KFUPM
Journal of Undergraduate Research International

Diamond Open Access

Article *Computer Science*

VGG-16- Based Deep Learning Architecture for Automated Chest X-Ray Diagnosis: Improving Clinical Accuracy and Reducing the Environmental Footprint

Md. Siam Ahmad[1], Md. Faruk Hossen[1] and Mohammad Hasan[1*]

Cite https://doi.org/10.64589/juri/209728

Submitted: June 06, 2025 **Revised:** July 21, 2025 **Accepted:** August 20, 2025

ABSTRACT

Interpretation of chest radiographs is a significant aspect of clinical diagnosis; however, traditional manual reporting systems are time-consuming, cumbersome, and prone to interobserver variability. With global demand for fast, high-quality diagnostic services, particularly in resource-limited settings, automated platforms are becoming increasingly critical. This study presents an artificial intelligence (AI)-based automated reporting system for chest X-ray using deep-learning algorithms to enhance diagnostic quality and support sustainable healthcare delivery. Five state-of-the-art convolutional neural network models, InceptionV3, EfficientNet, DenseNet, ConvNeXt, and VGG-16, were compared and trained on the National Institutes of Health chest X-ray dataset consisting of 10,000 front-view images with corresponding diagnostic labels. The models were evaluated using accuracy, precision, recall, F1 score and area under the curve. VGG-16 achieved the highest performance (93.88% accuracy) and showed superior stability in producing clinically applicable diagnostic reports. Unlike previous studies where models such as ResNet, U-Net with DeCovNet, and initial VGG-16 implementations faced overfitting and task restrictions, our enhanced VGG-16 model addressed these limitations through improved training and multilabel classification. This system increases radiology reporting productivity and reproducibility while promoting sustainable healthcare through paperless digital operations and, thus, reduced environmental impact.

Keywords: chest radiography, deep learning, VGG-16, NIH dataset, radiology automation, sustainability

1. INTRODUCTION

X-ray image interpretation and report generation are integral tasks for medical professionals and typically involve significant time and effort. Manual radiographic analysis can lead to inconsistent or delayed results. This study explores an artificial intelligence (AI)-based solution using convolutional neural networks (CNNs) to automate chest X-ray interpretation and generate radiologist-level reports. Medical imaging is essential for treatment and diagnostic procedures[1], and AI addresses the problems of increasing imaging volumes, staff shortages, and the need for fast yet accurate reporting.

Chest radiography remains one of the most frequently used diagnostic modalities in modern clinical practice because of its availability, low cost, and wide application in diagnosing thoracic diseases such as pneumonia, tuberculosis, and lung carcinoma. However, chest radiograph interpretation is a time-intensive process with risks of interobserver variability and human error. The growing demand for radiological examinations has placed enormous pressure on medical systems, particularly in regions with limited access to trained radiologists, leading to diagnostic and treatment delays.

Deep learning and artificial intelligence technologies provide promising avenues for automating medical image analysis. CNNs have demonstrated impressive performance in computer vision tasks, including medical diagnosis[2]. These models can efficiently learn complex features from medical images, facilitating accurate and efficient disease diagnosis[3].

Among the various CNN architectures, VGG-16 is a robust and efficient model because of its architectural simplicity and depth, achieving high accuracy in image classification tasks[4]. Building on these strengths, we developed an optimized AI model for automated chest X-ray interpretation. Our approach utilizes a pretrained VGG-16 model, trained and tested on 10,000 frontal-view chest X-ray images from the National Institutes of Health (NIR) chest X-ray dataset[5]; each image was linked to up to 14 thoracic disease labels. The primary goal was to enhance diagnostic precision, reduce interpretation inconsistency, and support clinical decision-making through reproducible, automated chest radiograph analysis. The experimental results demonstrate that VGG-16 outperforms several state-of-the-art CNN models with a classification accuracy of 93.88%, highlighting its capability for chest X-ray interpretation.

Additionally, integrating AI systems into radiological processes promotes environmental sustainability through digital, paperless reporting, reducing reliance on traditional film-based imaging methods, and decreasing the environmental footprint of healthcare services[6]. This study highlights the dual benefits of AI in improving diagnostic accuracy while enabling environmentally sustainable medical practices.

