especially during the transitional period, when assessment can be significantly subjective. Notwithstanding these difficulties, the models achieved an accuracy of around 81% and ROC AUC scores nearing 0.93.

This study highlights the potential of deep learning to generate clinical decision support tools to increase the accuracy and effectiveness in the diagnosis and staging of silicosis, whose early detection would allow the patient to be moved away from all sources of occupational exposure, therefore constituting a substantial advancement in occupational health diagnostics.

1. Introduction

Inhalation of silica is a primary cause of occupational respiratory illness globally [1]. Silicosis occurs after prolonged exposure to respirable crystalline silica, progressing in some cases, from simple silicosis (SS), the simplest clinical form of the disease, into progressive massive fibrosis (PMF), respiratory insufficiency, and, in severe cases, fatality. Historically associated with mining, quarrying, tunneling, and similar industries, silicosis has seen a decline in high-income countries due to

improved surveillance and occupational health measures, yet it remains a significant global health challenge [2].

In the 1990s, artificial quartz or silica conglomerates, also known as engineered stone or artificial stone (ES), emerged on the global market. These materials were developed as affordable and versatile alternatives to natural stone surfaces like marble and granite. Over time, ES gained popularity for its durability, stain resistance, and aesthetic diversity, becoming a common choice in construction and

^{*} Corresponding author at: Bioengineering, Automation and Robotics Research Group, Department of Automation Engineering, Electronics and Computer Architecture and Networks, School of Engineering, University of Cadiz, Puerto Real, 11519, Cádiz, Spain.

E-mail address: blanca.priego@uca.es (B. Priego-Torres).

<https://doi.org/10.1016/j.combiomed.2025.110153>

Received 4 November 2024; Received in revised form 9 March 2025; Accepted 4 April 2025

Available online 18 April 2025

0010-4825/© 2025 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC license (<http://creativecommons.org/licenses/by-nc/4.0/>).

interior design applications worldwide, including kitchens, bathrooms, flooring, and architectural embellishments.

ES compositions often contain over 85% silica, predominantly in the form of quartz or cristobalite [3], significantly higher than the silica content in natural stone counterparts. Consequently, silicosis caused by ES exhibits distinct characteristics compared to traditional mining-induced silicosis. Notably, the latency period between exposure and disease onset is shorter, often affecting younger individuals. Moreover, the disease progression is aggressive, with a considerable proportion of cases presenting in advanced stages [4], such as complicated silicosis or PMF. Even after cessation of exposure, the disease may rapidly progress, underscoring the urgency of early diagnosis and intervention [5].

Currently, silicosis diagnosis relies heavily on radiological findings, occupational history, and exclusion of alternative diagnoses, with chest radiography being the cornerstone for diagnosis and classification. However, the subjective nature of radiological interpretation contributes to intra- and interobserver variability [6], posing challenges in accurate diagnosis and staging.

Given the increasing number of exposed workers and the exponential rise in diagnosed cases [2], particularly in early disease stages, there is a pressing need for simplified and cost-effective diagnostic systems. In this context, the development of artificial intelligence (AI)-based approaches holds immense promise [7].

Recently, AI has achieved significant success in medical imaging tasks [8]. As a subset of AI, machine learning is utilized to predict the extent of histological tissue invasion and patient survival in lung cancer [9], classify pulmonary nodules [10], or diagnose and categorize COVID-19 [11]. However, research on the topic of silicosis is limited, leading to a lack of AI tools specifically designed for diagnosing this condition. This limitation stems from the complexity of silicosis and the limited availability of well-annotated chest X-rays for training AI models to accurately identify the disease. The state of the art on this topic reflects that while the diagnosis of pneumoconiosis has been primarily addressed [12–21], the staging has been marginally approached [12,16,17,20–22].

The most common procedure for binary disease classification (healthy or pneumoconiosis patients) or stratification into stages (0, I, II, III) established by the International Labour Organization (ILO) divides processing into two stages: first, lung delineation using a deep-learning architecture for automatic segmentation, followed by image resizing [12,14,16,17,21,22], with some studies additionally dividing each lung field vertically into three subfields for finer analysis [12,21,22]. The second stage consists of automatic classification, based on variations of pre-trained deep-learning models from the literature [12,14,16,17,21], through multi-scale feature mapping extracted from different layers in the deep-learning architecture [22], or a multistage classification approach using a pre-trained classification model [21]. Other studies, however, classify the images from the outset using models from the literature without prior segmentation, relying on pre-trained deep-learning networks directly [15,19,20], an ensemble of classifiers [13], or a pre-trained classification model followed by CNN-based feature aggregation methods or SVM-based classification [13].

The first strategy, however, presents two main challenges: first, it necessitates reliance on a sufficiently accurate lung segmentation algorithm to ensure it does not adversely affect the classifier's performance. Training a robust lung segmentation model requires a vast dataset of images. Most datasets found in the literature lack representative subsets of this disease, which, in its advanced stages, can produce severely affected lung areas, impairing respiratory function. These rare imaging patterns can lead to segmentation errors when applied to a silicosis dataset. Secondly, this approach assumes that all non-lung tissue is irrelevant in the screening and staging of the disease. However, in the interpretation of X-ray images for diagnosing silicosis caused by artificial stone, it is crucial to evaluate not only

the strictly pulmonary region but also other tissues and structures, such as the mediastinum. Surrounding areas such as the heart, pleura or large blood vessels can provide additional relevant information, as alterations may indicate complications associated with silicosis or other coexisting conditions [23].

Regarding the origin of the datasets used in the studies, all works describe the images as collected from hospitals in China, except for two studies, in which the hospitals providing the images are based in Australia [13,18]. Concerning the types of pneumoconiosis present in the images comprising the datasets, most studies do not specify this information. Exceptions include the study by [14], which contains cases of silicosis and coal workers' pneumoconiosis (CWP), as well as the studies by [13,18,22], which include only cases of CWP. The title of the study by [19] introduces some confusion, as it explicitly references a diagnosis of silicosis, despite using the same dataset as [15]. This dataset was previously publicly available, and the initial study describing it did not specify the particular type of pneumoconiosis affecting the included patients.

In the study by Sharma et al. [19], silicosis screening is indicated, achieving an accuracy of 81%. They utilized a dataset consisting of 706 chest X-ray images, including 114 with silicosis and 452 normal images. The study employed state-of-the-art classifiers, enhancing the models with multiple output layers and considering the entire chest X-ray image for classification. Similarly, as observed in this study, one of the most common limitations encountered in the referenced studies stems from the imbalance within chest X-ray image datasets. Consequently, the majority of these investigations solely focus on detection models, overlooking stratification due to the uneven distribution of pneumoconiosis/silicosis cases across stages.

In the realm of silicosis diagnosis, it is imperative to discern between SS, characterized by a diffuse and bilateral nodular pattern, with greater involvement of the upper lobes and posterior areas of the lung, and PMF, defined by the presence of larger diameter opacities in a process of nodule conglomeration, parenchymal retraction, and cicatricial emphysema. This differentiation poses a particular challenge, more complex than in other types of pneumoconiosis, such as coal worker's pneumoconiosis, due to several specific factors. First, the rapid progression of disease is notably more aggressive in artificial stone silicosis compared to other forms of silica or coal dust exposure. Patients can develop within a few years of initial exposure, even if the patient has ceased exposure. This pattern of rapid advancement is rarely seen in other forms of pneumoconiosis and is especially concerning in younger patients who face the progression of PMF shortly after diagnosis. Radiologically, the shift from SS to PMF is subtle and progresses in brief intervals, often presenting with atypical patterns and large opacities on radiographs. These atypical radiological signs can obscure differentiation between consolidating lesions and advanced fibrosis, complicating accurate interpretation. Clinically, distinguishing between SS and PMF is critical, as PMF is associated with a poorer prognosis and more severe respiratory complications. While the International Labour Organization stages (0, I, II, III) provide a useful framework for stratification, monitoring and classifying disease progression specifically from SS to PMF in artificial stone silicosis yields crucial insights for treatment decisions. Patients with PMF often require more intensive monitoring and may be candidates for advanced interventions, including oxygen therapy or, in severe cases, lung transplantation.

To address this need, this study offers several important contributions to the field of silicosis diagnosis:

- Unlike previous articles that primarily discussed pneumoconiosis in general, this article specifically examines how AI models contribute to the diagnosis and stratification of silicosis caused by artificial stone. This focus is particularly valuable, as silicosis presents distinct features and progression compared to other types of pneumoconiosis.

- We leverage a unique dataset from a hospital in southern Spain, comprising chest X-ray images acquired from the same patients at multiple time points and displaying varying degrees of silicosis severity. Specifically, the dataset includes 285 images representing simple silicosis (SS) and 212 images depicting progressive massive fibrosis (PMF), all curated from medical records collected over several years. Although our current study does not conduct a detailed longitudinal analysis, the repeated imaging of individual patients provides the opportunity to observe trends in disease progression. This design not only offers a robust foundation for training and evaluating deep learning-based classification models but also hints at the potential for further insights into the temporal evolution of silicosis in future studies.
- We introduce a preprocessing stage for X-ray images to isolate the region of interest, thereby enhancing the accuracy of the subsequent image processing steps for both screening and staging. This stage involves rib-cage segmentation using a deep learning-based algorithm, which ensures consistent delineation of the rib cage and minimizes noise, maintaining the visibility of critical regions such as the heart and mediastinum. Unlike most other studies, which often employ less specific segmentation, such as isolating only the lung area, our approach allows for a more refined focus on regions impacted by silicosis. Additionally, this article provides an analysis with detailed segmentation metrics, including mAP50 and mAP50-95, as well as comprehensive classification metrics (precision, sensitivity, specificity, AUC, F1-Score) for each stage of silicosis.
- We test state-of-the-art algorithms on our dataset to assess the potential of a complete pipeline for the automatic screening and staging of silicosis. By employing advanced deep learning techniques and transfer learning approaches, we aim to refine these models to improve diagnostic accuracy and reliability, ultimately contributing to better occupational health outcomes.
- We addressed the challenge of model interpretability in medical imaging by using the explainable artificial intelligence techniques GradCAM and GradCAM++ to visualize the decisions of one of the models with the best performance results. We collaborated with two radiologists and a pulmonologist from our research team. They evaluated the heatmaps generated by GradCAM and GradCAM++ for their correlation with pulmonary lesions using a Likert scale, providing relevant insights into the model's decision-making process. This deep approach to interpretability is a relatively unexplored feature in other studies, which primarily report performance metrics without a detailed analysis of the key image regions influencing model decisions. In cases where other studies do present heatmaps for explainability, they generally limit themselves to displaying them without validation from medical experts.

2. Material and methods

2.1. Participants and datasets

A total of 101 subjects agreed to participate in this study. All patients were males with a history of working in polishing, cutting, and finishing engineered stone countertops. They were all diagnosed at the Pulmonology and Radiology Units of the Puerta del Mar and Puerto Real University Hospitals in Cádiz, Spain.

Clinical, demographic data, and chest X-ray images were collected from hospital records retrospectively, from 2009 to 2016, and prospectively from 2017 to 2024.

