Explainable Deep Learning for Thoracic Radiographic Diagnosis: A COVID-19 Case Study Toward Clinically Meaningful Evaluation

Divine Nicholas-Omoregbe , Olamilekan Shobayo * , Obinna Okoyeigbo , Mansi Khurana , Reza Saatchi

Posted Date: 13 February 2026

doi: 10.20944/preprints202602.1063.v1

Keywords: COVID-19; chest X-ray; explainable artificial intelligence; deep learning; Grad-CAM; medical image analysis

 doi:10.20944/preprints202602.1063.v1

Article

Explainable Deep Learning for Thoracic Radiographic Diagnosis: A COVID-19 Case Study Toward Clinically Meaningful Evaluation

Divine Nicholas-Omoregbe [1], Olamilekan Shobayo [1,*], Obinna Okoyeigbo [2], Mansi Khurana [1] and Reza Saatchi [3]

[1] School of Computing and Digital Technologies, Sheffield Hallam University, 151 Arundel Street, Sheffield S1 2NU, UK

[2] Department of Engineering, Edge Hill University, Ormskirk L39 4QP, UK

[3] School of Engineering and Built Environment, Sheffield Hallam University, Pond Street, Sheffield S1 1WB, UK

* Correspondence: o.shobayo@shu.ac.uk

Abstract

COVID-19 still poses a global public health challenge, exerting pressure on radiology services. Chest X-ray (CXR) imaging is widely used for respiratory assessment due to its accessibility and cost effectiveness. However its interpretation is often challenging because of subtle radiographic features and inter-observer variability. Although recent deep learning (DL) approaches have shown strong performance in automated CXR classification, their black-box nature limits interpretability. This study proposes an explainable deep learning framework for COVID-19 detection from chest X-ray images. The framework incorporates anatomically guided preprocessing, including lung-region isolation, contrast-limited adaptive histogram equalization (CLAHE), bone suppression, and feature enhancement. A novel four-channel input representation was constructed by combining lung-isolated soft-tissue images with frequency-domain opacity maps, vessel enhancement maps, and texture-based features. Classification was performed using a modified Xception-based convolutional neural network, while Gradient-weighted Class Activation Mapping (Grad-CAM) was employed to provide visual explanations and enhance interpretability. The framework was evaluated on the publicly available COVID-19 Radiography Database, achieving an accuracy of 95.3%, an AUC of 0.983, and a Matthews Correlation Coefficient of approximately 0.83. Threshold optimisation improved sensitivity, reducing missed COVID-19 cases while maintaining high overall performance. Explainability analysis showed that model attention was primarily focused on clinically relevant lung regions.

Keywords: COVID-19; chest X-ray; explainable artificial intelligence; deep learning; Grad-CAM; medical image analysis

1. Introduction

Modern medical imaging has become a crucial component of modern medicine for diagnosing, treating, and detecting various diseases. Medical imaging technologies continue to evolve with advances in artificial intelligence (AI) and deep learning technologies, which have changed dramatically the way we create and produce medical imaging technologies. The World Health Organisation reported that over 3.6 billion diagnostic imaging procedures were conducted annually in global healthcare facilities and highlighted the importance diagnostic imaging on modern healthcare services and the importance it plays in assisting with the detection and management of conditions like cancer, cardiovascular disease, and neurological disorders. X-ray, MRI, CT and ultrasound are some of the technologies being used for this purpose. The high demand for diagnostic

 doi:10.20944/preprints202602.1063.v1

imaging coupled with the shortage of consultant radiologists is placing substantial strain on health care services around the world to meet the growing need for imaging [1–3].

With deep learning tools and systems currently being used to assist with this need, Deep Learning technologies are becoming a significant asset for the healthcare industry. Convolutional Neural Networks (CNNs) are proven to classify and segment images with exceptional accuracy [4–6]. These models can identify subtle patterns in images that would otherwise be undetectable by humans, providing clinicians with vast increases in their ability to make better treatment decisions. For example, dermatology is finding great promise in utilising Convolutional Neural Networks to aid the classification of skin lesions with an equivalent level of accuracy to Board Certified Dermatologists [7]. AI systems can match or exceed the performance of human radiologists in detecting pneumonia in patients based on chest X-rays [8].

Despite the progress, the clinical adoption of deep learning technologies is lagging demands, primarily due to the 'opaque' nature of the deep learning model(s). Due to the inability of healthcare professionals to easily understand the inner workings of these models, they are often referred to as 'Black Box' systems. However, their “black-box” nature is a barrier, since predictions are frequently not transparent or easily interpretable [9–11]. This lack of transparency undermines clinician confidence and gives rise to ethical and regulatory concerns. The inability to understand or explain the predictions observed with these systems created a distrust of Artificial Intelligence models by healthcare professionals that raises a host of ethical and legal questions as well as creating barriers to regulatory approval of these systems [12]. A recent survey of healthcare leaders reported that more than 60% of respondents indicated that the absence of 'explainability' is the primary barrier to the use of AI in hospitals [13].

To address these issues, Explainable Artificial Intelligence (XAI) has emerged, attempting to provide techniques such as saliency maps, feature detection, and concept-based models to produce interpretable outputs [14,15]. XAI aims to overcome the lack of interpretability by providing methods that make AI model predictions transparent, understandable, and clinically meaningful [16]. Techniques concluded SHAP (Shapley Additive Explanations) values, Concept Bottleneck Models, are attempting to make the connection between insight from an algorithm and the interpretation of how the AI concluded it's making, but many of these techniques are still untested and have not yet made it into the clinical workflow [17,18].

Many of these techniques remain to be evaluated in the clinical workflow, creating the need for the current study on explainable deep learning models applied to medical imaging. This study aims to develop and evaluate explainable deep learning models that enhance transparency, usability, and trust in Medical Imaging Analysis. This research aims to develop and evaluate an explainable deep learning framework for thoracic medical image analysis, integrating multi-channel feature preprocessing extraction to improve diagnostic accuracy and interpretability, guided by preprocessing and explainable artificial intelligence techniques.

Recent research has sought to include explainable AI into medical image analysis and chest radiograph interpretation. Multi-channel chest radiograph pipelines have been explored previously, for example, by constructing channels such as LBP, CLAHE, and contrast/edge-enhanced maps and learning from them using deep neural networks [19]. Reviews of deep learning for chest X-ray analysis highlighted the importance of preprocessing, enhancement, and segmentation or masking choices for model performance and reliability [20,21]. While related feature-fusion strategies exist, an identical combination of these anatomically and diagnostically motivated channels was not identified in the reviewed literature.

Beyond input representation, prior studies in explainable AI for medical imaging have highlighted that saliency maps alone are often insufficient for reliable interpretation, particularly for non-technical stakeholders [22,23]. Grad-CAM is an established explainability technique, however it is typically presented solely as a visual artefact, leaving interpretation to expert users [24,25]. While prior studies have explored either quantitative evaluation of saliency maps or textual explanation of model outputs, a structured rule-based translation of spatial attention metrics into clinician-oriented

 doi:10.20944/preprints202602.1063.v1

natural-language explanations has not been widely reported in the literature. The major contributions of this research are as follows:

- A novel four-channel input representation for thoracic chest X-ray analysis was proposed, integrating lung-region masking, frequency-domain enhancement, vesselness filtering, and texture-based features. While individual preprocessing and feature-enhancement techniques have been explored in prior chest radiograph studies, an identical combination of these anatomically and diagnostically motivated channels has not been identified in the reviewed literature. This novel four-channel combination provides complementary information beyond a single intensity channel, supporting improved sensitivity and anatomically aligned model attention when combined with explainable AI techniques.
- A structured, explainable deep learning framework was developed for thoracic medical image analysis, integrating the proposed multi-channel input representation with a modified deep convolutional neural network architecture. The framework was designed to balance diagnostic performance and interpretability, addressing the limitations of conventional single-channel pipelines and black-box deep learning models commonly reported in medical imaging literature.
- A novel approach to interpreting and communicating model attention was introduced by combining quantitative spatial attention analysis with rule-based natural-language explanations. Rather than presenting saliency maps solely as visual artefacts, this study quantified the distribution of model attention inside and outside lung regions and translated these measurements into concise, human-readable explanations. This structured explanation strategy improved the accessibility and interpretability of explainable AI outputs for non-technical users, including clinicians.
- The proposed framework enhanced clinical relevance and trustworthiness of AI-assisted diagnosis, by ensuring that model attention was anatomically meaningful and aligned with lung regions of interest. This design supported transparent decision-making and addressed key ethical, regulatory, and usability concerns associated with the deployment of deep learning models in real-world clinical settings.
- The study provided practical insights into the integration of explainable AI within medical imaging workflows, demonstrating how anatomically guided preprocessing, multi-channel learning, and explainability mechanisms can be combined into a cohesive and computationally feasible diagnostic system.

2. Literature Review

In recent years, deep learning has transformed the way medical images could be interpreted by allowing for an expert level of performance in the areas of segmentation and classification. Deep learning provides healthcare systems with the ability to automate and extract important diagnostic features from imaging datasets that improves both efficiency and quality of care [4,5]. Notably, deep learning has demonstrated diagnostic accuracy equivalent to that of human physicians working in radiology, ophthalmology, cardiology, and pathology [26,27].

2.1. Deep Learning for Chest X-Ray Classification

Classification of X-ray chest imaging is critical in diagnosing patients who may be experiencing respiratory failure at an early stage [20]. Several studies have found that convolutional neural networks (CNNs) have been particularly effective at analysing chest X-ray images because they can automatically learn feature representations directly from pixel data [4]. As the use of deep learning becomes increasingly accepted, knowledge and understanding of how deep learning can create results have become an issue [9,10,15].