The primary contributions of this research include:
Enhancing Radiology Efficiency and Diagnostic Precision: This study proposes an AI-driven system for automated chest X-ray interpretation using the VGG-16 model to improve diagnostic precision, reduce interpretive variability, and decrease radiologist workload through automated thoracic pathology detection.
Improving Diagnostic Accessibility in Under-resourced Settings: Through automated chest radiograph interpretation, the system enhances diagnostic services in rural, remote, and under-resourced healthcare centers lacking professional radiologists.
Standardizing Diagnostic Reporting through AI-Based Systems: The system generates consistent and standardized diagnostic reports that support clinical decision-making and facilitate interdisciplinary communication among healthcare teams.
Advancing Environmentally Sustainable Radiology Practices: The computerized, paperless reporting system promotes green healthcare by reducing dependence on conventional film-based radiology products, thereby supporting environmental sustainability in medical diagnosis.

1.1. Research Questions.

The following research questions were created to meet the stated objectives:
RQ1: How effective is the proposed AI-based chest X-ray interpretation model developed using the NIH Chest X-ray dataset in terms of diagnostic accuracy and consistency?
RQ2: How does the system ensure reliable diagnostic reporting of chest pathologies?
RQ3: What are the primary advantages of using AI-based reporting for chest X-rays compared to conventional manual radiology reporting?
RQ4: How can this AI-driven reporting system enhance healthcare delivery, particularly in settings with limited radiology infrastructure?
RQ5: How do the quality and reliability of AI-generated reports compare to those produced by expert radiologists?
RQ6: What are the quantifiable environmental benefits achievable through transitioning from conventional paper-based radiology practices to AI-based digital reporting systems?
RQ7: What ethical, legal, and clinical challenges must be addressed for the successful integration of AI-based diagnostic systems into routine clinical practice?

1.2. Related Works.

Magalhães et al.[7] utilized NIH and Indiana University (IU) chest X-ray (CXR) datasets with a model combining Swin transformer (image encoder) and GPT-2 (text decoder) with cross-attention (CA) mechanisms. This approach uses hierarchical features and bilingual training (English and Portuguese), achieving ROUGE-L score of 0.404 (NIH) and 0.748 (bilingual), and METEOR score of 0.393 and 0.741, respectively, demonstrating effective coherent report generation.

Dansana et al.[8] evaluated 360 images (18 *Streptococcus*, 295 COVID-19, and 16 SARS) using tuned VGG-19, Inception V2, and decision tree classifiers. VGG-19 achieved the highest performance at 91%, indicating potential utility for early COVID-19 diagnosis; however, this study has been retracted.

Yang et al.[9] presented a model combining a learned knowledge base with multimodal alignment mechanisms to enhance alignment between radiology reports, disease labels, and images. When evaluated on IU-X-ray and MIMIC-CXR datasets, the models achieved state-of-the-art language generation and clinical accuracy, demonstrating the capability for generating relevant and accurate radiology reports.

Pang et al.[10] surveyed deep learning methods for medical report generation using IU X-ray and MIMIC-CXR datasets . These methods include hierarchical recurrent neural networks (RNNs), attention mechanisms, CNN–long short-term memory (LSTM) architectures, and reinforcement learning approaches. The models achieved BLEU-2 scores of 0.829 and ROUGE-L score of 0.701 with high textual similarity. However, these metrics prioritize lexical overlap, requiring additional clinically relevant evaluations.

Liu et al.[11] proposed an innovative solution for automated chest X-ray report generation using a contrastive attention (CA) mechanism with aggregate and differentiate sub-components. When deployed on the IU-X-ray and MIMIC-CXR datasets, the model with multi-attention and CA achieved a clinical F1 score of 0.303 on MIMIC-CXR, demonstrating improved clinical accuracy in AI-based report generation.

El Asnaoui and Chawki et al.[12] developed COVID-19 classification models using chest X-ray and computerized tomography (CT) images from Kermany et al. (“5 856” images) and Cohen et al. (>6000 images) datasets for pneumonia, coronavirus, COVID-19, and normal classifications. They evaluated deep CNNs with transfer learning (VGG-16, DenseNet201, and Inception-ResNet-V2), finding that Inception-ResNet-V2 achieved the highest accuracy (approximately 92.18%), followed by DenseNet201 (approximately 88.09%), demonstrating strong potential for COVID-19 classification.

Zargari Khuzani et al.[13] analyzed 420 chest radiographs (140 each for normal, COVID-19, and non-COVID pneumonia cases). They extracted features (texture, fast-Fourier transform, wavelet, gray level co-occurrence matrix (GLCM) and gray level dependence matrix (GLDM)) with feature reduction and trained a multilayer neural network with two hidden layers. The model achieved approximately 94% accuracy for COVID-19 versus other pneumonia discrimination, proving useful for rapid and accurate diagnosis.

Gifani et al.[14] evaluated 349 COVID-19 positive and 397 negative CT scans using a weighted combination of 15 pretrained CNN models with transfer learning. The weighted ensemble achieved 85.0% accuracy, demonstrating effective automated COVID-19 detection.