The diagnosis was determined by reviewing the patient's history of silica exposure and conducting chest radiography and/or High-Resolution Computed Tomography (HRCT), along with, in some cases, lung or mediastinal lymph node biopsy. Radiological findings were identified following the ILO system and the International Classification

of HRCT for Occupational and Environmental Respiratory Diseases (ICOERD) [24]. HRCT confirmed the diagnosis of both PMF and simple silicosis. In the case of PMF, classification was performed according to the ICOERD international classification system, even though large opacities were not evident on simple X-rays.

Sociodemographic data included age at first diagnosis, age at starting exposure, duration of exposure, time since initial exposure to first diagnosis, time since initial exposure to PMF, and time since the end of exposure to PMF (in years). Data from clinical assessments consisted of respiratory function including the forced expiratory volume in 1 s (FEV1), forced vital capacity (FVC), FEV1/FVC ratio, and diffusing capacity of the lung for carbon monoxide (DLCO), determined using the single-breath procedure following global guidelines.

This study adhered to the principles outlined in the Declaration of Helsinki and received approval from the Research Ethics Committee of the province of Cadiz. All study participants provided informed consent.

The dataset for silicosis screening and staging consisted of 786 chest X-ray images in total, including 289 healthy control images, 285 representing SS, and 212 depicting PMF. The 497 images of SS and PMF are proprietary and were curated from medical records.

Given that the patients were consistently monitored by the clinical authors of this study for over 15 years in some instances, the dataset may feature images from the same patient categorized differently based on their disease status or progression (see Fig. 1).

2.1.1. Other datasets

The healthy control cases were sourced from the VinDr-CXR open dataset [25], which contains 18,000 scans in DICOM format annotated by seasoned radiologists. These scans encompass various types of thoracic abnormalities observed in chest radiographs, including images indicating the absence of abnormalities. These particular cases have been chosen as healthy control subjects for this study. The availability of scans in DICOM format facilitated the filtering of images by gender, restricting the selection to those depicting male subjects to maintain consistency with the dataset of patients affected by the disease.

As detailed in the following section, the image classification pipeline begins with a preprocessing stage that focuses on rib-cage segmentation. This crucial initial step ensures the proper extraction of relevant anatomical regions before the images undergo further analysis. The YOLOv8m-seg model was used for this purpose. The training and validation of this model were conducted using the VinDr-RibCXR dataset, a specialized subset of the broader VinDr-CXR dataset. VinDr-RibCXR consists of 245 chest X-ray (CXR) images, meticulously segmented and annotated for rib identification tasks. An expert radiologist manually annotated all 20 ribs in each image, categorizing them as left (L1 to L10) or right (R1 to R10) ribs.

Each image in the dataset provides a clear delineation and labeling of individual ribs. Fig. 2a shows an original X-ray image, while Fig. 2b highlights the segmentation annotations, with each of the 20 ribs labeled in distinct colors for easy differentiation.

2.2. Deep-neural-network-based classification pipeline

The proposed classification pipeline for automated silicosis screening and staging employs deep neural networks and comprises two principal phases: image preprocessing and classification (screening and staging).

The preprocessing phase begins with applying a deep learning-based rib segmentation algorithm. This model ensures consistent delineation of the rib cage across all patient images. By focusing on the region externally bounded by the ribs, the algorithm minimizes noise introduced by varying arm positions, achieving more uniform and reliable images for further analysis. Moreover, this approach maintains the visibility of critical regions, such as the heart and mediastinum, which are important for comprehensive thoracic assessment. After segmenting

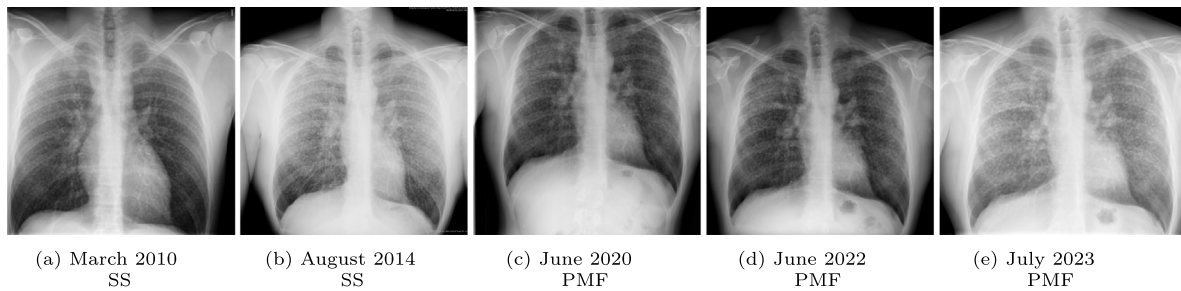


Fig. 1. X-ray images illustrating the progression of a patient from simple silicosis (SS) to progressive massive fibrosis (PMF).

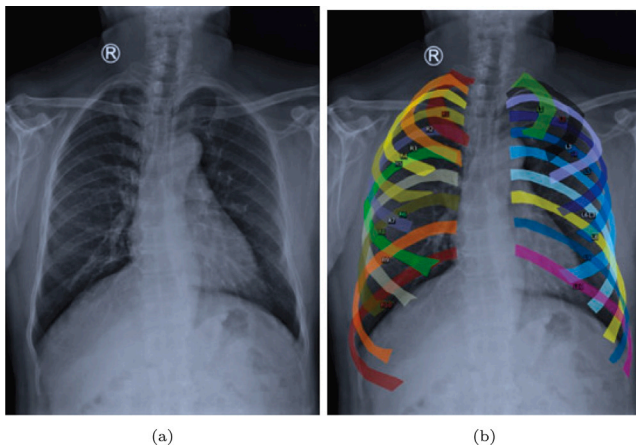


Fig. 2. Example of an X-ray image annotation from the VinDr-RibCXR dataset [26]: (a) original image, (b) original image with the 20 annotated ribs overlaid.

the ribs, the images undergo further processing. This involves tracing a convex hull around the segmented rib regions and removing the background. This step enhances image clarity by eliminating irrelevant information, thereby facilitating more accurate downstream classification. Next, from the segmented and cropped image, an RGB image version is created where each channel contains distinct information: the first channel includes the normalized image values scaled between 0 and 255, the second channel contains the result of applying CLAHE (Contrast Limited Adaptive Histogram Equalization) to the normalized image, and the third channel holds the image after applying gamma correction. This multi-channel approach provides diverse enhanced representations of the original image, enriching the input data for the classification models.

In the classification phase, the refined X-ray images are input into several state-of-the-art deep learning classifiers. These classifiers are re-trained using our dataset through a transfer learning approach. This method leverages pre-trained models and fine-tunes them with our specific dataset, enhancing their performance for silicosis screening and staging. Fig. 3 illustrates the processing pipeline followed.

2.2.1. Preprocessing: rib-cage segmentation and image masking

To achieve segmentation of the rib cage, we employed YOLOv8m-seg, a modified version of the YOLO (You Only Look Once) object detection architecture [27]. Recent studies have demonstrated that YOLO-based models markedly enhance the detection and localization of critical abnormalities in chest X-rays [28,29]. YOLOv8m-seg extends the capabilities of the original YOLO model by integrating advanced segmentation features while retaining its inherent speed and efficiency. This model utilizes a convolutional neural network (CNN) backbone for robust feature extraction, followed by a segmentation head that generates detailed masks for each detected object, thereby allowing for the precise delineation of rib structures within the thoracic cavity. Our

fine-tuned version was trained on a comprehensive dataset of thoracic images annotated specifically to highlight the rib structures.

Transfer learning was employed to leverage pre-trained models and expedite the training process. The YOLOv8m-seg model was initialized with weights pre-trained on a large-scale dataset. This approach enables the model to benefit from the extensive feature extraction capabilities developed from the pre-training phase, providing a robust starting point for the specific task of rib segmentation. Furthermore, various data augmentation techniques were applied to enhance the dataset and improve the model's ability to generalize to unseen data.

Once the ribs are segmented by YOLOv8m-seg, a convex hull is traced around the segmented regions. Subsequently, the background is removed and the final image is cropped to isolate the area of interest. Then, a multi-channel image is created to enhance the visual clarity of the processed image. This multi-channel image includes the first channel with the normalized image values scaled between 0 and 255, the second channel with CLAHE [30] applied to the normalized image using a clip limit factor of 2, significantly improving the contrast and making subtle features more discernible, and the third channel with gamma correction applied to the normalized image with a gamma value of 1.5, enhancing the visibility of details in both darker and brighter regions. These enhanced images focus solely on the rib cage and are used for subsequent classification tasks aimed at screening and staging silicosis. This approach ensures that only the relevant regions are analyzed, improving the accuracy and reliability of the classification process.

2.2.2. Classification: Silicosis screening and staging

In the classification phase for silicosis screening and staging, we utilized four convolutional neural network (CNN) architectures: DenseNet121 [31], InceptionV3 [32], MobileNetV2 [33], and ResNet152V2 [34]. These models were selected for their high performance in medical imaging and their ability to detect intricate patterns in chest X-ray images.

DenseNet121 employs a dense connectivity pattern where each layer is directly connected to every other layer in a feed-forward manner. This structure enhances feature propagation and reduces the number of parameters, making it highly efficient for deep learning tasks [31].

InceptionV3 is designed with inception modules that handle different scales of image regions, allowing for effective multi-scale feature capture [32].

MobileNetV2 is optimized for mobile and edge device applications, utilizing depthwise separable convolutions and inverted residuals to maintain high accuracy with reduced computational complexity [33].

ResNet152V2, an advanced variant of the Residual Network (ResNet), incorporates skip connections to alleviate the vanishing gradient problem, facilitating the training of very deep networks [34].

The depth of all models enables them to capture detailed features in chest X-rays, essential for distinguishing between different stages of silicosis.