In a study the researchers introduced CheXNet,a deep neural network with 121 layers, that was trained with the NIH ChestX-ray14 dataset to predict pneumonia [8]. However, CheXNet focused primarily on classification accuracy and did not provide mechanisms for uncertainty handling or localisation of pathological regions. Similarly, the CheXpert dataset and associated models

 doi:10.20944/preprints202602.1063.v1

introduced uncertainty-aware learning but did not explicitly address whether model attention aligned with clinically meaningful lung regions [28]. Other studies have explored alternative architecture and transfer learning strategies, demonstrating improved sensitivity and specificity in thoracic disease classification [29]. While these approaches improved performance, most did not assess whether increased accuracy translated into clinically interpretable model behaviour.

Research focused on COVID-19 diagnosis further demonstrated high reported accuracy using deep learning models such as COVID-Net and pre-trained CNNs, including ResNet50 and InceptionV3 [30,31]. However, concerns regarding dataset heterogeneity, overfitting, and shortcut learning limited confidence in clinical reliability [32,33]. Studies incorporating lung segmentation before classification demonstrated that constraining model input to anatomically meaningful regions improved diagnostic performance and reliability of visual explanations [34]. Conversely, Grad-CAM-based explanations often revealed attention outside pathological regions, highlighting the limitations of explainability when applied without appropriate preprocessing [14,22,23].

Overall, despite Deep learning models providing good diagnostic performance in chest X-ray classification, some challenges remain. Many of the published accuracy results for these deep learning models are from either small or unbalanced datasets, which present an increased likelihood of overfitting and poor clinical generalisability. Additionally, most deep learning models utilise unaltered, raw images for diagnosing patients without removing or addressing the influence of other anatomical structures in the same imaging area, which therefore limits the reliability of their predictions. Although explainability methods such as Grad-CAM are applied, highlighted regions frequently lack alignment with radiological findings. These limitations emphasise the need for research combining region-focused preprocessing, such as lung segmentation and bone suppression, with explainable deep learning methods to ensure predictions are accurate and clinically interpretable.

2.2. Explainable Artificial Intelligence in Medical Imaging

Despite strong diagnostic performance, the black-box nature of deep learning systems remains a major barrier to clinical adoption [9,11]. Explainable Artificial Intelligence (XAI) aims to address this limitation by offering explanations about how models generate predictions. It is therefore an essential element for both clinical decision support and regulatory approval for use within a healthcare environment [16]. Saliency-based methods such as Grad-CAM, Grad-CAM++, LIME, and SHAP are widely used to create visualisations of the areas of an image with the greatest impact on model prediction [14,35–37]. However, studies have revealed that areas that receive attention through these techniques do not consistently correspond to observable signs of disease on diagnostic imaging and may differ depending on the architecture and preprocessing strategy applied [22,23].

Concep and prototype-based, textual explanation approaches aim to link model predictions to clinical reasoning by matching the output of the model to rational explanations for decisions made [38–40]. These methods often require extensive human annotation, large multimodal datasets or creation of explanations that may not always be clinically valid [41]. Hybrid approaches that integrate several different methods of explainability have been presented to enhance robustness and enhance clinician trust. However, many of the same challenges associated with evaluating the consistency of the multiple methods, integrating them into the workflow, and establishing clinician trust persist [25,41].

The literature demonstrates progress in interpretability methods but highlights weaknesses limiting clinical applicability. Saliency-based methods may produce unstable heatmaps that do not reflect diagnostic reasoning [23]. Concept- and prototype-based methods require large, annotated datasets or fail to generalise. Textual explanation systems align with clinical language but are constrained by limited multimodal datasets and risk producing clinically irrelevant text [41].

Most XAI research remains retrospective, with limited validation in real clinical workflows. As noted in prior work, interpretability must be evaluated in collaboration with human experts to ensure that explanations are clinically meaningful and trustworthy [21]. The lack of standardised evaluation

 doi:10.20944/preprints202602.1063.v1

frameworks limits comparison and adoption. Overall, despite progress, existing approaches remain fragmented and inconsistently evaluated. These gaps motivate research that develops and evaluates explainability methods in terms of usability, reliability, and clinician trust. This article therefore positions explainable deep learning as a roadmap for transparent AI systems suitable for integration into health care practice.

2.3. Addressing Gaps and Advancing Knowledge

This research project addresses gaps in the current literature on deep learning applied to medical imaging. While convolutional neural networks (CNNs) have demonstrated strong performance across diagnostic problems, studies often note difficulties arising from limited dataset diversity, class imbalance, and lack of interpretability for clinical decision-making [34,42]. These problems become evident particularly in the classification of chest X-ray, where overlapping anatomical structures mask disease-relevant features, leading to a lack of visual clarity. This has had an impact on performance on thoracic diagnoses, leading to a lack of confidence in deployment in real-world healthcare situations [43].

The present research aims to address these shortcomings by incorporating a feature-focused preprocessing pipeline, including lung segmentation, bone suppression, and contrast enhancement, to improve clarity of diagnostically relevant regions. This is in line with recent work, which emphasises the importance that CNN attention is steered towards meaningful anatomical features instead of background patterns [34,44]. In addition, the project utilises explainable AI methods such as Grad-CAM, which provide insight regarding areas that contribute to classification outcomes, thus improving transparency and clinical trust [14,45].

3. Proposed Methodology

This study introduces an explainable deep learning framework that integrates lung-focused preprocessing, multi-channel feature construction, and visual explanation techniques to address interpretability and reliability challenges in automated chest X-ray classification. The proposed method comprises systematic pulmonary region isolation, feature-specific image enhancement, and robust feature extraction using a modified Xception-based convolutional neural network, followed by interpretable predictions generated with Grad-CAM. Each component is designed to improve diagnostic performance while simultaneously ensuring anatomical relevance, clinical interpretability, and transparency. The proposed methodology is illustrated in Figure 1, with each phase of the solution formulation described in the following sections.

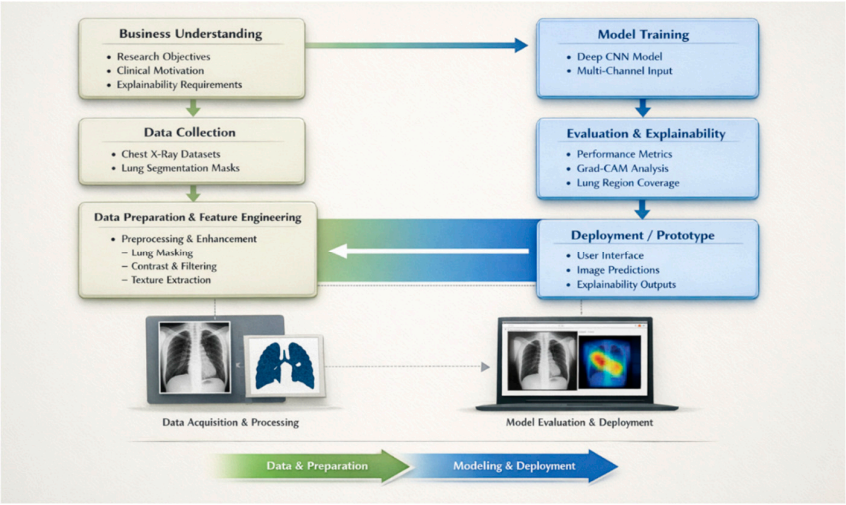


Figure 1. Workflow Diagram of the Proposed Explainable DL Framework.

 doi:10.20944/preprints202602.1063.v1

This research is based on secondary data and utilises publicly available radiographic datasets. The data set provided annotated images suitable for this study [4]. The primary dataset selected for this study is the COVID-19 Radiography Database, a publicly available benchmark dataset compiled from publicly available clinical repositories [46]. The dataset contains 21,165 chest X-ray images taken from a posterior–anterior (PA) projection classified into four diagnostic categories: COVID-19, Normal, Lung Opacity, and Viral Pneumonia [46]. An important element in determining dataset appropriateness is the presence of segmented lung masks, made for each image, so that lung areas can be isolated and samples pre-processed consistently. This aspect of lung segmentation masks is crucial to explainable deep learning, whereby a model is limited to areas of anatomic significance that provide clinical insights. The details of these operations are included in the methodology section. A summary of COVID-19 radiography database dataset composition used in this study is provided in Table 1 and illustrated in Figure 2.

Table 1. Summary of COVID-19 radiography database dataset composition used in this study.

Diagnostic Class	Number of Images	Description
COVID-19	3,616	Confirmed COVID-19 radiographs sourced from curated public repositories.
Normal	10,192	Chest radiographs with clear lung fields and no radiographic abnormalities.
Lung Opacity	6,012	Images showing non-COVID pulmonary opacities caused by various conditions.
Viral Pneumonia	1,345	Radiographs depicting viral pneumonia distinct from COVID-19.

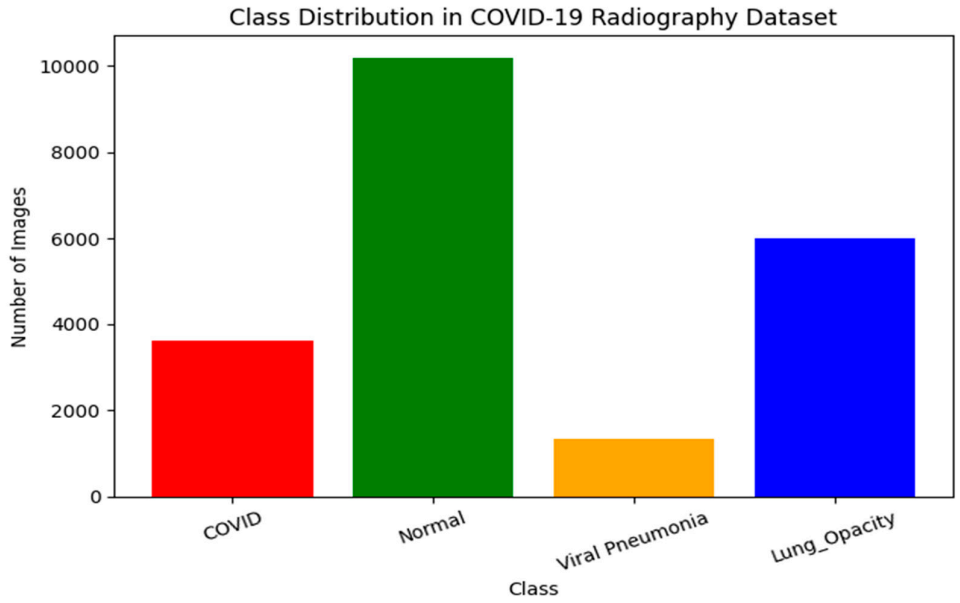


Figure 2. Distribution of the COVID-19 Radiography Database.