Song et al.[15] analyzed 88 COVID-19 cases, 101 bacterial pneumonia cases, and 86 normal cases using a deep-learning pneumonia model with automated deep learning for diagnosis. The model achieved 86.0% accuracy, confirming its effectiveness in diagnosing COVID-19 using CT scans.

2. METHODOLOGY

The main principle of our approach is depicted in the block diagram in Figure 1. In this section, we present a systematic methodology for model creation, data preprocessing, data collection, and

Table 1. *Summary of related works*

Ref.	Dataset	Techniques	Accuracy
Magalhães et al. [7]	NIH Chest X-ray, IU X-ray	Multimodal transformer (Swin Transformer + GPT-2), bilingual datasets	ROUGE-L ≈ 0.404–0.748, METEOR ≈ 0.393-0.741
Dansana et al. [8]	X-ray and CT scan images (360 images)	Tuned VGG-19, Inception V2, decision tree	Best 91% (VGG-19)
Yang et al. [9]	IU-Xray, MIMIC-CXR	Learned knowledge base, multimodal alignment	79.5% (MIMIC-CXR)
Pang et al. [10]	IU X-ray, MIMIC-CXR, COVID-19, spinal and skin image sets	Hierarchical RNN, attention-based frameworks, reinforcement learning	Up to 0.829
Liu et al. [11]	IU X-ray, MIMIC-CXR	Contrastive attention (aggregate and differentiate attention)	F1 score = 0.303
El Asnaoui & Chawki et al. [12]	Chest X-ray and CT (Kermany et al., 2018)	VGG-16, VGG-19, DenseNet201, InceptionV3,ResNet50, Inception-ResNet-V2	Inception-ResNet-V2 ≈ 92.18%
Zargari Khuzani et al. [13]	420 Chest X-rays (normal, COVID-19, non-COVID pneumonia)	Texture, FFT, wavelet, GLCM, GLDM, MLP classifier	≈ 94%
Gifani et al. [14]	349 COVID-19 (+) and 397 COVID-19 (-) CT images	Fifteen pre-trained CNN models	85.00%
Song et al. [15]	88 COVID-19, 101 bacterial pneumonia, 86 normal CT images	Deep pneumonia model	86.00%

performance evaluation to address the research objectives. It discusses the rationale for selecting particular deep learning models and tools to ensure clarity, reproducibility, and contextualization of results within AI-based medical diagnosis.

2.1. Data Collection. We used a subset of the NIH chest X-ray dataset, specifically a collection of 10,000 frontal-view images available on Kaggle for free download. Although the complete dataset contains 112,120 images, we only used a subset. The images were labeled with up to 14 thoracic disease categories (e.g., atelectasis, cardiomegaly, and pneumonia) using natural language processing (NLP) techniques with greater than 90% accuracy. The accompanying comma-separated value (CSV) metadata includes patient ID, age, gender, view position, and image dimensions, which were utilized during preprocessing. Image were resized to 224 × 224 pixels to match the input requirements of the deep learning models (InceptionV3, EfficientNet, ConvNeXt, and VGG-16). All experiments were conducted using Google Colab for efficient model training and testing.

2.2. Preprocessing. Data preprocessing is essential to ensure data quality, consistency, cleanliness, and proper structure for successful analysis.

2.3. Handling Missing Values. The missing value matrix plot (Figure 2) of the X-ray dataset provides a graphical representation of data completeness for each row and feature. Features are depicted as columns and data entries as rows in the plot. The absence of white lines or gaps indicates no missing values in any dataset column. In this image, dark bars represent non-missing values. The dataset contains 10,000 rows and 11 columns, as indicated by the axis labels, and the dark shading across all rows and columns confirms the dataset is complete with on missing values . This ensures that no imputation, deletion, or special treatment of missing values is required during preprocessing, simplifying data preparation and improving subsequent analysis quality.

The bar chart in Figure 3 shows that all columns in the X-ray dataset are complete because all bars reach full height, indicating that no values are missing in any of the features. The plot shows that none of the attributes, Image Index, Finding Labels, Follow-up #, Patient ID, Patient Age, View Position, Original Image Width, Original Image Height, and Original Image Pixel Spacing (x and y), contain missing values, as all bars are at maximum height (value = 1.0) across 10,000 records. This confirms the dataset is complete with no missing data, eliminating the need for deletion or imputation during preprocessing.

2.4. Outlier Handling.

2.4.1. Before Outlier Treatment. Figure 4 shows the original distribution of numerical features in the X-ray dataset using a blue boxplot on the left. The light blue rectangle represents the interquartile range (IQR), which contains 50% of the data (Q1–Q3). The black dashed lines, known as whiskers, extend to 1.5-times the IQR from the quartiles. Points that lie beyond the whiskers, marked as red dots, indicate potential outlier values that deviate significantly from the data.