PREPROCESSING: RIBCAGE SEGMENTATION AND IMAGE MASKING

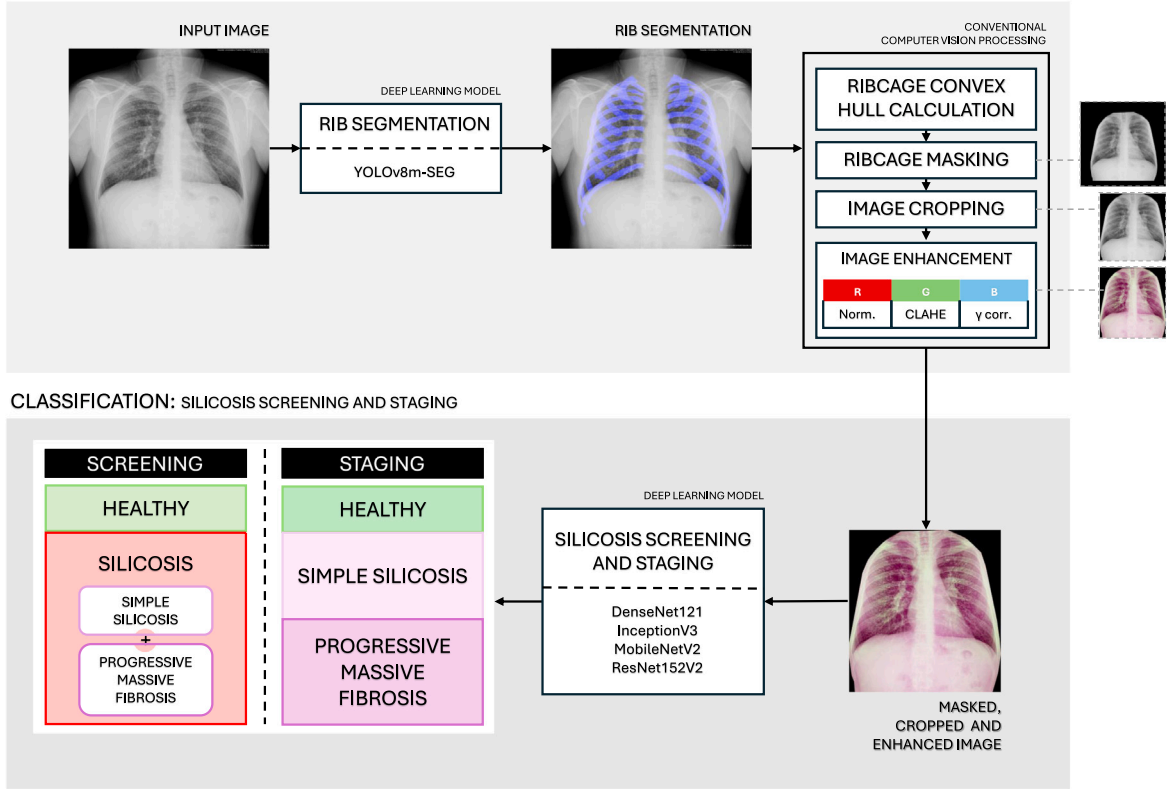


Fig. 3. Processing pipeline followed in the classification of automated silicosis screening and staging.

2.3. Ablation study: Evaluating preprocessing techniques

This study was conducted to assess the impact of different preprocessing techniques on classification performance. Four image preprocessing configurations were evaluated: (B) the baseline grayscale image with normalization, (M1) the image preprocessed with rib-cage segmentation, (M2) the full image processed in multiple channels incorporating normalization, CLAHE, and gamma correction, and (M3) the multi-channel image with rib-cage segmentation. The models tested (DenseNet121, InceptionV3, MobileNetV2, and ResNet152V2) were trained using an 80%–20% train-validation split.

2.4. Evaluation metrics

2.4.1. Segmentation metrics

The performance of the rib segmentation model based on YOLOv8m-seg was assessed using several metrics: precision, recall, mean average precision at 50% intersection over union (mAP_{50}), and mean average precision across multiple intersections over union thresholds (mAP_{50-95}), for both mask and bounding boxes.

Intersection over Union (IoU) is a metric used to evaluate the accuracy of an object detector on a particular dataset. It measures the overlap between the predicted bounding box or mask and the ground truth bounding box or mask. The IoU is defined as:

$$IoU = \frac{\text{Area of Overlap}}{\text{Area of Union}} \quad (1)$$

where the Area of Overlap is the area of intersection between the predicted bounding box or mask and the ground truth bounding box or mask, and the Area of Union, is the total area covered by both the predicted and ground truth bounding boxes or masks.

Precision measures the ratio of correctly identified positive samples to all samples detected as positive. It reflects the model's ability to

avoid false positives.

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2)$$

where TP and FP represent true and false positives, respectively.

Recall quantifies the ratio of correctly identified positive samples to all actual positive samples, indicating the model's capacity to capture relevant instances.

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3)$$

where FN denotes false negatives.

Mean Average Precision at 50% IoU (mAP_{50}) is the average precision computed at a single IoU threshold of 50%. It balances precision and recall to provide a single metric for model performance at this threshold. The average precision (AP) for a single class is the area under the precision–recall curve. For multiple classes, mAP_{50} is calculated as:

$$mAP_{50} = \frac{1}{N} \sum_{i=1}^N AP_i \quad (4)$$

where N is the number of classes and AP_i is the average precision for class i .

Mean Average Precision across multiple IoU thresholds (mAP_{50-95}) extends the evaluation to a range of IoU thresholds, typically from 0.5 to 0.95 in increments of 0.05. This provides a more comprehensive evaluation of the model's performance across various levels of strictness.

$$mAP_{50-95} = \frac{1}{|T|} \sum_{t \in T} mAP_t \quad (5)$$

where T is the set of IoU thresholds, typically $T = \{0.5, 0.55, \dots, 0.95\}$, and mAP_t is the mean average precision at IoU threshold t .

Combined, these metrics provide a comprehensive assessment of the segmentation model performance, ensuring accuracy and comprehensiveness in detecting and segmenting rib structures in chest X-rays. Further details on these metrics can be found in [35] and [36].

2.4.2. Classification metrics

To evaluate the performance of the classification models, we employed a comprehensive set of metrics. These include class-wise performance metrics such as precision, recall, and F1-score, and overall performance metrics including accuracy, ROC AUC, PR AUC, and log-loss.

Class-wise Performance Metrics

Precision measures the ratio of correctly predicted positive instances to the total predicted positives. The formula is described in Eq. (2).

Recall, also known as sensitivity, measures the ratio of correctly predicted positive instances to all actual positives. The formula is described in Eq. (3).

The F1-score is the harmonic mean of precision and recall, providing a balance between the two metrics. It is particularly useful when the class distribution is imbalanced.

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (6)$$

Overall Performance Metrics

Accuracy measures the ratio of correctly predicted instances (both positive and negative) to the total instances. It is a straightforward metric to assess the overall correctness of the model.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (7)$$

The Receiver Operating Characteristic (ROC) Area Under the Curve (AUC) metric evaluates the model's ability to distinguish between classes. It plots the true positive rate (recall) against the false positive rate at various threshold settings. The ROC AUC value ranges from 0 to 1, with higher values indicating better performance. The formula for the area under the ROC curve is typically computed using numerical integration methods, such as the trapezoidal rule.

The Precision–Recall Area Under the Curve (PR AUC) measures the trade-off between precision and recall for different threshold values. This metric is particularly useful for imbalanced datasets. Like ROC AUC, PR AUC is calculated using numerical integration.

Log-loss, or logarithmic loss, measures the performance of a classification model by penalizing false classifications. It calculates the uncertainty of predictions based on the predicted probability.

$$\text{Log-Loss} = -\frac{1}{N} \sum_{i=1}^N [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)] \quad (8)$$

where N is the number of instances, y_i is the actual label, and p_i is the predicted probability of the positive class.

2.5. Interpretability of classification models

Interpreting deep learning models in medical imaging is essential for clinical adoption. To this end, we utilized GradCAM (Gradient-weighted Class Activation Mapping) and GradCAM++ to visualize the decision-making process of our model. GradCAM generates heatmaps that highlight the most influential regions of an image used by the model for its classification. It achieves this by computing the gradients of the target class flowing into the last convolutional layer, creating a coarse localization map that indicates the importance of different regions in the image [37]. GradCAM++ builds upon this method by providing more accurate localization, especially useful when dealing with small or overlapping features, by considering the contribution of each pixel more precisely [38].

To validate the interpretability of the deep learning models used in the automatic screening and staging of silicosis, we generated GradCAM and GradCAM++ heatmaps for a subset of 29 chest X-ray images, corresponding to the longitudinal evolution of five patients over time, using the trained MobileNetV2 model. These heatmaps illustrate the areas in the images that contributed most to the model's classification decisions. To assess the clinical relevance of these highlighted regions, two radiologists and a pulmonologist from our research team evaluated the correspondence between the heatmap regions and the actual

locations of pulmonary lesions. This evaluation was conducted using a Likert scale ranging from 0 to 4, where 0 indicated poor alignment and 4 indicated excellent alignment.

To measure the degree of agreement between the assessments made by the evaluators, Kendall's coefficient of concordance (Kendall's W) was estimated. This coefficient assesses the level of agreement among multiple raters by comparing the ranks assigned to each subject. A value of Kendall's W ranges from 0 to 1, where 0 indicates no agreement and 1 indicates perfect agreement. It is particularly useful in cases where multiple evaluators independently rank or score the same set of items, as it quantifies the overall concordance among their rankings.

2.6. Comparison with previous methods

To evaluate the robustness and generalization of our proposed approach, we implemented and tested three previous methods on our dataset, following the methodologies described by Wang et al. [14], Sun et al. [17], and Yang et al. [16]. These methods were selected because they incorporate a preprocessing stage (specifically, lung segmentation) and are applicable to our dataset, allowing for a fair comparison with our pipeline.

Wang et al. [14] applies lung segmentation using U-Net to extract the regions of interest. The lung fields are masked, and the image is cropped to encompass only the pulmonary region. The segmented images are then classified using InceptionV3. For the classification, the images are resized to 299×299 pixels. Yang et al. [16] method employs U-Net for lung segmentation and ResNet34 for classification, following a multi-stage pipeline for pneumoconiosis detection and staging. The segmentation stage isolates the lung regions, and in the classification stage, the images are resized to 224×224 pixels. Sun et al. [17] uses ED-Net for lung segmentation, where lung bounding boxes (instead of masks) are extracted, concatenated, and resized to 512×512 pixels. The classification stage employs ResNet18, trained with Label Distribution Learning (LDL) and Focal Staging Loss to improve staging performance.

Given the challenges of lung segmentation, we employed a more advanced segmentation algorithm based on a Pyramid Scene Parsing Network (PSPNet) [39] to improve segmentation accuracy. This decision was made after observing poor segmentation quality when using U-Net on our dataset. By integrating a more refined segmentation model, we ensured that all state-of-the-art methods were given an advantage in the preprocessing stage, thus eliminating potential bias due to segmentation quality that might affect the screening and staging phases. For all these approaches, models were trained and validated using our dataset, ensuring a direct and fair comparison under the same evaluation metrics: accuracy, precision, recall, F1-score, and AUC. An 80%–20% training-validation split with a hold-out validation strategy was applied.

3. Results

3.1. Participants

The mean follow-up time of the patient group was 10.5 ± 3.2 years. Table 1 shows sociodemographic and diagnosis information of the 101 study participants, and details parameters related to the first clinical examination of respiratory function after diagnosis.

In the initial stage of this study, 97 out of the 101 patients had a diagnosis of SS and 4 were diagnosed with PMF. 46.4% of the participants progressed to PMF throughout the study.

Table 1

Sociodemographic and diagnosis information of the participants. Pulmonary function data of patients collected at the first clinical examination of respiratory function after diagnosis. Data are expressed as mean value $\pm$ standard deviation. PMF: progressive massive fibrosis. FEV1: forced expiratory volume in 1 s. FVC: forced vital capacity. DLCO: diffusing capacity of the lung for carbon monoxide.