Rahman et al. established the COVID-19 Radiography Database to create an accessible database of COVID-related images for classification research [45]. Initial versions combined images from open-access datasets to reduce background noise, standardise metadata, and simplify analysis. A distinguishing feature is the inclusion of lung masks for every image, which enables the isolation of the lung region during preprocessing. This capability improves the reliability of preprocessing steps such as lung cropping, contrast enhancement, and XAI-based localisation. Studies suggest that a more tightly curated dataset with standardised PA-view images and corresponding lung masks supports more reliable preprocessing and region-of-interest extraction [45]. The dataset does not

 doi:10.20944/preprints202602.1063.v1

contain patient metadata, such as age, sex, or clinical history, but it includes consistent diagnostic labels and image quality, which supports reproducibility for deep learning research. Examples of chest x-ray class with corresponding lung mask in the dataset and their corresponding masks are shown in Figure 3.

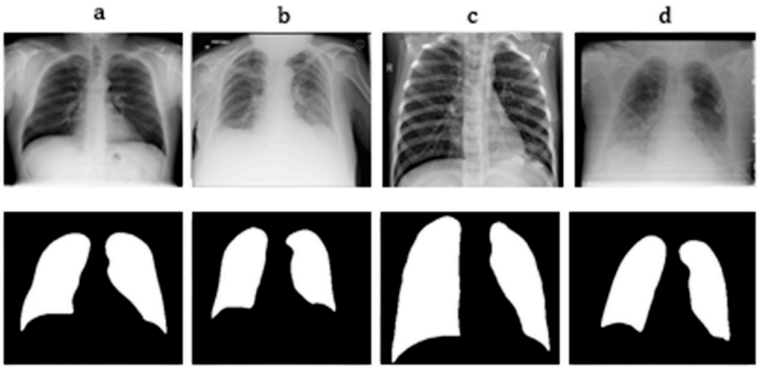


Figure 3. *Examples of* Chest X-ray images (top row) and their corresponding masked lung shapes (bottom images) for different clinical categories: (a) Normal, (b) Lung Opacity, (c) Pneumonia, and (d) COVID-19.

Normal chest radiographs show clear lung fields with normal vascular markings. Lung Opacity and Viral Pneumonia cases display abnormal density patterns, including localised or diffuse opacification. Bilateral ground-glass opacities and haziness are often present in most cases of COVID-19; however, early-stage infection may not present with clearly visible radiographic abnormalities [47]. Therefore, the preprocessing of images and the use of XAI techniques will be necessary for the developed methods to yield clinically useful results. lung-region masks obtained directly from the COVID-19 Radiography Database were used to isolate pulmonary regions during preprocessing. These masks were generated by the dataset authors using deep learning–based lung segmentation models trained on curated chest X-ray data and are provided as paired annotations for each image [45]. Although not manually delineated by medical practitioners, the masks exhibit consistent and anatomically plausible lung localisation and have been widely used in prior chest X-ray analysis studies [34,44].

3.1. Image Preprocessing and Normalisation

Image processing is a crucial stage in medical imaging analysis, where data quality significantly impacts the performance and reliability of deep learning models. The preprocessing pipeline in this study was directed towards enhancing the visibility of clinically useful lung structures by reducing noise and undesirable features within the image. This process consisted of lung segmentation and suppression of surrounding bony structures, enhancement of contrast, normalisation of the data, and data augmentation. Figure 4 shows a visual representation of the preprocessing workflow implemented in this research.

 doi:10.20944/preprints202602.1063.v1

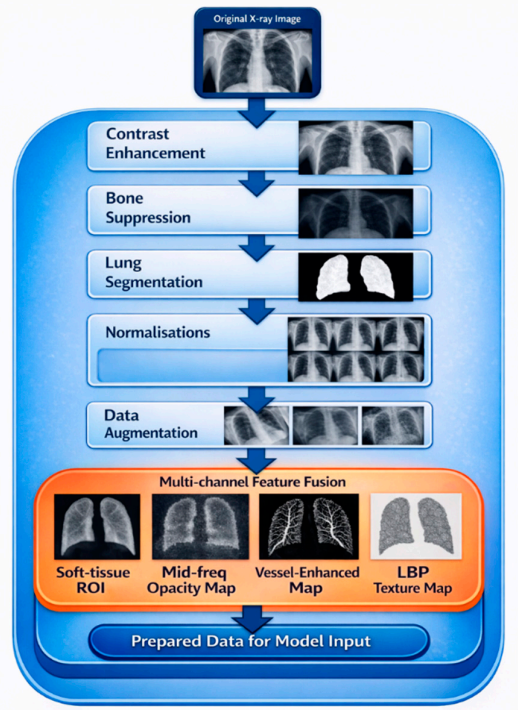


Figure 4. A visual representation of the preprocessing workflow implemented in this re-search.

3.1.1. Pulmonary Region of Interest (ROI) Extraction

Pulmonary region of interest (ROI) extraction was applied to isolate lung fields and enhance clinically relevant soft-tissue information while suppressing non-diagnostic structures. Lung segmentation was first used to restrict analysis to pulmonary regions, reducing the influence of surrounding anatomy and imaging artefacts [34,44]. Bone suppression was then applied to minimise the visual dominance of ribs and the spine, improving visibility of underlying lung pathology [43]. Finally, Contrast Limited Adaptive Histogram Equalisation (CLAHE) was used to increase local contrast, highlighting small patterns that can be seen in chest radiographs, such as ground-glass opacities or Consolidation [45,46].

I. Lung Region Isolation

A key limitation observed in previous deep-learning studies for chest X-ray classification is the model's over-reliance on non-lung artefacts. Several studies have shown that classifiers used irrelevant cues like laterality markers, collars, and scanner-specific background noise, leading to inflated accuracy and limited clinical value [32,33]. To mitigate this issue, some researchers have employed automatic lung segmentation networks to constrain model attention to pulmonary regions [34,44]. However, such approaches may introduce additional uncertainty and segmentation-induced errors. In contrast, the availability of ground-truth lung masks within the COVID-19 Radiography Database enables precise pulmonary isolation without reliance on automated segmentation methods [43]. Consequently, this study adopted anatomically guided masking to explicitly remove non-pulmonary structures, following prior findings that lung-field extraction improves robustness and supports the reliability of explainable techniques such as Grad-CAM [14,30,34]. By reducing shortcut learning and constraining model attention to clinically meaningful regions, this approach enhanced interpretability and strengthens the validity of explanation-based performance assessment. A comparison between dataset-provided lung segmentation masks and algorithmically generated lung segmentation, visualised as overlays on the original chest X-ray image is provided in Figure 5.

 doi:10.20944/preprints202602.1063.v1

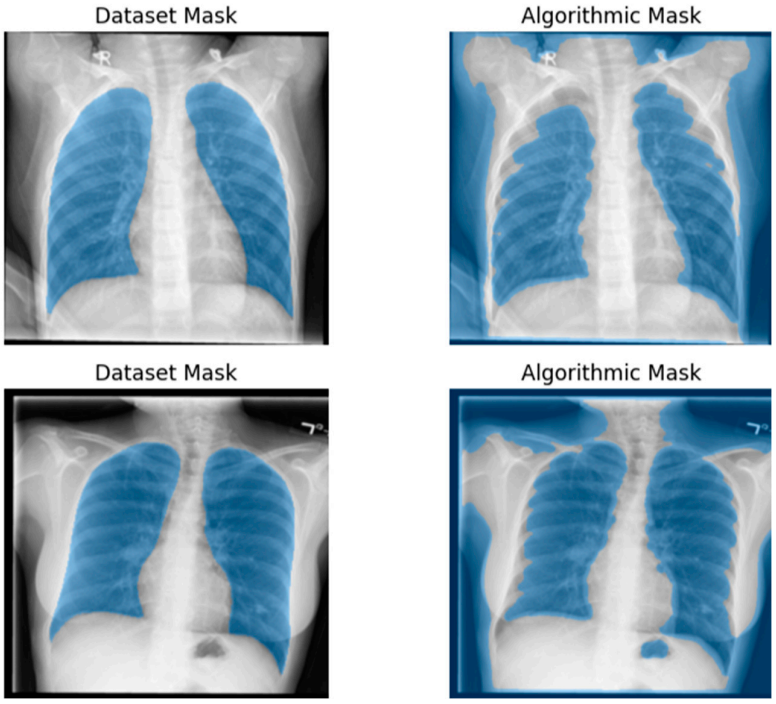


Figure 5. Comparison between dataset-provided lung segmentation masks and algorithmically generated lung segmentation, visualised as overlays on the original chest X-ray image.

As shown in Figure 5, the algorithmically generated lung masks occasionally over-segment into non-lung regions, particularly around the chest wall and shoulder areas. This variability can lead to inconsistent region-of-interest extraction, which can affect the model's performance. In contrast, the dataset-provided paired lung masks offer more stable and reliable lung localisation across samples with smoother edges, making them more suitable for ROI preprocessing and model training in this study.