2.4.2. After Outlier Treatment. Figure 5 shows the distribution following IQR-based outlier capping, using a green boxplot on the right. The green rectangle still represents the middle 50% of the data, but the spread has been reduced. The

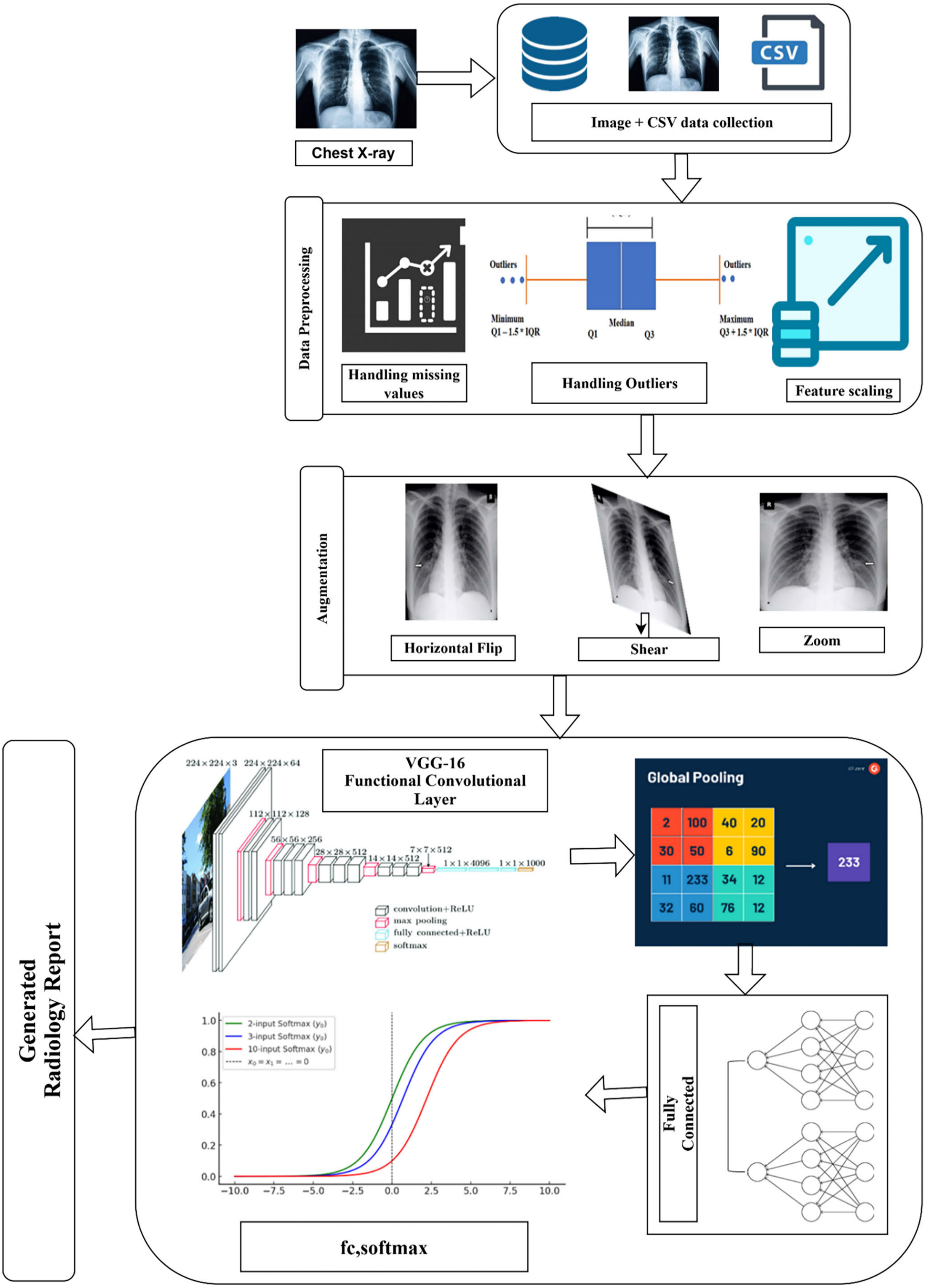


Figure 1. Methodology[16–25]