Variable	Participants (n=101)
Age at first diagnosis (years)	37.0 $\pm$ 7.4
Starting exposure age (years)	20.9 $\pm$ 6.1
Duration of exposure (years)	12.9 $\pm$ 6.1
Time from initial exposure to first diagnosis (years)	15.8 $\pm$ 5.9
Time from initial exposure to PMF diagnosis (years)	20.4 $\pm$ 6.0
Time from end of exposure to PMF (years)	6.5 $\pm$ 4.1
FEV1/FVC	0.80 $\pm$ 0.06
FEV1 (mL)	3554.29 $\pm$ 675.32
FEV1 (%)	88.14 $\pm$ 14.86
FVC (mL)	4428.10 $\pm$ 779.68
FVC (%)	87.86 $\pm$ 14.36
DLCO (mmol/min/kPa)	8.82 $\pm$ 1.78
DLCO (%)	81.42 $\pm$ 18.24

3.2. Experimental setup

For the rib segmentation task, during the training phase, the YOLOv8m-seg model (245 layers, 110.0 GFLOPs) was fine-tuned for chest X-ray segmentation. Images were resized to 600 $\times$ 600 pixels, and a comprehensive data augmentation pipeline was implemented to enhance model generalization. This included scaling with a factor of 0.1, allowing zoom variations between 90% and 110%, and HSV transformations with parameters set to Hue = 0.01, Saturation = 0.5, and Value = 0.3. The model was optimized using the AdamW optimizer (learning rate = 0.0004, momentum = 0.9) and trained over 150 epochs with a batch size of 16. Early stopping and checkpointing were applied to prevent overfitting and retain the best-performing model. Validation was performed using a hold-out procedure with 20% of the dataset reserved for validation.

For the classification task, the DenseNet121, InceptionV3, MobileNetV2, and ResNet152V2 models were initialized with pre-trained weights from ImageNet and fine-tuned for this specific task. Each base model was used as the feature extractor, incorporating global average pooling, dense layers, dropout for regularization, and a softmax output layer on top. The architecture included a dense layer with 1024 neurons, using the ReLU (Rectified Linear Unit) activation function to introduce non-linearity and improve the model's ability to capture complex patterns within the chest X-ray images. Dropout regularization was applied with a probability of 0.5, randomly deactivating 50% of the neurons in the dense layer during training to mitigate the risk of overfitting. The training process involved image resizing to 512 $\times$ 512 pixels and data augmentation techniques to improve model generalization, including rescaling, rotation with a range of 2 degrees, width and height shifts up to 10%, shear transformations up to 10%, and zoom variations up to 10%. Stratified K-Fold cross-validation with 5 splits was applied to ensure robust model evaluation, splitting the dataset into training and validation sets for each fold while preserving class distributions across folds. For the ablation study, a separate 80%–20% split was employed to analyze the impact of different preprocessing techniques on classification performance. The models were compiled using the Adam optimizer [40], with a learning rate set to 0.001, β_1 to 0.9, β_2 to 0.999, ϵ to 1e-07, and AMSGrad set to False. Adam (Adaptive Moment Estimation) combines the benefits of AdaGrad and RMSProp by calculating adaptive learning rates for each parameter and maintaining two moving averages of the gradients and their squares, which are controlled by the hyperparameters β_1 and β_2 , respectively. The loss function used was categorical cross-entropy, and class weights were computed based on the class distribution to mitigate class imbalance.

The class weights were calculated using the following formula to ensure that classes with fewer samples receive higher weights,

thereby reducing the bias towards more prevalent classes during model training.

$$w_{class_i} = \frac{N}{k \times n_{class_i}} \quad (9)$$

where w_{class_i} denotes the weight assigned to class i ; N is the total number of samples in the dataset; k corresponds to the number of distinct classes; and n_{class_i} is the number of samples belonging to class i .

The training process was conducted for 15 epochs, with early stopping applied to prevent overfitting and model checkpointing implemented to save the best-performing model based on validation accuracy. Each model was trained with a batch size of 4. Additionally, the cross-validation splits were stored to ensure consistency and reproducibility across different experimental runs.

3.3. Performance results

3.3.1. Rib cage segmentation

The performance of the YOLOv8m-seg model was evaluated using a validation set of 49 annotated rib images. The model demonstrated high accuracy across various metrics, indicating its effectiveness in segmenting rib structures within chest X-rays. Key performance metrics are summarized in Table 2. The model achieved a precision of 1.00 and a recall of 1.00 for bounding boxes, with a mAP_{50} of 1.00 and a mAP_{50-95} of 0.78. For mask segmentation, the precision and recall were 0.81 and 0.81, respectively, with a mAP_{50} of 0.82 and a mAP_{50-95} of 0.32.

The rib cage was correctly segmented in all dataset images. Fig. 4 presents examples of healthy, SS, and PMF chest X-ray images after the rib structures were accurately segmented using the model trained and validated on the VinDr-RibCXR dataset. Following this segmentation, the rib cage was delineated, and the images were masked based on the segmentation results.

3.3.2. Impact of preprocessing on classification performance

The results of this study are presented in Table 3, showcasing the performance differences across the different preprocessing techniques.

The results highlight that the M1 configuration (rib-cage segmentation applied) consistently achieves the best performance across multiple evaluation metrics. This suggests that segmenting the rib cage enhances classification by focusing on the most relevant anatomical regions while removing background noise. Although multi-channel processing is beneficial in some medical imaging applications, its impact in this specific task (silicosis staging) is marginal at best and may even introduce unnecessary complexity. The slight drop in performance across accuracy, ROC AUC, and PR AUC suggests that segmentation alone (M1) is a more effective preprocessing step than applying additional image enhancement techniques.

For this reason, multi-channel processing (M2 and M3) is not selected for further experimentation, and M1 is used as the final preprocessing method for silicosis classification.

3.3.3. Screening and staging results

For silicosis classification, the DenseNet121, InceptionV3, MobileNetV2, and ResNet152V2 models were evaluated using stratified five-fold cross-validation. The performance metrics are presented in Tables 4 and 5. All models achieved a class-wise precision in the screening task at or above 96%, and a recall and F1-score higher than or equal to 0.93 and 0.96, respectively. Concerning the overall performance metrics, accuracy ranged from 0.97 to 0.99 and ROC AUC was 1.0 in all models, indicating outstanding discriminatory power.

For the staging task, the same four deep-learning architectures were evaluated. The performance metrics are summarized in Tables 6 and 7. Performance varied across classes, with the most significant challenges observed in distinguishing PMF cases. Class-wise precision varied between 0.97 and 0.99 for the healthy cases, 0.69 and 0.75 for the SS

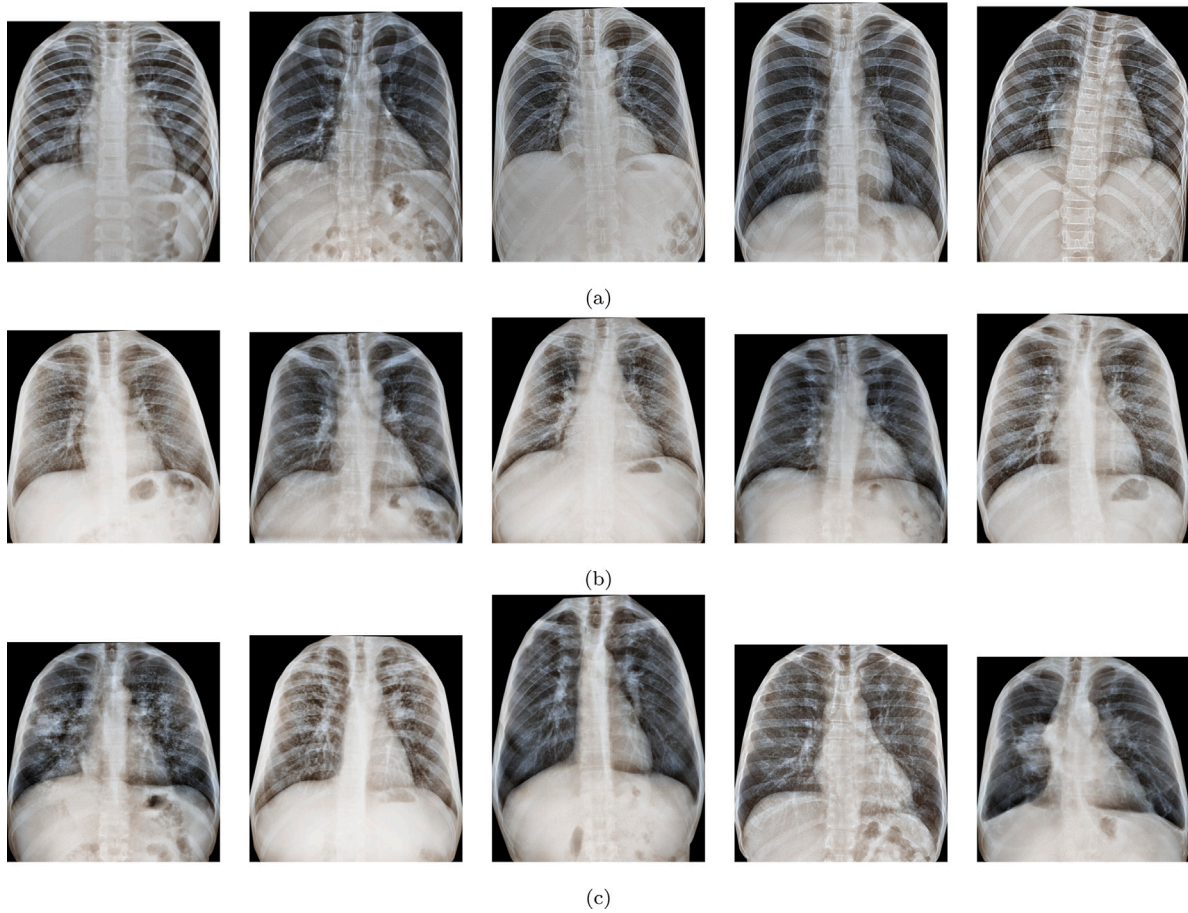


Fig. 4. Samples of X-ray images from the dataset, once they have been preprocessed after the rib-cage extraction phase: (a) healthy samples; (b) simple silicosis samples; (c) progressive massive fibrosis samples.

Table 2
Performance metrics for rib segmentation.

Class	Images	Instances	Box				Mask			
			Precision	Recall	mAP_{50}	mAP_{50-95}	Precision	Recall	mAP_{50}	mAP_{50-95}
all	49	980	1.00	0.98	1.00	0.78	0.81	0.81	0.82	0.32
0	49	49	1.00	1.00	1.00	0.66	0.98	0.98	0.98	0.53
1	49	49	1.00	1.00	1.00	0.67	0.96	0.96	0.98	0.37
2	49	49	1.00	0.97	1.00	0.75	0.87	0.85	0.87	0.32
3	49	49	1.00	1.00	1.00	0.81	0.75	0.76	0.77	0.24
4	49	49	1.00	1.00	1.00	0.82	0.63	0.63	0.66	0.13
5	49	49	1.00	1.00	1.00	0.81	0.69	0.69	0.74	0.18
6	49	49	1.00	1.00	1.00	0.80	0.78	0.78	0.74	0.17
7	49	49	1.00	1.00	1.00	0.79	0.80	0.80	0.76	0.30
8	49	49	1.00	1.00	1.00	0.83	0.82	0.82	0.82	0.40
9	49	49	1.00	1.00	1.00	0.82	0.78	0.78	0.81	0.43
10	49	49	1.00	1.00	1.00	0.67	0.92	0.92	0.95	0.52
11	49	49	1.00	1.00	1.00	0.70	0.94	0.94	0.95	0.42
12	49	49	1.00	1.00	1.00	0.78	0.86	0.86	0.88	0.32
13	49	49	1.00	1.00	1.00	0.79	0.75	0.76	0.76	0.24
14	49	49	1.00	1.00	1.00	0.82	0.73	0.74	0.65	0.17
15	49	49	1.00	1.00	1.00	0.80	0.75	0.76	0.78	0.21
16	49	49	1.00	0.99	1.00	0.77	0.77	0.76	0.79	0.25
17	49	49	1.00	1.00	1.00	0.81	0.86	0.86	0.86	0.36
18	49	49	1.00	1.00	1.00	0.85	0.84	0.84	0.83	0.43
19	49	49	0.98	0.98	1.00	0.78	0.76	0.76	0.78	0.42

Table 3
Ablation study: Impact of image preprocessing on silicosis staging performance.