The safe_lung_mask function follows a structured three-stage process consisting of intensity-based segmentation, geometric filtering, and anatomical validation.

i. Segmentation Phase: Otsu's Thresholding

The otsu_lung_mask_simple step determines an optimal threshold t that separates darker lung regions from brighter surrounding anatomy by maximising the between-class variance. The optimal threshold is obtained by solving:

$$t^* = \arg\max_t \sigma\_b^2(t) \quad (1)$$

where $\sigma\_b^2(t)$ denotes the between-class variance associated with a candidate threshold t.

ii. Formula for Between-Class Variance:

$$\sigma_b^2(t) = \omega_0(t)\omega_1(t)[\mu_0(t) - \mu_1(t)]^2 \quad (2)$$

where:

$\omega_0(t), \omega_1(t)$ are the probabilities of the background and foreground classes, respectively

$\mu_0(t), \mu_1(t)$ are the mean intensities of the two classes.

iii. Binary Mask Result:

$$M(x, y) = \begin{cases} 1, \text{if } I(x, y) < t \\ 0, \text{otherwise} \end{cases} \quad (3)$$

where $I(x, y)$ denotes the image intensity at pixel location (x, y).

iv. Geometric Phase: Connected Components

The binary mask was treated as a set of connected components C_i. Each component's area is computed as:

 doi:10.20944/preprints202602.1063.v1

$$A_i = \sum_{(x,y)\in C_i} M(x,y) \quad (4)$$

where $M(x,y) = 1$ for foreground pixels and 0 otherwise.

v. Assuming the lungs correspond to the two largest contiguous dark regions, the mask was filtered by retaining the two largest components:

$$M_{\text{filtered}} = \cup\{C_i \mid \text{rank}\,(A_i) \in \{1,2\}\} \quad (5)$$

vi. Formula for Component Area:

$$A_i = \Sigma(x,y) \in C_i 1 \quad (6)$$

vii. Validation Phase: Area Fraction Heuristic

To ensure anatomical plausibility, the fraction of the image occupied by the detected lung region was computed:

$$\text{Area Fraction} = \frac{\sum_{x=1}^{W} \sum_{y=1}^{H} M_{\text{filtered}}\,(x,y)}{H \times W} \quad (7)$$

where H and W denote the image height and width.

viii. Decision Rule

$$M_{\text{final}} = \begin{cases} M_{\text{filtered}}, & \text{if } 0.05 < \text{Area Fraction} < 0.80 \\ 1_{H\times W}, & \text{otherwise (fallback)} \end{cases} \quad (8)$$

This constraint ensures that only anatomically reasonable lung masks are accepted. Although algorithmically generated lung masks were not adopted for ROI preprocessing, they were still used during Grad-CAM analysis to constrain percentage activation measurements to the pulmonary region, ensuring that saliency measurements reflected model attention within the lung fields while avoiding the introduction of segmentation-related noise into the training pipeline.

II. Soft-Tissue Enhancement: CLAHE and Bone Suppression

Soft-tissue visibility is fundamental for detecting diffuse opacities associated with COVID-19 pneumonia. Contrast Limited Adaptive Histogram Equalisation (CLAHE) is widely used in radiographic image enhancement, and multiple COVID-19 studies have demonstrated that moderate local contrast enhancement improves convolutional neural network sensitivity without distorting anatomical structures or excessively amplifying noise [48,49].

Bones such as the ribs and clavicles can obscure subtle parenchymal changes; bone suppression techniques have therefore been explored to reduce this effect. Early work showed that suppressing rib shadows improves diagnostic accuracy, while more recent studies have confirmed that soft-tissue emphasis benefits the detection of COVID-19-related lesions [43].

Theory: CLAHE applies Histogram Equalisation to small image tiles, and limits contrast amplification using a clipping threshold (clipLimit=3.0).

Formula (General Histogram Equalisation): The transformation function $T(r)$ maps the input intensity r to the output intensity s: The general histogram equalisation transformation is defined as:

$$s = T(r) = (L-1)\sum_{j=0}^{r} p_r\,(j) \quad (9)$$

where:

- $p_r(j)$ is the normalised histogram
- $L = 256$ is the number of grayscale levels

CLAHE applies this transformation locally to image tiles with contrast clipping (clipLimit = 3.0).

where pr(j) is the normalised histogram of the image (or tile), and L is the number of intensity levels (256).

Bone suppression uses lightweight approximation to enhance fine details and reduce high-contrast bony structures.

Theory: The image was decomposed into a base (low-frequency structures) and a detail component (high-frequency texture). Reducing the detail component suppresses bone structures. A bilateral filter was used as it preserves edges better than Gaussian filtering.

Filter Used (Bilateral Filter): The output intensity I_F at pixel x is: Bilateral Filter

 doi:10.20944/preprints202602.1063.v1

$$I_F(x) = \frac{1}{W_x} \sum_{\xi \in \Omega} I(\xi)\, f(\| x - \xi \|)\, g(| I(x) - I(\xi) |) \qquad (10)$$

where:

W_x is a normalisation factor

$f(\cdot)$is the spatial Gaussian kernel (controlled by σ_{space})

$g(\cdot)$is the range Gaussian kernel (controlled by σ_{color}).

Approximation Formula in Code:

Base = BilateralFilter(Img) (11)

Detail = Img – Base (12)

Output = Base + (0.18 Detail) (13)

This study, therefore, combined CLAHE with bilateral-filter-based bone suppression to enhance soft-tissue visibility while reducing structural noise, improving interpretability and discriminative ability with lower computational cost than learning-based suppression models. Figure 6 represents raw chest X-ray images compared with pulmonary region of interest (ROI) representations following lung segmentation, bone suppression, and contrast enhancement. Typical raw chest X-ray images VS pulmonary region of interest (ROI) representations are shown in Figure 6.

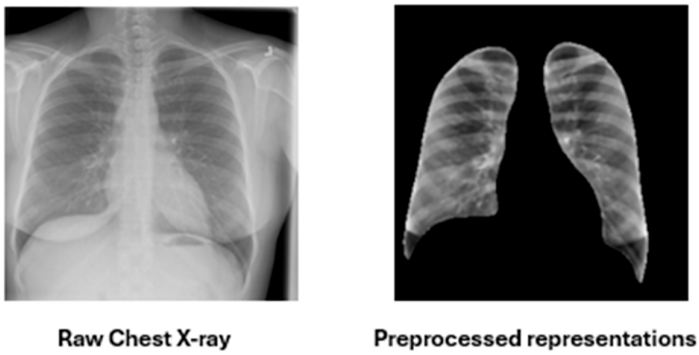


Figure 6. Raw chest X-ray images VS pulmonary region of interest (ROI) representations.

3.1.2. Multi-Channel Feature Construction

Following core preprocessing, a multi-channel representation is constructed to provide complementary views of the same lung region. Rather than relying on a single greyscale image, this approach encodes multiple radiologically meaningful characteristics into separate channels, supporting richer and more anatomically aligned feature learning [19,20]. The lung-masked region of interest (ROI) forms the base channel, ensuring that all derived representations focus exclusively on clinically relevant pulmonary tissue. Frequency-based representations are included to emphasise diffuse opacity patterns commonly associated with infectious lung disease [50]. A vessel-enhanced channel highlights pulmonary vascular and airway-related structures linked to inflammatory changes [51–53]. Finally, a texture-based channel captures local lung texture variations that support differentiation between normal and pathological radiographic patterns [54].

I. Mid-frequency opacity mapping (Fourier band-pass filtering)

Mid-frequency textures associated with ground-glass opacities have been identified using radiomics-based texture analysis techniques [50]. Band-pass filtering improves visualisation by suppressing low-frequency illumination variations and high-frequency noise, thereby enhancing diagnostically relevant structural patterns. Prior medical imaging studies supported the diagnostic value of frequency-based feature decomposition for disease characterisation.

Theory: The 2D Discrete Fourier Transform (DFT) converts images to the frequency domain. A band-pass filter preserves frequencies within a radial range (r_1, r_2).

Formulas:

 doi:10.20944/preprints202602.1063.v1

a) 2D DFT:

$$F(u,v) = \sum_{x=0}^{M-1}\sum_{y=0}^{N-1} f(x,y)exp\left[-j2\pi\left(\frac{ux}{M}+\frac{vy}{N}\right)\right] \quad (14)$$

b) Distance to Center:

$$d(i,j) = \sqrt{(i-c_{row})^2+(i-c_{col})^2} \quad (15)$$

c) Filter Mask (Ideal Band-Pass):

$$Mask(i,j) = \begin{cases} 1, & r_1 < d(i,j) < r_2 \\ 0, & otherwise \end{cases} \quad (16)$$

d) Filtered Transform:

$$F_{filtered}(i,j) = F_{shifted}(i,j) \cdot Mask(i,j) \quad (17)$$

e) Inverse DFT:

$$Output(x,y) = \left|\mathcal{F}^{-1}\{F_{unshifted}(i,j)\}\right| \quad (18)$$

II. Vessel enhancement using the Frangi filter

The Frangi Vesselness filter enhances tubular structures by emphasising vascular morphology and suppressing background noise [51]. COVID-19 chest imaging studies have reported vascular thickening and dilation associated with inflammatory and thrombotic processes [52]. Prior pulmonary imaging research has demonstrated that vessel-enhanced representations can improve diagnostic sensitivity for pulmonary disease patterns [53].

Theory: The Frangi filter analyses the Hessian matrix and its eigenvalues to identify line-like structures across multiple scales. The eigenvalues (λ_1, λ_2) of the Hessian indicate the directions and magnitudes of maximum and minimum curvature.