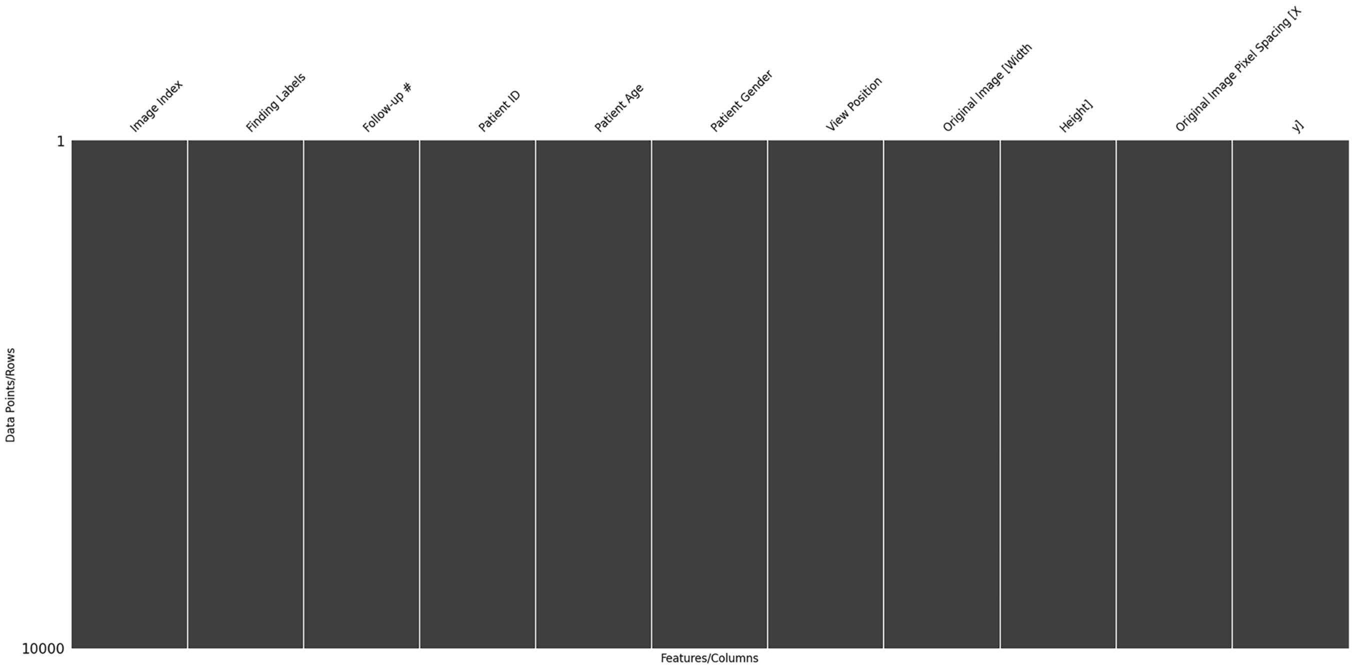


Figure 2. Matrix plot of missing values in the X-ray dataset

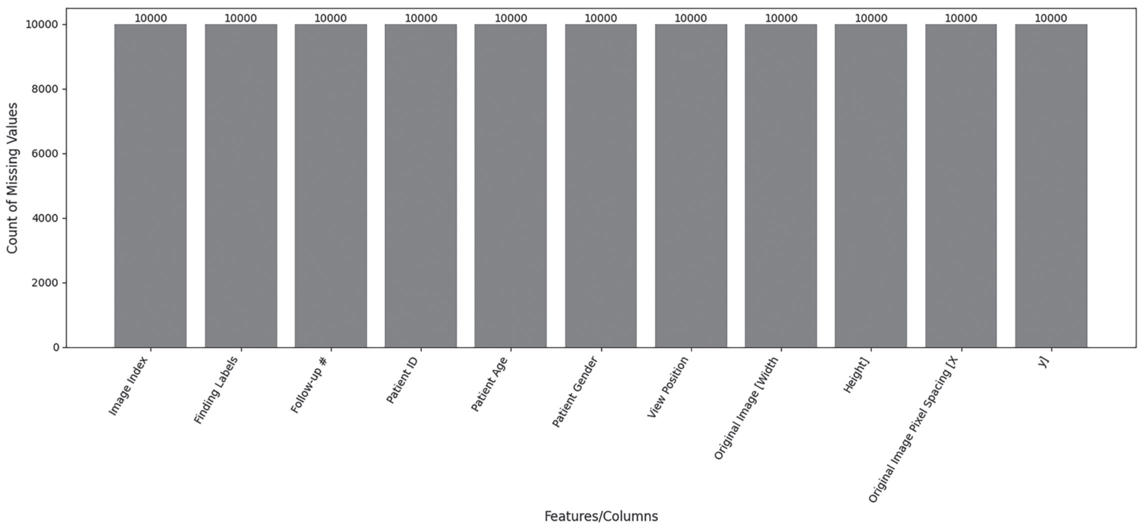


Figure 3. Bar plot of missing values in the X-ray dataset

whiskers are now shorter, reflecting the removal of extreme values. Notably, no red dots appear beyond the whiskers, confirming that outliers have been truncated to the upper and lower bounds.