Model	Overall performance metrics											
	Accuracy				ROC AUC				PR AUC			
	B	M1	M2	M3	B	M1	M2	M3	B	M1	M2	M3
DenseNet121	0.80	0.81	0.79	0.77	0.91	0.91	0.91	0.89	0.85	0.84	0.85	0.79
InceptionV3	0.76	0.82	0.75	0.77	0.89	0.90	0.84	0.89	0.79	0.83	0.72	0.80
MobileNetV2	0.79	0.82	0.79	0.81	0.92	0.93	0.92	0.91	0.84	0.87	0.85	0.83
ResNet152V2	0.81	0.81	0.81	0.78	0.92	0.93	0.92	0.91	0.85	0.87	0.85	0.84

Comparison of overall performance metrics for silicosis staging using different preprocessing techniques. (B) Original grayscale image with normalization; (M1) Rib-cage segmentation applied; (M2) Full image with multi-channel processing (normalization, CLAHE, and gamma correction); (M3) Multi-channel processing with rib-cage segmentation. Results are reported for accuracy, ROC AUC, and PR AUC across different models.

Table 4
Screening class-wise performance metrics.

Model	Silicosis			Healthy		
	Precision	Recall	f1-score	Precision	Recall	f1-score
DenseNet121	0.97	1.0	0.98	0.99	0.95	0.97
InceptionV3	0.96	0.99	0.98	0.98	0.93	0.96
MobileNetV2	0.99	0.99	0.99	0.99	0.98	0.98
ResNet152V2	0.99	0.99	0.99	0.98	0.98	0.98

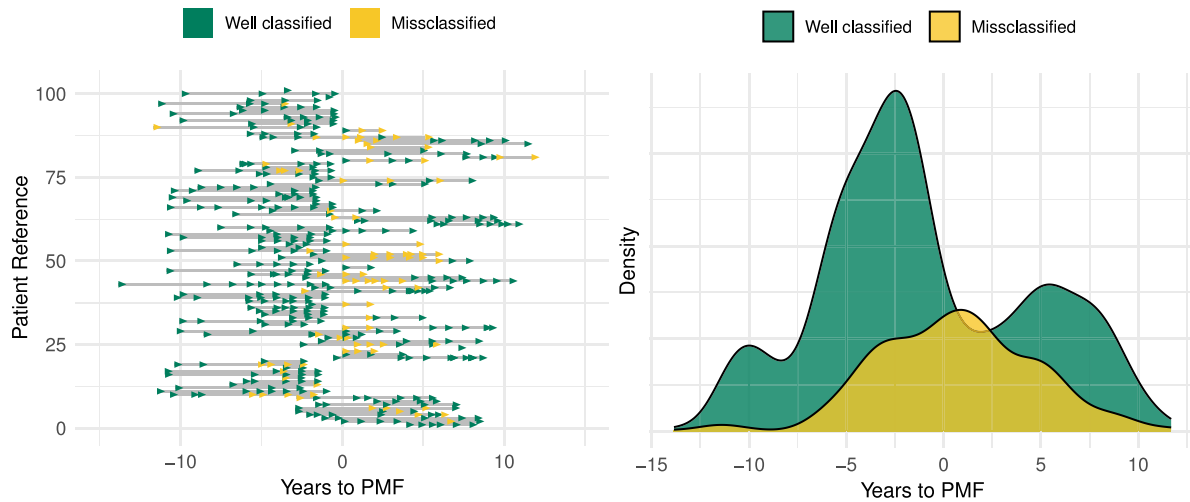


Fig. 5. Analysis of classification errors in distinguishing between simple silicosis (SS) and progressive massive fibrosis (PMF). The graph on the left represents the timeline relative to the PMF diagnosis of patient diagnoses with green triangles indicating correct classifications and red triangles indicating misclassifications; The graph on the right shows the density or frequency of correct and incorrect classifications by the models over the disease progression timeline.

Table 5
Screening overall performance metrics.

Model	Overall performance metrics			
	Accuracy	ROC AUC	PR AUC	Log-loss
DenseNet121	0.98	1.0	0.99	0.08
InceptionV3	0.97	1.0	1.0	0.08
MobileNetV2	0.99	1.0	1.0	0.04
ResNet152V2	0.98	1.0	1.0	0.06

cases, and between 0.73 and 0.82 in the PMF cases. Overall, all models displayed robust performance with accuracy rates around 0.81–0.84 and ROC AUC values ranging from 0.91 to 0.94, as summarized in Table 7. MobileNetV2 exhibited the highest balance between precision and recall, achieving 82% precision and 58% recall for PMF detection, and an overall accuracy of 0.84.

To investigate whether there is an underlying reason for the errors in distinguishing between SS and PMF, we analyzed the classification results considering only the images of patients diagnosed with silicosis.

We took into account the temporal evolution of the patients. Fig. 5 (left panel) shows, on the x-axis, the timeline relative to the patient's diagnosis as PMF. The y-axis represents individual patients, with each fixed value corresponding to a different patient. The x-axis shows the time (in years) between each medical imaging test and the moment the patient was diagnosed with PMF. Each imaging test is represented by a triangle along the timeline. A green triangle indicates a correct classification, by most of the trained classification models, of an X-ray image as SS or PMF when it was evaluated in the validation fold. A red triangle, therefore, indicates an incorrect classification. In Fig. 5 (right panel), again, the x-axis refers to the timeline of the disease progression of the patients, relative to the moment they were diagnosed with PMF. The y-axis shows the density or frequency of correct or incorrect classifications by the algorithms when the images were evaluated in their corresponding validation folds. This analysis shows that most of the incorrect classifications occur around the time when the patient transitions from SS to PMF. The transition between these stages is highly subjective, which is consistent with the validation results of the classifiers.

Table 6
Staging class-wise performance metrics.

Model	Healthy			SS			PMF		
	Precision	Recall	f1-score	Precision	Recall	f1-score	Precision	Recall	f1-score
DenseNet121	0.98	0.95	0.97	0.72	0.78	0.74	0.73	0.65	0.67
InceptionV3	0.98	0.94	0.96	0.69	0.88	0.77	0.81	0.54	0.64
MobileNetV2	0.99	0.98	0.98	0.73	0.89	0.8	0.82	0.58	0.67
ResNet152V2	0.97	0.98	0.97	0.75	0.8	0.77	0.77	0.66	0.69

Expert annotations highlighting key diagnostic features

.....	Minimal small rounded opacities	-----	Small rounded opacities
-----	Small rounded and irregular opacities	-----	Predominant small irregular opacities
-----	Incipient large opacity	-----	Large opacity

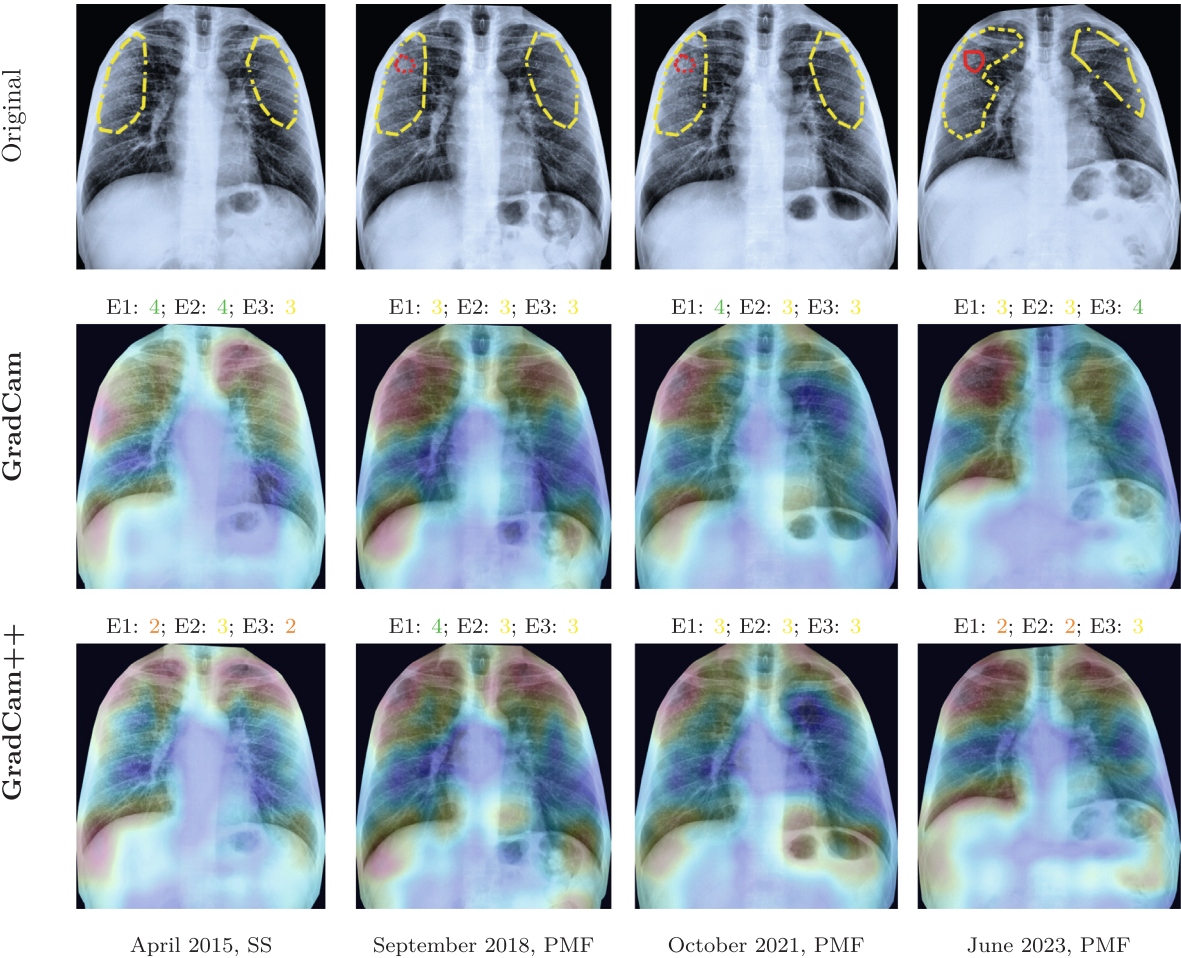


Fig. 6. X-ray images illustrating the progression of a patient from simple silicosis (SS) to progressive massive fibrosis (PMF), along with heatmaps generated by the GradCAM and GradCAM++ models, using the trained MobileNetV2 classification model for image diagnosis. All images were classified correctly by MobileNetV2. Above each heatmap, the score assigned by each evaluator is displayed, on a scale from 0 to 4, where E1, E2, and E3 denote the ratings given by each evaluator, respectively.