Hessian Matrix:

$$\boldsymbol{H_I}(x,y)\begin{bmatrix} \frac{\partial^2 I(x,y)}{\partial x^2} & \frac{\partial^2 I(x,y)}{\partial x \partial y} \\ \frac{\partial^2 I(x,y)}{\partial y \partial x} & \frac{\partial^2 I(x,y)}{\partial y^2} \end{bmatrix} \quad (19)$$

The maximum vesselness response across scales (1–8) is retained.

III. Texture encoding using Local Binary Patterns (LBP)

LBP are established texture descriptor for capturing local intensity variations and micro-texture information in images [54]. Prior chest X-ray studies have demonstrated that LBP features complement convolutional neural network representations by capturing fine-grained texture variations present in COVID-19 radiographs [19,55].

Theory: LBP compares a centre pixel with neighbouring pixels to generate a binary code.

LBP Formula (General):

$$LBP_{P,R}(x_c,y_c) = \sum_{p=0}^{P-1} s(i_p\ i_c)2^p \quad (20)$$

where:

i_c is the gray value of the center pixel (x_c, y_c).

i_p is the gray value of the p-th neighbor.

P is the number of neighbors (P=8).

R is the radius of the circle (R=1).

$s(x)$ is the step function: $s(x) = \begin{cases} 1\ if\ x \geq 0 \\ 0\ if\ x < 0 \end{cases}$

Collectively, these channels augment the soft-tissue ROI to form a 4-channel tensor (ROI, Frequency, Vessel, LBP). This approach advances beyond single-channel studies and aligns with hybrid feature-learning strategies [19,55]. Figure 7 represents Four-channel lung-focused feature representations.

 doi:10.20944/preprints202602.1063.v1

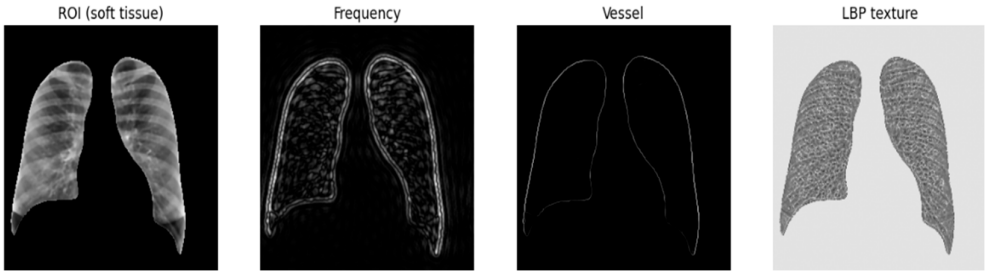


Figure 7. Four-channel lung-focused feature representations prepared for model input.

3.2. Normalisation, Class Balancing, and Augmentation

Minimum–maximum scaling was applied to preserve relative tissue intensity while ensuring numerical stability during training. Class imbalance was addressed using a controlled over-sampling strategy to balance COVID-positive and non-COVID samples. Conservative data augmentation techniques, including horizontal flipping and small rotations, were applied to simulate acquisition variability while preserving radiological realism.

3.3. Model Architecture

The modelling phase involved constructing a deep convolutional neural network based on the Xception architecture. Xception was selected due to its effectiveness in medical imaging tasks requiring fine-grained feature extraction and its use of depth wise separable convolutions. The architecture was modified to accept four-channel input tensors corresponding to the proposed multi-channel representation. The first convolutional layer was adapted accordingly, and the classification head was replaced with a fully connected layer producing a single output logit for binary classification. Transfer learning was employed, and optimisation strategies were selected to ensure stable convergence and robustness to class imbalance. Conceptual architecture of the proposed 4-channel xception-based model for this study is provided in Figure 8.

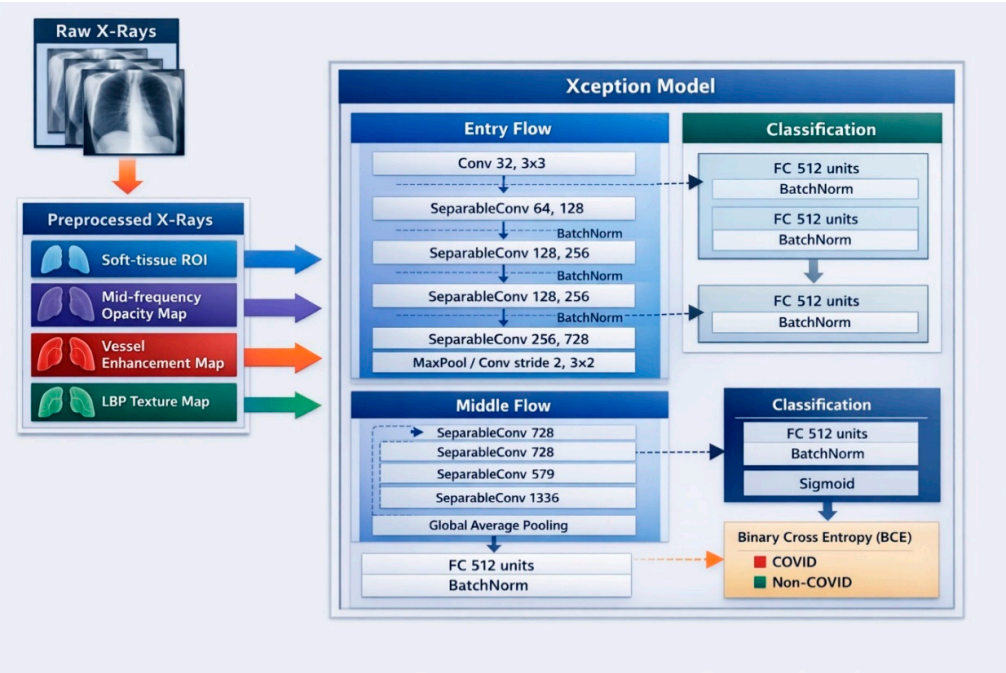


Figure 8. Conceptual Architecture of the Proposed 4-Channel Xception-Based Model for this Study.

 doi:10.20944/preprints202602.1063.v1

3.4. Explainability Integration

Explainability was incorporated as a core component of the proposed framework rather than as a post hoc addition. Gradient-weighted Class Activation Mapping (Grad-CAM) was used to generate localisation heatmaps highlighting image regions contributing to model predictions.

The explainability framework is strengthened through lung-field masking, ensuring that activation patterns reflect clinically meaningful pulmonary regions using lung-region coverage analysis. Both qualitative inspection and quantitative lung-region coverage analysis were used to assess the anatomical relevance of model attention, supporting transparent and clinically grounded interpretation.

3.5. Quantitative Classification Performance

Explainability is a critical requirement in medical AI systems, particularly in radiology, where model predictions must be grounded in clinically meaningful image regions. In this study, explainability was implemented using Gradient-weighted Class Activation Mapping (Grad-CAM), which produced spatial heatmaps indicating image regions that contribute most strongly to the model's prediction by backpropagating gradients from the final convolutional layer. While visual inspection of Grad-CAM heatmaps provides intuitive insight, visual explanations alone are subjective and insufficient for rigorous evaluation. To address this limitation, this study incorporated a quantitative explainability assessment based on lung-region coverage and CAM energy distribution, enabling objective measurement of anatomical relevance.

Two complementary quantitative measurements are used:

(a) CAM Energy

CAM energy represents the total activation strength of the Grad-CAM heatmap and is computed as the sum of all pixel intensities in the heatmap:

$$E_{\text{CAM}} = \sum_{x,y} H(x,y) \qquad (21)$$

where $H(x,y)$is the Grad-CAM activation value at pixel location (x, y).

This value reflects how strongly the model attends to image regions overall for a given prediction.

(b) Lung-Region CAM Energy Coverage

To assess anatomical relevance, CAM energy is separated into contributions **inside** and **outside** the lung region using a binary lung mask $M(x,y)$:

$$E_{\text{lung}} = \sum_{x,y} H(x,y) \cdot M(x,y) \qquad (22)$$

$$E_{\text{total}} = \sum_{x,y} H(x,y) \qquad (23)$$

The lung-region coverage ratio is then defined as:

$$\text{Lung Coverage (\%)} = \frac{E_{\text{lung}}}{E_{\text{total}}} \times 100 \qquad (24)$$

This metric quantifies the proportion of model attention focused within anatomically valid pulmonary regions.

4. Experimental Results and Analysis

4.1. Implementation Setup

This section presents the experimental evaluation results for the proposed explainable deep learning framework for Covid binary classification. The dataset was divided into training, validation, and test sets using stratified sampling to preserve class proportions. A 70%–15%–15% split was adopted, which is widely used in medical AI to provide a reliable assessment of model generalisation when evaluating heterogeneous clinical data. Model performance is assessed using standard classification metrics and threshold analysis, including accuracy, precision, recall (sensitivity), F1 score, Matthews Correlation Coefficient (MCC), and area under the receiver operating characteristic

 doi:10.20944/preprints202602.1063.v1

curve (AUC), while explainability was evaluated through qualitative visualisation and quantitative lung-region attention analysis. The objective was to demonstrate diagnostic performance, robustness, and clinically meaningful interpretability.

4.2. Model Training Configuration

The model was trained for 20 epochs using a balanced training dataset, with performance monitored on the validation set. Training incorporated Automatic Mixed Precision (AMP), gradient clipping (threshold = 5.0), cosine annealing learning rate scheduling with warm-up, and Early Stopping was based on the validation Matthews Correlation Coefficient (MCC) rather than the F1 score, because MCC provides a more reliable performance indicator under class imbalance conditions [56]. The model was compiled using the Focal Loss function, selected over binary cross-entropy due to robustness to class imbalance and ability to down-weight easy negatives. Optimisation was performed using AdamW, shown to improve convergence stability in medical imaging networks. Training metrics included loss, accuracy, precision, recall, and F1 score. The training history demonstrated stable convergence, with training and validation losses decreasing in parallel, indicating appropriate model capacity and optimisation strategy.