2.4.3. Dataset Splitting. The dataset was split into 80% training and 20% validation subsets using train-test-split function from Scikit-learn to enable performance validation and mitigate overfitting. This split ensures reproducibility through a fixed random state. The InceptionV3, EfficientNet, ConvNeXt, DenseNet, and VGG-16 models were trained on the training set, whereas the validation set was reserved for performance evaluation and hyperparameter fine-tuning.

2.4.4. Feature Scaling. Feature scaling ensures all input features receive equal weight in the model's learning process by normalizing values to the same range, preventing high-scale features from dominating training. Feature scaling is particularly important when datasets contain features with varying units or scales. Although standardization and min-max scaling are commonly used, this research employed normalization, which rescales pixel intensity values from [0, 255] to [0, 1] by dividing each pixel by 255.0. This method preserved the relative image data structure, improved convergence rates, enhanced numerical stability, and achieved better performance across all investigated architectures (InceptionV3, EfficientNet, ConvNeXt, and VGG-16).

2.4.5. Feature Selection. Feature selection identified the most significant features from the NIH chest X-ray dataset, including image data and metadata, to optimize deep model training. The features presented in Table 2 were selected based on their clinical relevance, classification impact, and data stability.

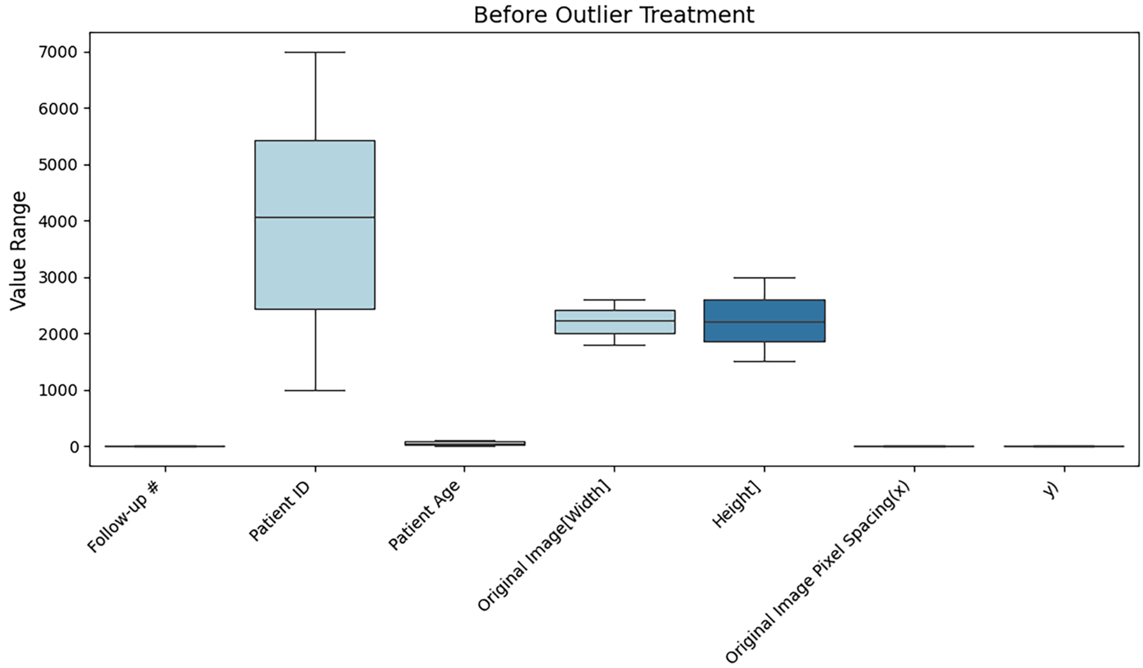


Figure 4. Before outlier treatment

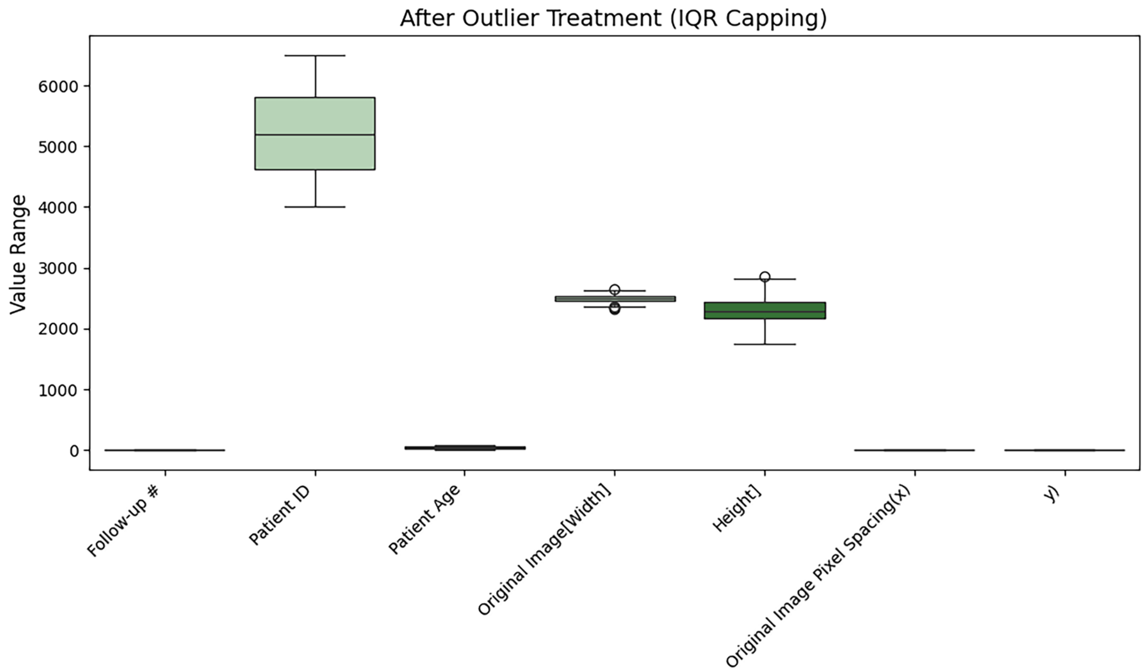


Figure 5. After outlier treatment

Note: Nondiagnostic data, such as Follow-up # and Patient ID, were eliminated because they did not aid diagnosis. The dimensions of the images were normalized by resizing and were not directly utilized for training.

2.5. Data Augmentation. To enhance dataset variability and diversity, we employed an extensive data augmentation strategy during preprocessing. Data augmentation is a common technique used to improve deep-learning-model robustness against variations in image quality while addressing issues related to limited dataset size and overfitting. Our approach utilized various augmentation techniques, including rotation, zooming, height and width translation, horizontal and vertical flipping, and shearing transformations. These augmentations significantly enhanced the training dataset and substantially improved model performance across diverse imaging conditions.