Table 7
Staging overall performance metrics.

Model	Overall performance metrics			
	Accuracy	ROC AUC	PR AUC	Log-loss
DenseNet121	0.81	0.92	0.85	0.45
InceptionV3	0.81	0.91	0.83	0.45
MobileNetV2	0.84	0.94	0.88	0.38
ResNet152V2	0.83	0.93	0.86	0.40

3.4. Interpretability results

As previously mentioned, three medical professionals (two radiologists and one pulmonologist) participated in evaluating the correspondence between the lesions identified in the 29 chest X-ray images used as samples for interpretability analysis and the heatmaps generated by the GradCAM and GradCAM++ algorithms when using the pre-trained MobileNetV2 model for classification. Each evaluator rated the alignment of the highlighted regions in the heatmaps with the actual lesions using a Likert scale from 0 to 4.

Table 8
Overview of studies on pneumoconiosis, including year, disease type, classification task, and dataset details.

Ref.	Year	Disease	Task	Dataset (No. images and origin)
[14]	2020	Pneumoconiosis (silicosis, CWP and pneumo. due to low-toxicity inorganic dust)	Screening	Total 1881 (China) 958 normal; 923 pneumo.
[12]	2021	Pneumoconiosis (unspecified)	Screening and Staging (ILO classification)	Total 1216 (China) 704 normal; 236 S-I; 146 S-II; 130 S-III
[22]	2021	Pneumoconiosis (silicosis and CWP)	Staging (ILO classification)	Total 190 (China) 50 S-I; 40S-II; 50 S-III; 50 S-IV
[13]	2021	Pneumoconiosis (CWP)	Screening	Total 153 (Australia) 82 normal; 71 pneumo.
[15]	2021	Pneumoconiosis (unspecified)	Screening	Total 706 (China) 564 normal, 142 pneumo.
[16]	2021	Pneumoconiosis (unspecified)	Screening and Staging (ILO classification)	Total 1760 (China) 512 normal; 352 S-I; 478 S-II; 418 S-III
[17]	2022	Pneumoconiosis (unspecified)	Screening and Staging (ILO classification)	Total 1372 (China) 845 normal; 293 S-I; 132 S-II; 102 S-III
[18]	2022	Pneumoconiosis (CWP)	Screening	Total 153 (Australia) 82 normal; 71 pneumo.
[19]	2024	Pneumoconiosis (unspecified)	Screening	Total 706 (China) 564 normal, 142 pneumo.
[20]	2024	Pneumoconiosis (unspecified)	Screening and Staging (ILO classification)	Total 1250 (China) 318 normal; 322 S-I; 299 S-II; 311 S-III
[21]	2024	Pneumoconiosis (unspecified)	Screening and Staging (ILO classification)	Total 498 (China) 193 normal; 147 S-I; 108 S-II; 50 S-III
Current study	2024	ES silicosis	Screening and Staging (SS or PMF)	Total 786 (Spain) 289 normal; 285 SS; 212 PMF

Abbreviations: CWP: Coal Workers' Pneumoconiosis; pneumo.: Pneumoconiosis; ES: Engineered Stone; ILO: International Labour Organization; S: Stage; SS: Simple Silicosis; PMF: Progressive Massive Fibrosis.

For the GradCAM algorithm, the average scores provided by the evaluators were 2.42, 2.31, and 2.69, respectively. For GradCAM++, the average scores were 2.41, 2.04, and 3.00. When considering the overall mean and standard deviation of the scores from all three evaluators, GradCAM achieved a mean score of 2.47 with a standard deviation of 0.98, while GradCAM++ had a mean score of 2.48 with a standard deviation of 0.92.

For the GradCAM results, the evaluators exhibited a moderate-to-high level of concordance. Kendall's coefficient (W) was 0.642, suggesting that they generally agreed in their evaluations. This level of agreement was statistically significant ($p = 0.00229$), indicating that the observed concordance is unlikely to be due to chance. In the case of GradCAM++, the level of agreement was moderate, with $W = 0.478$. The statistical significance of this concordance was not as strong (p -value = 0.0637), suggesting that, while there is some tendency towards agreement, it is not sufficient to reject the possibility that the observed concordance occurred by chance at a 95% confidence level.

Additionally, notable differences were observed when analyzing the pairwise concordance between the evaluators for both GradCAM and GradCAM++. For GradCAM, the agreement between the two evaluators was the highest, with $W = 0.84$, indicating a very high level of concordance. The lowest concordance was found between one of the evaluators and another specialist, with $W = 0.63$, suggesting a moderate-to-high level of agreement. These results suggest that evaluators with similar expertise tend to agree more closely on the highlighted regions compared to those with different specializations.

Figs. 6 and 7 show the chest X-ray images of two different patients, illustrating their progression from SS to PMF, along with the heatmaps provided by the GradCAM and GradCAM++ models. On each heatmap, the Likert scale rating provided by each evaluator is displayed, where E1, E2, and E3 denote the ratings given by each evaluator, respectively. The first figure presents a sequence where the interpretability results of the MobileNetV2 classification algorithm provided by GradCAM and GradCAM++ are considered favorable by the medical experts, as reflected by their scores, with an average of 3.05 and 2.86 for the two models, respectively. In contrast, the second sequence was selected for

demonstration purposes due to the low scores given by the evaluators. In this particular temporal sequence, the last two X-ray images were misclassified by MobileNetV2 when evaluated within the test sets to assess the robustness of the algorithm through cross-validation. In this case, the average scores of the evaluators for this sequence were 1.61 and 2.28 for GradCAM and GradCAM++, respectively.

4. Discussion

The findings of this study highlight the strong potential of deep learning models for automated engineering-stone silicosis screening and staging. Our method demonstrates advancements in this area compared to previous approaches focused on pneumoconiosis (see Tables 8 and 9). Unlike other studies that limit segmentation to lung fields, the rib segmentation model (YOLOv8m-seg) enhanced anatomical segmentation by preserving critical regions such as the heart and mediastinum and performed exceptionally well across various metrics, providing a solid basis for the image processing tasks in our classification pipeline.

The high performance of the screening models underscores their effectiveness in distinguishing healthy individuals from those with silicosis. This screening task benefited from multi-channel image inputs and showed high discriminatory power in all the convolutional neural network architectures tested. DenseNet121, InceptionV3, MobileNetV2, and ResNet152V2 each achieved near-perfect accuracy, indicating their ability to effectively differentiate between silicosis and healthy cases. This outperforms the typical accuracy reported in previous studies (80–98%) as shown in Table 9.

However, the challenges observed in PMF classification suggest that disease staging remains a complex task. In this staging task, performance varied between disease stages. Although all models exhibited high accuracy in identifying healthy cases, they showed greater difficulty in distinguishing between SS and PMF. DenseNet121 exhibited excellent precision and recall in identifying healthy cases. However, its performance dipped in more severe conditions, especially in PMF, with relatively lower scores, indicating challenges in accurately identifying more advanced stages of the disease. InceptionV3 showed a strong

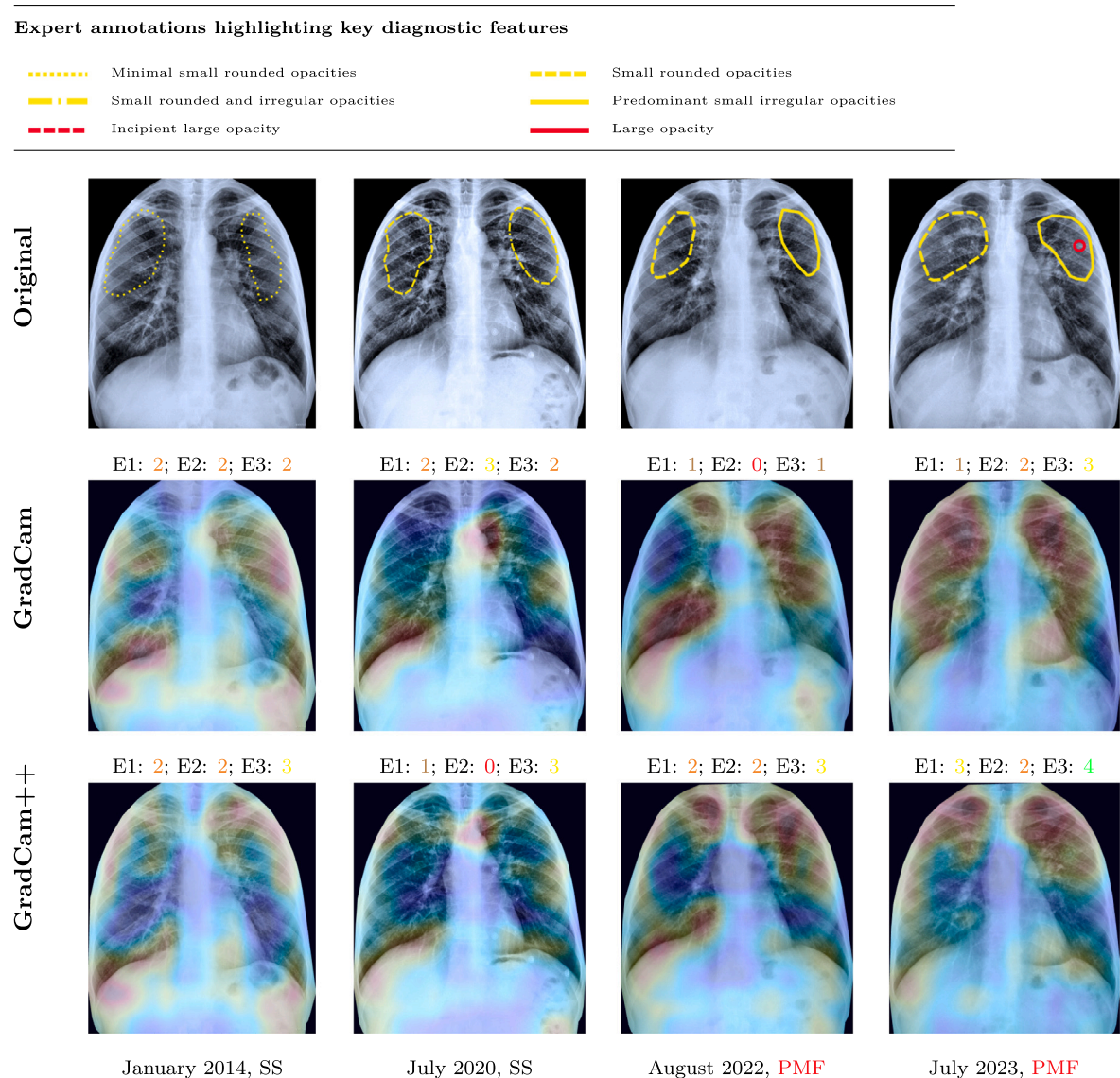


Fig. 7. X-ray images illustrating the progression of a patient from simple silicosis (SS) to progressive massive fibrosis (PMF), along with heatmaps generated by the GradCAM and GradCAM++ models, using the trained MobileNetV2 classification model for image diagnosis. Two out of the three images in the temporal sequence were misclassified by MobileNetV2 (those from August 2022 and July 2023). Above each heatmap, the score assigned by each evaluator is displayed, on a scale from 0 to 4, where E1, E2, and E3 denote the ratings given by each evaluator, respectively.

recall for the healthy cases, but struggled with PMF. This indicates a sensitivity to the Healthy class but a potential overfitting or misclassification issue when dealing with PMF cases. MobileNetV2 demonstrated a balanced performance with high metrics in the healthy and SS categories and moderate success in PMF detection. The model maintained a consistent F1-score across SS and PMF, showing its utility in more evenly handling both initial and severe stages of silicosis. Finally, ResNet152V2 also showed strong performance in the healthy cases but faced challenges with PMF, yet slightly better than other models, suggesting a modest advantage in dealing with complex cases. These results highlight the challenge of accurately staging silicosis, which is consistent with findings previous research reported, such as [12,16].