4.3. Quantitative Classification Performance

The results are summarised in Table 2 and the confusion matrix shown in Figure 9. Initial evaluation was conducted using a default probability threshold of 0.50. At this threshold, the model achieved an accuracy of 95.3%, precision of 88.2%, recall of 83.8%, F1 score of 85.9%, MCC of 0.83, and an AUC of 0.983. The confusion matrix indicated that 2,572 non-COVID images were correctly classified, while 454 COVID-positive cases were correctly identified, with 88 false negatives and 61 false positives observed at this operating point.

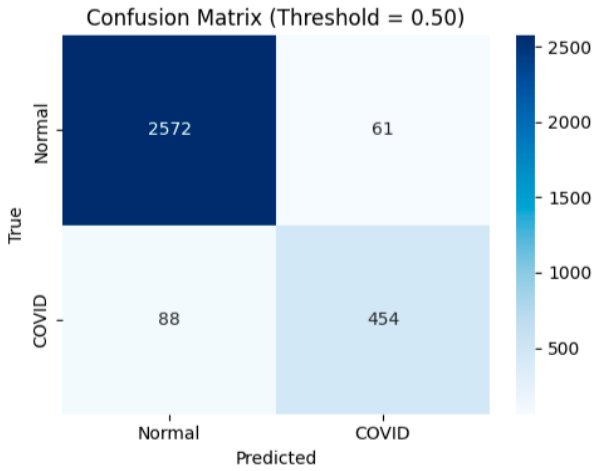


Figure 9. Confusion matrix at default threshold (0.50).

While the default threshold yielded strong overall performance, medical diagnosis often requires prioritising sensitivity. False negative predictions, corresponding to missed COVID-19 cases, pose a greater clinical risk than false positive classifications, particularly in screening-oriented applications. The model was therefore evaluated across probability thresholds ranging from 0.05 to 0.95 to identify an operating point that better balances sensitivity and specificity.

 doi:10.20944/preprints202602.1063.v1

Table 2. Model Evaluation Metrics at different thresholds.

Threshold	Accuracy	Precision	Recall	F1	MCC
0.05	0.70	0.37	0.99	0.53	0.48
0.10	0.82	0.49	0.99	0.65	0.61
0.15	0.88	0.58	0.98	0.73	0.69
0.20	0.91	0.67	0.97	0.80	0.76
0.25	0.93	0.71	0.95	0.81	0.78
0.30	0.94	0.76	0.93	0.84	0.81
0.35	0.95	0.81	0.91	0.86	0.83
0.40	0.95	0.83	0.89	0.86	0.83
0.45	0.95	0.85	0.87	0.86	0.83
0.50	0.95	0.88	0.84	0.86	0.83
0.55	0.95	0.90	0.82	0.86	0.83
0.60	0.95	0.93	0.78	0.85	0.82
0.65	0.95	0.94	0.74	0.83	0.80
0.70	0.94	0.96	0.69	0.80	0.78
0.75	0.94	0.98	0.65	0.78	0.77
0.80	0.92	0.98	0.57	0.72	0.71
0.85	0.91	0.99	0.48	0.65	0.66
0.90	0.89	1.00	0.37	0.54	0.57
0.95	0.86	0.99	0.20	0.34	0.41

At a threshold of 0.40, the model achieved an accuracy of 95.1%, a precision of 83.4%, a recall of 89.1%, an F1 score of 86.2%, and the highest MCC (approximately 0.83). Lowering the threshold from 0.50 to 0.40 reduced the number of missed COVID-19 cases from 88 to 59, while increasing false positives from 61 to 96. This trade-off is clinically acceptable in screening contexts, as it substantially improves sensitivity at the cost of a moderate increase in false alarms. The confusion matrix at selected operating threshold 0.40 is shown in Figure 10.

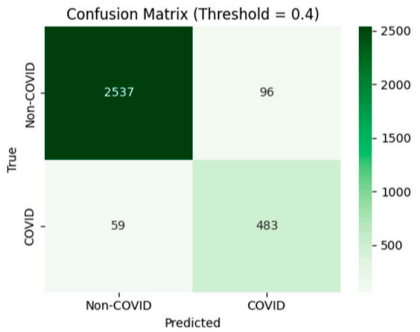


Figure 10. Confusion matrix at selected operating threshold (0.40),.

 doi:10.20944/preprints202602.1063.v1

The ROC curve (Figure 11) demonstrates strong class separability, with an AUC of approximately 0.98.

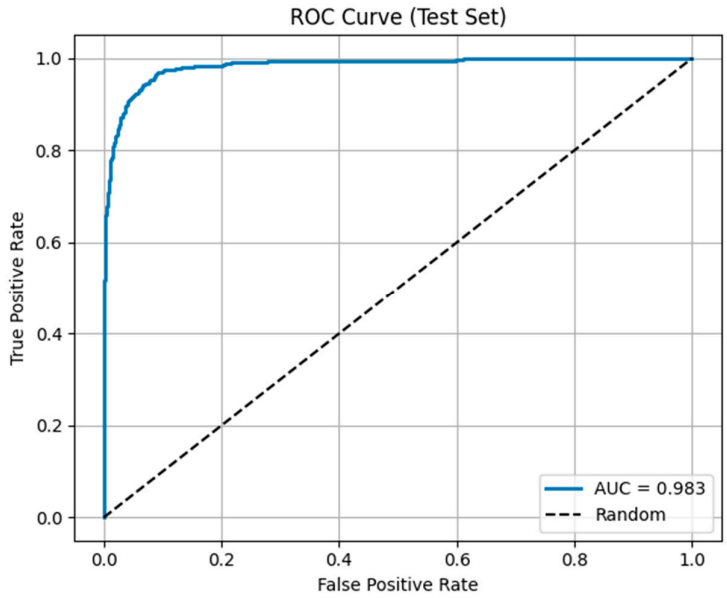


Figure 11. Receiver Operating Characteristic (ROC) curve on the test set.

Further analysis of MCC across thresholds (shown in Figure 12) revealed a broad optimal operating region between approximately 0.35 and 0.45, indicating robustness to minor threshold variations.

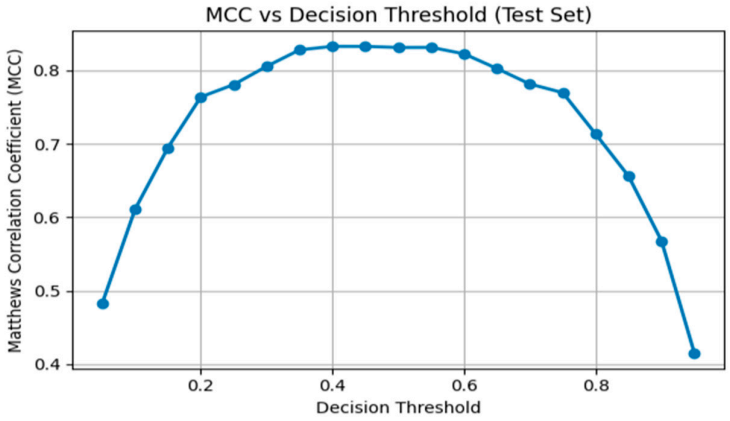


Figure 12. Matthews Correlation Coefficient (MCC) as a function of decision threshold.

4.4. Quantitative Classification Performance

Explainability is a critical requirement in medical AI systems, particularly in radiology, where model predictions must be grounded in clinically meaningful image regions. In this study, explainability was implemented using Gradient-weighted Class Activation Mapping (Grad-CAM), which produced spatial heatmaps indicating image regions that contribute most strongly to the model's prediction by backpropagating gradients from the final convolutional layer. While visual inspection of Grad-CAM heatmaps provides intuitive insight, visual explanations alone are subjective and insufficient for rigorous evaluation. To address this limitation, this study incorporated

 doi:10.20944/preprints202602.1063.v1

a quantitative explainability assessment based on lung-region coverage and CAM energy distribution, enabling objective measurement of anatomical relevance.

Two complementary quantitative measurements are used:

(a) CAM Energy

CAM energy represents the total activation strength of the Grad-CAM heatmap and is computed as the sum of all pixel intensities in the heatmap:

$$E_{\text{CAM}} = \sum_{x,y} H(x,y) \qquad (20)$$

where $H(x,y)$is the Grad-CAM activation value at pixel location (x, y).

This value reflects how strongly the model attends to image regions overall for a given prediction.

(b) Lung-Region CAM Energy Coverage

To assess anatomical relevance, CAM energy is separated into contributions **inside** and **outside** the lung region using a binary lung mask $M(x,y)$:

$$E_{\text{lung}} = \sum_{x,y} H(x,y) \cdot M(x,y) \qquad (21)$$

$$E_{\text{total}} = \sum_{x,y} H(x,y) \qquad (22)$$

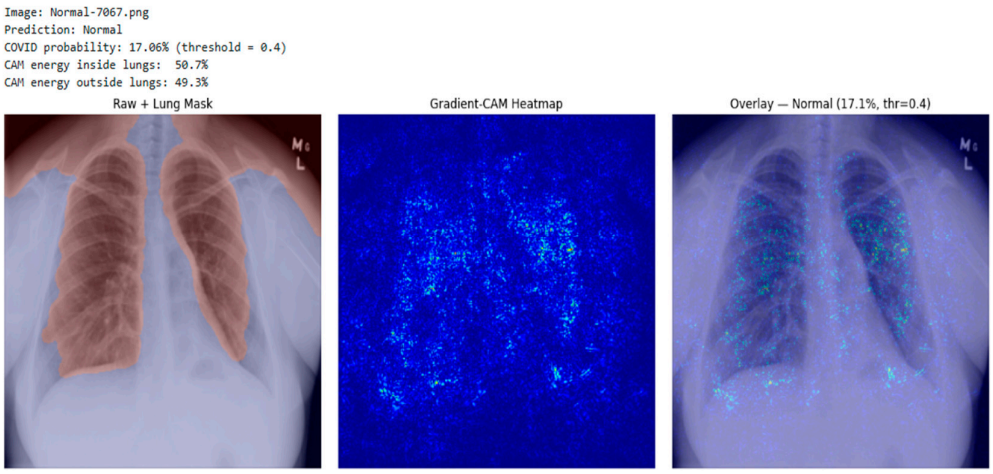
The lung-region coverage ratio is then defined as:

$$\text{Lung Coverage (\%)} = \frac{E_{\text{lung}}}{E_{\text{total}}} \times 100 \qquad (23)$$

This metric quantifies the proportion of model attention focused within anatomically valid pulmonary regions.