- Total training images (before augmentation): 10,000.
- Total validation images: 2000.

Table 2. *Dataset features*

Feature Name	Description
Image Index	Filename of the chest X-ray image (e.g., 00000001_000.png)
Finding Labels	Text-mined disease labels associated with the image. Multiple labels are separated by the '|' character. A value of 'No Finding' indicates no detected pathology
Follow-up #	Indicates the sequence number of the visit for a patient (e.g., 1 for the first visit)
Patient ID	A unique identifier assigned to each patient in the dataset
Patient Age	Age of the patient at the time the X-ray was taken, in years
Patient Gender	Gender of the patient; values are 'M' for male and 'F' for female
View Position	Position of the patient during the X-ray capture; common values include 'PA' (posteroanterior) and 'AP' (anteroposterior)
Original Image Width	Width of the original image in pixels
Original Image Height	Height of the original image in pixels
Original Image Pixel Spacing X	Physical distance between the centers of adjacent pixels along the x-axis, in millimeters
Original Image Pixel Spacing Y	Physical distance between the centers of adjacent pixels along the y-axis, in millimeters

- Data augmentation applied dynamically during training.

Total training images (after augmentation): 10,000 (size unchanged, content augmented on-the-fly).

2.6. Model Selection.

Five state-of-the-art CNN architectures were selected for developing an effective automated chest X-ray interpretation system based on their diagnostic accuracy, efficiency, and suitability for medical imaging tasks: InceptionV3, EfficientNet, ConvNeXt, DenseNet, and VGG-16.

InceptionV3: Developed by Google, InceptionV3 employs Inception modules with filter banks of varying sizes (1×1, 3×3, and 5×5) to capture features at different scales, making it ideal for diverse chest X-ray pathologies. It utilizes factorized convolutions to reduce computational requirements without compromising performance. This model was selected for its effectiveness in processing high-resolution X-rays and detecting widespread and subtle pathologies.