To assess the generalization capacity of our model, we applied state-of-the-art segmentation and classification methods [14,16,17] to our dataset, allowing a direct performance comparison (see Table 10). The results indicate that our proposed pipeline, which includes ribcage segmentation and classification with MobileNetV2, consistently outperforms models that rely on lung segmentation alone.

Error analysis revealed that most misclassifications occurred during the transition from SS to PMF, reflecting the inherent subjectivity in

disease progression assessment. This issue is consistent with clinical observations, where distinguishing between SS and early-stage PMF remains a diagnostic challenge.

The current understanding of deep learning interpretability is limited and determining reliable interpretability methods for case-specific applications is currently a challenging task. In this study, the application of explainable AI (XAI) validated by domain experts provided a level of interpretability and transparency absent in most studies, where heatmaps are often displayed without clinical validation (Table 9). Expert validation of GradCAM and GradCAM++ heatmaps confirmed that the highlighted regions were clinically relevant, providing valuable insights into its classification decisions, but also exposed variability among evaluators. This variability suggests the need for a multidisciplinary approach to further refine model interpretability, ultimately increasing their utility and reliability for a wider range of healthcare professionals.

A significant challenge in this study was the collection and diversity of the dataset. There are few public repositories with representative silicosis images, which shows an existing barrier to accessing curated

Table 9

Summary of studies on pneumoconiosis classification methods, including Region-of-Interest extraction approaches, classification models, best performance metrics, validation methods, and use of Explainable Artificial Intelligence. Performance metrics are reported for screening and staging tasks where available.

Ref.	Region-of-Interest (ROI) Extraction	Classification model	Performance (Best results)	Validation	XAI
[14]	No ROI; lung mask (using U-Net); and lung bounding box.	InceptionV3	Scr-AUC: 0.878	Hold-out (~5% test)	–
[12]	Lungs are masked (using U-Net) and divided into 3 subregions.	Each subregion is classified (ResNet) into categories 0, 1, 2, or 3 based on opacity level. Each radiograph is classified as 0, I, II, or III based on regional classifications.	Scr-Acc: 0.97 Scr-Sens: 0.98 Scr-Spec: 0.97 Stg-Acc: 0.92	Hold-out (~34% test) Class proportions differ between the training and test sets.	–
[22]	Lungs are masked (using U-Net) and divided into 3 subregions.	Multi-scale feature mapping model for opacity extraction (model trained using 1686 images labeled with fibrosis category). This trained model is then tested on their dataset. Output is used to classify opacities by size and location per ILO guidelines.	Stg-AUC: 0.84	All pneumo. images are used for testing.	–
[13]	–	CheXNet followed by CNN-based feature aggregation methods or SVM-based classification method.	Scr-Acc: 0.92 Scr-Sens: 0.96 Scr-Spec: 0.87	3-fold CV	–
[15]	–	Several models (best DenseNet53).	Scr-Acc: 0.88 Scr-Prec: 0.83 Scr-Rec: 0.53	Hold-out (~20% test)	–
[16]	Lung mask segmentation (using U-Net).	ResNet34.	Scr-Acc: 0.92 Scr-AUC: 0.8 Stg-Acc: 0.70	10-fold CV	–
[17]	Bounding boxes of lung masks (segmented using ED-Net) are concatenated.	Several models (best ResNet18) using Log-Normal Label Distribution Learning and Focal Staging Loss.	Stg-Acc: 0.90 Stg-Prec: 0.84 Stg-Sens: 0.78 Stg-Spec: 0.95 Stg-F1: 0.80 Stg-AUC: 0.96	10-fold CV	–
[18]	–	Ensemble learning using CheXNet, DenseNet, InceptionV3, Xception, and ResNet50, with methods including average probability, weighted averaging, and majority voting.	Scr-Acc: 0.91 Scr-Sens: 0.90 Scr-Spec: 0.93 Scr-Prec: 0.91 Scr-F1: 0.91	LOOCV RCV (~27% test)	–
[19]	–	Several models (best ResNet101).	Scr-AUC: 0.85 Scr-Acc: 0.79 Scr-Sens: 0.35 Scr-Spec: 0.90 Scr-Prec: 0.77 Scr-F1: 0.78	Hold-out (20% test)	GradCAM. Maps are displayed without analysis.
[20]	–	Several models (best Resnet101).	Scr-Acc: 0.98 Stg- μ -avg-AUC: 0.93 Stg-M-avg-AUC: 0.94 Stg-Acc: 0.72 Stg-M-F1: 0.78	Hold-out (20% test)	ResNet101 heat maps shown without analysis.
[21]	Lungs are masked (using U-Net) and divided into 3 subregions.	MultiStage join approach based on EfficientNetB0V1.	Scr-Acc: 0.97 Scr-Sens: 0.98 Scr-Spec: 0.97 Stg-Acc: 0.88 Stg-M-F1: 0.78	Hold-out (20% test)	–
Current study	Rib cage segmentation based on instance segmentation of ribs (using YOLOv8).	Several models (best MobileNetV2).	Scr-Acc: 0.99 Scr-AUC: 1.00 Scr-PR-AUC: 1.00 Stg-Acc: 0.84 Stg-AUC: 0.94 Stg-PR-AUC: 0.88	5-fold CV	GradCAM and GradCAM++ analyzed and evaluated by specialized clinics.

Abbreviations: ROI: Region of Interest; Scr: Screening (metrics with the prefix “Scr-” refer to the screening task, i.e., distinguishing between diseased and non-diseased cases); Stg: Staging (metrics with the prefix “Stg-” refer to the staging task, i.e., distinguishing between different stages of silicosis); AUC: Area Under the Curve (for the Receiver Operating Characteristic); Acc: Accuracy; Sens: Sensitivity; Spec: Specificity; Prec: Precision; F1: F1-Score; CV: Cross Validation; LOOCV: Leave-One-Out Cross Validation; RCV: Randomized Cross Validation; μ -avg: Micro Average; M-avg: Macro Average; PR-AUC: Precision-Recall Area Under the Curve; XAI: Explainable Artificial Intelligence.

datasets. Additionally, engineering-stone silicosis tends to be concentrated in localized areas, particularly around factories where artificial stone is processed. This geographical and occupational concentration further restricted the diversity of available data. The nature of silicosis

as an occupational disease also introduces a significant gender bias, with the majority of affected individuals being male due to the type of work associated with exposure to artificial stone dust. This bias is evident in our dataset, where all the patients are men. While this

Table 10

Results of state-of-the-art methods, evaluated using our own dataset.

Model	Staging overall performance metrics			
	Accuracy	ROC AUC	PR AUC	Log-loss
Ribcage seg. (YOLO) + MobileNetV2 (ours)	0.82	0.93	0.87	0.38
Lung seg. (PSPNet) + InceptionV3 [14]	0.78	0.92	0.86	0.45
Lung seg. (PSPNet) + ResNet34 [17]	0.61	0.81	0.65	0.70
Lung seg. (PSPNet) + ResNet18 (FSL) [16]	0.50	0.66	0.52	0.96
Lung seg. (PSPNet) + ResNet18 [16]	0.49	0.82	0.69	0.93

reflects the real-world distribution of the disease, it also highlights the importance of considering such biases in the development and evaluation of diagnostic models.

Our longitudinal methodological design provided a unique advantage over previous studies that rely on cross-sectional data. This design enabled further study of the model robustness at different stages of the disease, offering information on how models perform as the disease advances.

Addressing these challenges requires ongoing efforts to improve data collection and model validation. Collaborations with multiple institutions could help enhance the diversity and representativeness of the dataset. Additionally, incorporating techniques to mitigate bias and improve generalization will be crucial for developing robust models that perform well across different populations and settings.

5. Conclusions

This study demonstrates the feasibility and effectiveness of deep learning models for the automatic screening and staging of silicosis. The strong performance of the YOLOv8m-seg model in rib segmentation and the high accuracy of the classification models underscore the potential of these techniques in supporting the diagnosis and management of engineering-stone silicosis.

However, the challenges in accurately distinguishing between SS and PMF suggest that further research is needed. The results indicate that AI-based approaches can support clinical decision-making in silicosis diagnosis, but improvements are needed to enhance robustness in disease staging. The integration of additional clinical data, such as blood biomarkers [41,42] could refine predictive capabilities.

Moreover, the interpretability analysis highlighted the importance of expert validation in ensuring model reliability. The observed variability in specialist assessments suggests that further efforts should focus on improving explainability techniques to foster clinical adoption.

Despite its limitations, this study contributes to the application of deep learning for occupational lung diseases. Future work should explore larger, more diverse datasets and multidisciplinary collaborations to enhance model generalizability and clinical impact.

CRedit authorship contribution statement

Blanca Priego-Torres: Writing – original draft, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Daniel Sanchez-Morillo:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization. **Ebrahim Khalili:** Writing – review & editing, Validation, Software, Methodology, Investigation, Formal analysis. **Miguel Ángel Conde-Sánchez:** Writing – review & editing, Validation, Data curation. **Andrés García-Gómez:** Writing – review & editing, Validation, Data curation. **Antonio León-Jiménez:** Writing – review & editing, Validation, Data curation.

Ethical statements

The research followed the guidelines of the Declaration of Helsinki and received approval from the Institutional Research Ethics Committee of the province of Cadiz, Spain (register number 151.22).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was funded in the “Convocatoria 2021 de Ayudas a Proyectos de Excelencia, en régimen de concurrencia competitiva, destinadas a entidades calificadas como Agentes del Sistema Andaluz del Conocimiento, en el ámbito del Plan Andaluz de Investigación, Desarrollo e Innovación (PAIDI 2020). Consejería de Universidad, Investigación e Innovación de la Junta de Andalucía”, under grant number ProjExcel_00942.