5. Discussion

Examples of the grad-CAM explainability analysis for chest X-ray classification are included in Figure 13. Higher lung-region CAM energy indicated that the model relies predominantly on clinically relevant lung structures such as parenchymal textures and opacity patterns when making predictions. Conversely, lower coverage suggests reliance on non-diagnostic cues such as image borders, background artefacts, or acquisition markers. Increased concentrations of CAM energy within lung fields therefore provided objective evidence that the model's decision-making process is anatomically aligned and clinically meaningful, rather than driven by spurious correlations. This quantitative evaluation strengthens confidence in the interpretability and reliability of the model beyond qualitative heatmap inspection alone.



 doi:10.20944/preprints202602.1063.v1

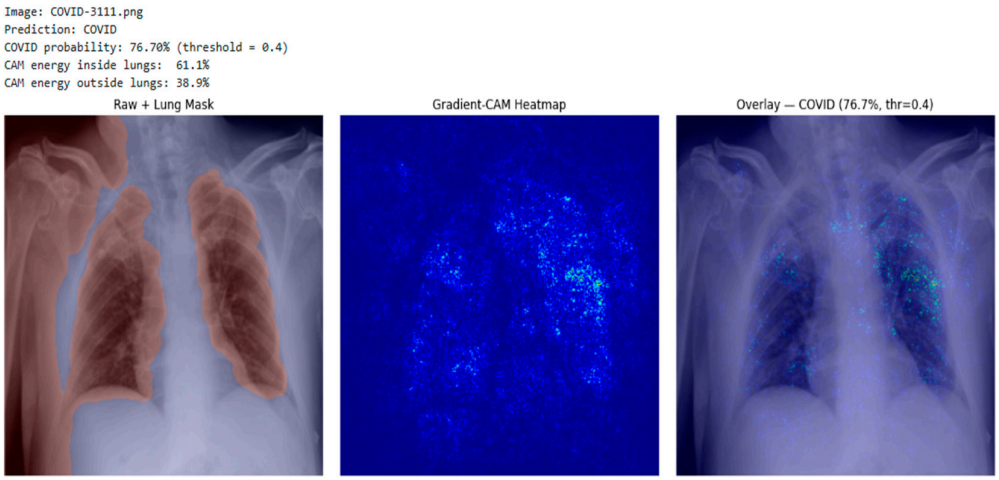


Figure 13. Grad-CAM explainability analysis for chest X-ray classification. Shown are (left) lung-masked chest X-ray, (centre) Grad-CAM heatmap, and (right) heatmap overlaid on the original image.

Quantitative evaluation was based on CAM energy distribution and lung-region coverage, measuring the proportion of model attention located within anatomically valid pulmonary regions. This explainability stage ensured that the developed COVID-19 classification model demonstrates behaviour that is interpretable, clinically grounded, and suitable for integration into diagnostic support workflows.

Explainability was incorporated as a core component of the framework through the integration of Grad-CAM and lung-region coverage analysis. Visual explanations and quantitative attention measurements confirmed that model predictions were predominantly driven by anatomically meaningful pulmonary regions rather than background artefacts. This combination of qualitative and quantitative explainability supports transparent and clinically grounded interpretation of model outputs and increases diagnostic confidence. A lightweight prototype deployment was also developed to demonstrate user interaction with the trained model. The interface enables users to upload chest X-ray images and receive classification results accompanied by probability scores, visual explanations, and concise textual interpretation. This demonstration highlights the feasibility of integrating explainable deep learning models into practical diagnostic support workflows.

Overall, the findings indicate that combining anatomical guidance, a novel four-channel feature representation, and explainable artificial intelligence techniques can yield robust predictive performance while maintaining high interpretability. The proposed framework provides a foundation for explainable chest X-ray classification systems and supports the safe and trustworthy adoption of deep learning methods in medical imaging.

Future research should focus on extending the framework to multi-class respiratory disease classification, performing external and multi-centre validation to assess generalisability, incorporating clinical metadata to enhance contextual relevance, and evaluating performance through prospective clinical studies. Continued development of human-centred explainability methods will further strengthen the role of explainable deep learning systems in real-world healthcare applications.

The proposed explainable DL framework offered a robust and interpretable solution for COVID-19 detection from chest X-rays, supporting its potential integration into clinical decision-support workflows.

6. Conclusions

This paper presented an explainable deep learning framework for COVID-19 detection from chest X-ray images, addressing the need for accurate, interpretable, and clinically relevant diagnostic support systems. The proposed approach integrates anatomically guided preprocessing, a novel

 doi:10.20944/preprints202602.1063.v1

four-channel input representation, and explainable artificial intelligence techniques to enhance both predictive performance and transparency.

A modified Xception-based convolutional neural network was developed to process a four-channel representation comprising lung-isolated soft-tissue images, mid-frequency opacity maps, vessel enhancement maps, and texture-based features. This multi-channel formulation extends beyond conventional single-channel chest X-ray analysis and provides complementary anatomical, frequency-domain, vascular, and texture information. Experimental evaluation demonstrated strong classification performance, achieving high accuracy, recall, F1 score, Matthews Correlation Coefficient, and AUC. Threshold analysis further identified an operating point that prioritised sensitivity, reducing missed COVID-19 cases and aligning model behaviour with screening-oriented clinical requirements.

Author Contributions: Conceptualization, D.N-O. and O.S.; methodology, D.N-O. and O.S.; software, D.N-O. and O.S.; validation, O.S., M.K., R.S and O.O.; formal analysis, D.N-O. and O.S.; investigation, D.N-O. and O.S.; resources, O.S.; data curation, D.N-O.; writing—original draft preparation, D.N-O. and O.S.; writing—review and editing, O.S., M.K., R.S and O.O.; supervision, O.S.; project administration, O.O., M.K., R.S., and O.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not Applicable.

Data Availability Statement: The dataset is available at https://www.kaggle.com/datasets/tawsifurrahman/covid19-radiography-database (accessed on 25 September 2025).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Mercaldo, F.; Belfiore, M.P.; Reginelli, A.; Brunese, L.; Santone, A. Coronavirus covid-19 detection by means of explainable deep learning. *Scientific Reports* **2023**, *13*, 462.
2. Chadaga, K.; Prabhu, S.; Sampathila, N.; Chadaga, R.; Umakanth, S.; Bhat, D.; GS, S.K. Explainable artificial intelligence approaches for COVID-19 prognosis prediction using clinical markers. *Scientific Reports* **2024**, *14*, 1783.
3. Pham, N.T.; Ko, J.; Shah, M.; Rakkiyappan, R.; Woo, H.G.; Manavalan, B. Leveraging deep transfer learning and explainable AI for accurate COVID-19 diagnosis: Insights from a multi-national chest CT scan study. *Comput. Biol. Med.* **2025**, *185*, 109461.
4. Litjens, G.; Kooi, T.; Bejnordi, B.E.; Setio, A.A.A.; Ciompi, F.; Ghafoorian, M.; Van Der Laak, J.A.; Van Ginneken, B.; Sánchez, C.I. A survey on deep learning in medical image analysis. *Med. Image Anal.* **2017**, *42*, 60–88.
5. Esteva, A.; Robicquet, A.; Ramsundar, B.; Kuleshov, V.; DePristo, M.; Chou, K.; Cui, C.; Corrado, G.; Thrun, S.; Dean, J. A guide to deep learning in healthcare. *Nat. Med.* **2019**, *25*, 24–29.
6. Shobayo, O.; Saatchi, R. Developments in Deep Learning Artificial Neural Network Techniques for Medical Image Analysis and Interpretation. Diagnostics **2025**, 15, 1072.
7. Esteva, A.; Kuprel, B.; Novoa, R.A.; Ko, J.; Swetter, S.M.; Blau, H.M.; Thrun, S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **2017**, *542*, 115–118.
8. Rajpurkar, P.; Irvin, J.; Zhu, K.; et al. CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning. *arXiv* **2017**, arXiv:1711.05225.
9. El-Magd, L.M.A.; Dahy, G.; Farrag, T.A.; Darwish, A.; Hassnien, A.E. An interpretable deep learning based approach for chronic obstructive pulmonary disease using explainable artificial intelligence. *International Journal of Information Technology* **2025**, *17*, 4077–4092
10. Adadi, A.; Berrada, M. Peeking inside the black box: A survey on explainable artificial intelligence (XAI). *IEEE Access* **2018**, *6*, 52138–52160.

 doi:10.20944/preprints202602.1063.v1

11. Solayman, S.; Aumi, S.A.; Mery, C.S.; Mubassir, M.; Khan, R. Automatic COVID-19 prediction using explainable machine learning techniques. *International Journal of Cognitive Computing in Engineering* **2023**, *4*, 36–46
12. Wachter, S.; Mittelstadt, B.; Floridi, L. Why a right to explanation of automated decision-making does not exist in the GDPR. *Int. Data Priv. Law* **2017**, *7*, 76–99.
13. Singh, J.; Sillerud, B.; Yednock, J.; Larson, C.; Steffen, A.; Singh, A. Healthcare leaders' attitudes and perceptions on the use of artificial intelligence and artificial intelligence enabled tools in healthcare settings. *Journal of Medical Artificial Intelligence* **2025**, *8*, 41.
14. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. *Proc. IEEE ICCV* **2017**, 618–626.
15. Gulum, M.A.; Trombley, C.M.; Kantardzic, M. Explainable deep learning for medical image analysis: A survey. *J. Imaging* **2021**, *7*, 102.
16. Samek, W.; Montavon, G.; Vedaldi, A.; Hansen, L.K.; Müller, K. *Explainable AI: interpreting, explaining and visualizing deep learning;* Springer Nature: 2019; Vol. 11700.
17. Holzinger, A.; Langs, G.; Denk, H.; Zatloukal, K.; Müller, H. Causability and explainability of artificial intelligence in medicine. *WIREs Data Min. Knowl. Discov.* **2019**, *9*, e1312.
18. Slack, D.; Hilgard, A.; Jia, E.; Singh, S.; Lakkaraju, H. Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. *Proc. AAAI Conf. Artif. Intell.* **2020**, *34*, 180–187.
19. Nneji, G.U.; Cai, J.; Deng, J.; Su, S. Multi-channel deep learning framework for chest X-ray classification. *Comput. Biol. Med.* **2022**, *145*, 105410.
20. Çallı, E.; Sogancioglu, E.; van Ginneken, B.; van Leeuwen, K.G.; Murphy, K. Deep learning for chest X-ray analysis: A survey. *Med. Image Anal.* **2021**, *72*, 102125.
21. Ait Nasser, A.; Jilbab, A.; Bourouhou, A. Deep learning for chest X-ray analysis: A survey. *Artif. Intell. Rev.* **2023**, *56*, 1243–1290.
22. Arun, N.; Gaw, N.; Singh, P.; Chang, K.; Aggarwal, M.; Chen, B.; Hoebel, K.; Gupta, S.; Patel, J.; Gidwani, M.; et al. Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging. *Radiol. Artif. Intell.* **2021**, *3*, e200267.
23. Singh, P.; Gaw, N.; Chang, K.; et al. Are saliency maps trustworthy? A large-scale evaluation in medical imaging. *Med. Image Anal.* **2022**, *75*, 102327.
24. Adebayo, J.; Gilmer, J.; Muelly, M.; Goodfellow, I.; Hardt, M.; Kim, B. Sanity checks for saliency maps. *Advances in neural information processing systems* **2018**, *31*.
25. Tjoa, E.; Guan, C. A survey on explainable artificial intelligence (XAI): Toward medical XAI. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *32*, 4793–4813.
26. Wani, N.A.; Kumar, R.; Bedi, J. DeepXplainer: An interpretable deep learning based approach for lung cancer detection using explainable artificial intelligence. *Comput. Methods Programs Biomed.* **2024**, *243*, 107879.
27. Wang, X.; Peng, Y.; Lu, L.; Lu, Z.; Bagheri, M.; Summers, R.M. ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly supervised classification and localization of common thorax diseases. *Proc. IEEE CVPR* **2017**, 2097–2106.
28. Irvin, J.; Rajpurkar, P.; Ko, M.; Yu, Y.; Ciurea-Ilcus, S.; Chute, C.; et al. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *Proc. AAAI Conf. Artif. Intell.* **2019**, 33, 590–597.
29. Tang, Y.; Tang, Y.; Peng, Y.; Yan, K.; Bagheri, M.; Redd, B.A.; Brandon, C.J.; Lu, Z.; Han, M.; Xiao, J. Automated abnormality classification of chest radiographs using deep convolutional neural networks. *NPJ digital medicine* **2020**, *3*, 70.
30. Wang, L.; Lin, Z.Q.; Wong, A. COVID-Net: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images. *Sci. Rep.* **2020**, 10, 19549.
31. Narin, A.; Kaya, C.; Pamuk, Z. Automatic detection of coronavirus disease (COVID-19) using X-ray images and deep convolutional neural networks. *Pattern Anal. Appl.* **2021**, 24, 1207–1220.
32. Cohen, J.P.; Hashir, M.; Brooks, R.; Bertrand, H. On the limits of cross-domain generalization in automated X-ray prediction. *Proc. Med. Imaging Deep Learn.* (MIDL) **2020**, 136-155.

 doi:10.20944/preprints202602.1063.v1

33. Maguolo, G.; Nanni, L. A critic evaluation of methods for COVID-19 automatic detection from X-ray images. *Inf. Fusion* **2021**, 76, 1–7.
34. Zhang, Y.; Chen, H.; Liu, Q. Lung segmentation improves explainability of deep learning models for chest X-ray classification. *Comput. Methods Programs Biomed.* **2021**, 207, 106173.
35. Babu, P.A.; Rai, A.K.; Ramesh, J.V.N.; Nithyasri, A.; Sangeetha, S.; Kshirsagar, P.R.; Rajendran, A.; Rajaram, A.; Dilipkumar, S. An explainable deep learning approach for oral cancer detection. *Journal of Electrical Engineering & Technology* **2024**, *19*, 1837–1848.
36. Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. Adv. Neural Inf. Process. Syst. **2017**, 30, 4765–4774.
37. Aasem, M.; Javed Iqbal, M. Toward explainable AI in radiology: Ensemble-CAM for effective thoracic disease localization in chest X-ray images using weak supervised learning. *Front. Big Data* **2024**, *7*, 1366415, 10.3389/fdata.2024.1366415.
38. Chen, C.; Li, O.; Tao, D.; Barnett, A.; Rudin, C.; Su, J.K. This looks like that: deep learning for interpretable image recognition. *Advances in neural information processing systems* **2019**, *32*, 8930–8941.
39. Koh, P.W.; Nguyen, T.; Tang, Y.S.; Mussmann, S.; Pierson, E.; Kim, B.; Liang, P. In *In Concept bottleneck models;* International conference on machine learning; PMLR: 2020; 5338–5348.
40. Sultana, S.; Hossain, A.A.; Alam, J. COVID-19 detection from optimized features of breathing audio signals using explainable ensemble machine learning. *Results in Control and Optimization* **2025**, *18*, 100538.
41. Pino, C.; Carrara, F.; Bellio, M.; Neri, E. Influence of explainable AI on trust in medical image analysis. *Artif. Intell. Med*. **2021**, 115, 102079.
42. Mohammed, M.A.; Abdulkareem, K.H.; Garcia-Zapirain, B.; Mostafa, S.A.; Maashi, M.S. A comprehensive investigation of deep learning-based COVID-19 detection from chest X-ray images. *Comput. Biol. Med*. **2022**, 136, 104730.
43. Suzuki, K.; Abe, H.; MacMahon, H.; Doi, K. Image-processing technique for suppressing ribs in chest radiographs by means of massive training artificial neural network. *IEEE Trans. Med. Imaging* **2006**, 25, 406–416.
44. Harrison, A.P.; Xu, Z.; George, K.; Lu, L.; Summers, R.M.; Mollura, D.J. Progressive and multi-path holistic lung segmentation from chest X-ray images. *Med. Image Anal*. **2021**, 67, 101840.
45. Saha, M.; Chakraborty, C.; Racoceanu, D. Efficient deep learning model for explainable COVID-19 detection from chest X-ray images. *Diagnostics.* **2021**, 11, 1772.
46. Rahman, T.; Chowdhury, M.E.H.; Khandakar, A.; et al. COVID-19 radiography database. *arXiv.* **2020**, arXiv:2005.06794.
47. Jacobi, A.; Chung, M.; Bernheim, A.; Eber, C. Portable chest X-ray in coronavirus disease-19 (COVID-19): A pictorial review. *Clin. Imaging.* **2020,** 64, 35–42.
48. Pisano, E.D.; Zong, S.; Hemminger, B.M.; et al. Contrast limited adaptive histogram equalization image processing to improve the detection of simulated spiculations in dense mammograms. *J. Digit. Imaging* **1998**, 11, 193–200.
49. Pizer, S.M.; Amburn, E.P.; Austin, J.D.; et al. Adaptive histogram equalization and its variations. *Comput. Vis. Graph. Image Process.* **1990**, 39, 355–368.
50. Chung, A.G.; Shafiee, M.J.; Khalvati, F.; Haider, M.A.; Wong, A. Radiomics-based multi-scale texture analysis for infectious lung disease characterization. *Comput. Biol. Med.* **2021**, 133, 104372.
51. Frangi, A.F.; Niessen, W.J.; Vincken, K.L.; Viergever, M.A. Multiscale vessel enhancement filtering. *Lect. Notes Comput. Sci. (MICCAI)* **1998**, 130–137.
52. Carotti, M.; Salaffi, F.; Sarzi-Puttini, P.; Agostini, A.; Borgheresi, A.; Minorati, D.; Galli, M.; Giovagnoni, A. Chest CT features of coronavirus disease 2019 (COVID-19) pneumonia: Key points for radiologists. *Radiol. Med.* **2020**, 125, 636–646.
53. Zhang, Y.; Wang, S.; Dong, D.; Tian, J.; Zhou, X. Vascular feature learning for pulmonary disease diagnosis. *Med. Image Anal.* **2019**, 58, 101541.
54. Ojala, T.; Pietikäinen, M.; Mäenpää, T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, 24, 971–987.

 doi:10.20944/preprints202602.1063.v1

55. Zhou, S.K.; Greenspan, H.; Davatzikos, C.; Duncan, J.S.; van Ginneken, B.; Madabhushi, A.; Prince, J.L.; Rueckert, D.; Summers, R.M. A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. *Proc. IEEE* **2021**, 109, 820–838.
56. Chicco, D.; Jurman, G. The advantages of the Matthews correlation coefficient over F1 score and accuracy in binary classification evaluation. *BMC Genomics* **2020**, 21, 6.