Data availability

The data are not publicly available due to privacy or ethical restrictions. The data that support the findings of this study are available on request from the corresponding author for researchers who meet the criteria for confidential data access as stipulated by participant informed consent and the Institutional Research Ethics Committee of the province of Cadiz, Spain.

References

- [1] P. Cullinan, X. Muñoz, H. Suojalehto, R. Agius, S. Jindal, T. Sigsgaard, A. Blomberg, D. Charpin, I. Annesi-Maesano, M. Gulati, et al., Occupational lung diseases: from old and novel exposures to effective preventive strategies, *Lancet Respir. Med.* 5 (5) (2017) 445–455.
- [2] R.F. Hoy, M.F. Jeebhay, C. Cavalin, W. Chen, R.A. Cohen, E. Fireman, L.H. Go, A. León-Jiménez, A. Menéndez-Navarro, M. Ribeiro, et al., Current global perspectives on silicosis—Convergence of old and newly emergent hazards, *Respirology* 27 (6) (2022) 387–398.
- [3] A. León-Jiménez, J.M. Manuel, M. García-Rojo, M.G. Pintado-Herrera, J.A. López-López, A. Hidalgo-Molina, R. García, P. Muriel-Cueto, N. Maira-González, D. Del Castillo-Otero, et al., Compositional and structural analysis of engineered stones and inorganic particles in silicotic nodules of exposed workers, Part. Fibre Toxicol. 18 (2021) 1–16.
- [4] K. Newbiggin, R. Parsons, D. Deller, R. Edwards, R. McBean, Stonemasons with silicosis: Preliminary findings and a warning message from Australia, *Respirology* 24 (12) (2019).
- [5] A. León-Jiménez, A. Hidalgo-Molina, M.Á. Conde-Sánchez, A. Pérez-Alonso, J.M. Morales-Morales, E.M. García-Gómez, J.A. Córdoba-Doña, Artificial stone silicosis: rapid progression following exposure cessation, *Chest* 158 (3) (2020) 1060–1068.
- [6] J. Yi, K. Tapia, J.W. Robinson, Z. Gandomkar, E.S. Mo'ayyad, P.C. Brennan, N. Sommerfeld, S.T. Taba, Radiologists' performance in diagnosing silicosis on high-resolution computed tomography (HRCT) scans: an online platform, in: *Medical Imaging 2024: Image Perception, Observer Performance, and Technology Assessment*, vol. 12929, SPIE, 2024, pp. 32–38.
- [7] V. Gowda, G. Cheng, K. Saito, The b reader program, silicosis, and physician workload management: a niche for AI technologies, *J. Occup. Environ. Med.* 63 (7) (2021) e471–e473.
- [8] I. Castiglioni, L. Rundo, M. Codari, G. Di Leo, C. Salvatore, M. Interlenghi, F. Gallivanone, A. Cozzi, N.C. D'Amico, F. Sardanelli, AI applications to medical images: From machine learning to deep learning, *Phys. Medica* 83 (2021) 9–24.
- [9] A. Davri, E. Birbas, T. Kanavos, G. Ntritos, N. Giannakeas, A.T. Tzallas, A. Batistatou, Deep learning for lung cancer diagnosis, prognosis and prediction using histological and cytological images: a systematic review, *Cancers* 15 (15) (2023) 3981.

- [10] M.A. Balci, L.M. Batrancea, Ö. Akgüller, A. Nichita, A series-based deep learning approach to lung nodule image classification, *Cancers* 15 (3) (2023) 843.
- [11] Y.H. Bhosale, K.S. Patnaik, Application of deep learning techniques in diagnosis of covid-19 (coronavirus): a systematic review, *Neural Process. Lett.* 55 (3) (2023) 3551–3603.
- [12] L. Zhang, R. Rong, Q. Li, D.M. Yang, B. Yao, D. Luo, X. Zhang, X. Zhu, J. Luo, Y. Liu, et al., A deep learning-based model for screening and staging pneumoconiosis, *Sci. Rep.* 11 (1) (2021) 2201.
- [13] L. Devnath, S. Luo, P. Summons, D. Wang, Automated detection of pneumoconiosis with multilevel deep features learned from chest X-Ray radiographs, *Comput. Biol. Med.* 129 (2021) 104125.
- [14] X. Wang, J. Yu, Q. Zhu, S. Li, Z. Zhao, B. Yang, J. Pu, Potential of deep learning in assessing pneumoconiosis depicted on digital chest radiography, *Occup. Environ. Med.* 77 (9) (2020) 597–602.
- [15] C. Hao, N. Jin, C. Qiu, K. Ba, X. Wang, H. Zhang, Q. Zhao, B. Huang, Balanced convolutional neural networks for pneumoconiosis detection, *Int. J. Environ. Res. Public Heal.* 18 (17) (2021) 9091.
- [16] F. Yang, Z.-R. Tang, J. Chen, M. Tang, S. Wang, W. Qi, C. Yao, Y. Yu, Y. Guo, Z. Yu, Pneumoconiosis computer aided diagnosis system based on X-rays and deep learning, *BMC Med. Imaging* 21 (2021) 1–7.
- [17] W. Sun, D. Wu, Y. Luo, L. Liu, H. Zhang, S. Wu, Y. Zhang, C. Wang, H. Zheng, J. Shen, et al., A fully deep learning paradigm for pneumoconiosis staging on chest radiographs, *IEEE J. Biomed. Heal. Informatics* 26 (10) (2022) 5154–5164.
- [18] L. Devnath, S. Luo, P. Summons, D. Wang, K. Shaukat, I.A. Hameed, F.S. Alrayes, Deep ensemble learning for the automatic detection of pneumoconiosis in coal worker's chest X-ray radiography, *J. Clin. Med.* 11 (18) (2022) 5342.
- [19] G.K. Sharma, P. Harjule, B. Agarwal, R. Kumar, Silicosis detection using extended transfer learning model, in: *Recent Trends in Image Processing and Pattern Recognition*, Springer Nature Switzerland, 2024, pp. 111–126.
- [20] Y. Zhang, B. Zheng, F. Zeng, X. Cheng, T. Wu, Y. Peng, Y. Zhang, Y. Xie, W. Yi, W. Chen, et al., Potential of digital chest radiography-based deep learning in screening and diagnosing pneumoconiosis: An observational study, *Medicine* 103 (25) (2024) e38478.
- [21] C. Liu, Y. Fang, Y. Xie, H. Zheng, X. Li, D. Wu, T. Zhang, Deep learning pneumoconiosis staging and diagnosis system based on multi-stage joint approach, *BMC Med. Imaging* 24 (1) (2024) 165.
- [22] Y. Zhang, Computer-aided diagnosis for pneumoconiosis staging based on multi-scale feature mapping, *Int. J. Comput. Intell. Syst.* 14 (1) (2021) 191.
- [23] R.F. Álvarez, C.M. González, A.Q. Martínez, J.J.B. Pérez, L.C. Fernández, A.P. Fernández, Guidelines for the diagnosis and monitoring of silicosis, *Arch. de Bronconeumol. (Engl. Edition)* 51 (2) (2015) 86–93.
- [24] N. Suganuma, Y. Kusaka, K.G. Hering, T. Vehmas, T. Kraus, H. Arakawa, J.E. Parker, L. Kivisaari, M. Letourneux, P.A. Gevenois, et al., Reliability of the proposed international classification of high-resolution computed tomography for occupational and environmental respiratory diseases, *J. Occup. Health* 51 (3) (2009) 210–222.
- [25] H.Q. Nguyen, K. Lam, L.T. Le, H.H. Pham, D.Q. Tran, D.B. Nguyen, D.D. Le, C.M. Pham, H.T. Tong, D.H. Dinh, et al., VinDr-CXR: An open dataset of chest X-rays with radiologist's annotations, *Sci. Data* 9 (1) (2022) 429.
- [26] H.C. Nguyen, T.T. Le, H.H. Pham, H.Q. Nguyen, VinDr-RibCXR: A benchmark dataset for automatic segmentation and labeling of individual ribs on chest X-rays, 2021, arXiv preprint [arXiv:2107.01327](https://arxiv.org/abs/2107.01327).
- [27] G. Jocher, A. Chaurasia, J. Qiu, YOLO by ultralytics, 2023, URL <https://github.com/ultralytics/ultralytics>.
- [28] S. Hao, X. Li, W. Peng, Z. Fan, Z. Ji, I. Ganchev, YOLO-CXR: A novel detection network for locating multiple small lesions in chest X-Ray images, *IEEE Access* 12 (2024) 156003–156019, <http://dx.doi.org/10.1109/ACCESS.2024.3482102>.
- [29] Z. Mustafa, H. Nsour, Using computer vision techniques to automatically detect abnormalities in chest X-rays, *Diagnostics* 13 (18) (2023) 2979.
- [30] K. Zuiderveld, Contrast limited adaptive histogram equalization, in: P.S. Heckbert (Ed.), *Graphics Gems IV*, Academic Press Professional, Inc., 1994, pp. 474–485.
- [31] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, *Proc. the IEEE Conf. Comput. Vis. Pattern Recognit.* (2017) 4700–4708.
- [32] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2818–2826.
- [33] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.-C. Chen, MobileNetV2: Inverted residuals and linear bottlenecks, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4510–4520.
- [34] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.
- [35] M. Everingham, L. Van Gool, C.K. Williams, J. Winn, A. Zisserman, The pascal visual object classes (voc) challenge, *Int. J. Comput. Vis.* 88 (2) (2010) 303–338.
- [36] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft coco: Common objects in context, in: *European Conference on Computer Vision*, Springer, 2014, pp. 740–755.
- [37] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 618–626.
- [38] A. Chattopadhyay, A. Sarkar, P. Howlader, V.N. Balasubramanian, Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks, in: *2018 IEEE Winter Conference on Applications of Computer Vision, WACV, IEEE*, 2018, pp. 839–847.
- [39] J.P. Cohen, J.D. Viviano, P. Bertin, P. Morrison, P. Torabian, M. Guarrera, M.P. Lungren, A. Chaudhari, R. Brooks, M. Hashir, H. Bertrand, TorchXRyVision: A library of chest X-ray datasets and models, in: *Medical Imaging with Deep Learning*, 2022, URL <https://github.com/mlmed/torchxrayvision>.
- [40] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, 2014, arXiv preprint [arXiv:1412.6980](https://arxiv.org/abs/1412.6980).
- [41] D. Sanchez-Morillo, A. León-Jiménez, M. Guerrero-Chanivet, G. Jiménez-Gómez, A. Hidalgo-Molina, A. Campos-Caro, Integrating routine blood biomarkers and artificial intelligence for supporting diagnosis of silicosis in engineered stone workers, *Bioeng. Transl. Med.* e10694 (2024).
- [42] A. García-Núñez, G. Jiménez-Gómez, A. Hidalgo-Molina, J.A. Córdoba-Doña, A. León-Jiménez, A. Campos-Caro, Inflammatory indices obtained from routine blood tests show an inflammatory state associated with disease progression in engineered stone silicosis patients, *Sci. Rep.* 12 (1) (2022) 8211.