EfficientNet: EfficientNet scales network depth, width, and resolution proportionally using a compound scaling method controlled by a single hyperparameter, achieving superior accuracy and efficiency. Its lightweight architecture provides low computational overhead for high-resolution image training. This model was chosen for its ability to detect fine features such as early-stage infiltrations or effusions.

ConvNeXt: ConvNeXt modernizes CNNs by incorporating transformer-inspired design elements, including larger 7×7 kernels, inverted bottlenecks, layer normalization, and Gaussian error liner unit (GELU) activation function. It delivers transformer-level performance while maintaining CNN computational efficiency. ConvNeXt was selected for its ability to learn complex thoracic anatomy and generate more informative radiology reports.

DenseNet: DenseNet connects each layer to all preceding layers, promoting feature reuse and improved gradient flow. This architecture excels at identifying subtle variations and shared features in chest X-rays. It was utilized to capture intricate patterns and co-occurring pathologies effectively.

VGG-16: VGG-16 features a straightforward architecture with 13 convolutional layers using small 3×3 filters and three fully connected layers. Despite its simplicity, VGG-16 provides robust feature extraction capabilities through its uniform, deep structure. This model was selected for its proven effectiveness in medical image classification tasks.

All models were set up using pre-trained ImageNet weights and fine-tuned on the NIH chest X-ray dataset. This ensemble of architectures provides performance diversity and robustness for automated report generation.

2.7. Prediction and Evaluation.

The AI-based deep learning models for X-ray reporting underwent a comprehensive evaluation for accuracy, generalizability, and diagnostic performance. Model accuracy was assessed using training data(known) and validation data (unseen). Classification results were analyzed using confusion matrices that recognized true positives, false positives, true negatives, and false negatives. As well evaluation metrics include precision recall (sensitivity) F1 score, and specificity. Sensitivity measured the model's ability to identify positive cases correctly, precision estimated the proportion of correct positive predictions, the F1 score represented the harmonic mean of precision and recall, and specificity evaluated the correct classification of negative cases. These comprehensive metrics ensured robust diagnostic performance assessment of the AI system, supporting healthcare automation and contributing to more efficient, environmentally sustainable healthcare delivery.

3. RESULTS

The results section presents the findings from experimental validation according to the proposed methodology. Empirical results are presented objectively and systematically through tables, figures, and graphs, providing a factual foundation for further discussion and analysis.

3.1. Environmental Setup.

The proposed system was implemented in Python using Keras and TensorFlow and executed on Google Colab. Data preprocessing, training, validation,

Table 3. *Hyperparameters used in the proposed model*

Hyperparameter	Description
Input Size	224 × 224 pixels (resized X-ray images for model compatibility)
Batch Size	16 (images processed per iteration)
Learning Rate	0.0001 (initial rate for feature extraction); 0.00001 (for fine-tuning)
Optimizer	Adam (adaptive stochastic gradient descent)
Loss Function	Binary cross-entropy (multi-label classification)
Epochs	10 (initial training) + 10 (fine-tuning)
Classes	15 thoracic disease classes (multi-label output)
Base Model	VGG-16 (pre-trained on ImageNet, applied for transfer learning)
Pooling Layer	GlobalAveragePooling2D (dimensionality reduction prior to dense layers)
Regularization	Dropout (rate 0.3 to prevent overfitting in dense layers)

Table 4. *Comparison to existing systems*

Model	Accuracy (%)	Val Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)	F1 Score (%)
VGG-16	93.88	88.42	71.76	37.38	98.74	49.15
EfficientNet	93.17	85.49	58.19	48.67	97.00	53.01
ConvNeXt	93.49	86.32	65.13	38.01	98.25	48.07
InceptionV3	93.68	87.31	68.51	37.13	98.53	48.16
DenseNet	93.67	87.68	71.70	33.50	98.83	45.56

and testing were performed using an NVIDIA Tesla T4 GPU with 12.72GB of RAM and 68.40 GB of disk space.

3.2. Model Evaluation Metrics.

3.2.1. Confusion Matrix. A confusion matrix is an essential tool for evaluating how well classification models perform. It displays true positives, false positives, true negatives, and false negatives, resulting misclassification pattern detection and model improvement. For classification tasks, this $N \times N$ matrix (where N represents the number of target classes) enables a direct comparison between actual and predicted outputs. Binary classification utilizes a simplified 2×2 matrix, capturing all four significant outcomes.

3.2.2. Key Components of a Confusion Matrix.

- True Positive (TP): Occurs when the model correctly predicts the positive class; both actual and predicted values are positive. For example, when a patient has a disease and the model correctly predicts this condition.
- True Negative (TN): Occurs when the model correctly predicts negative class; both actual and predicted values are negative. For example, when a healthy patient is correctly predicted as having no disease.
- False positive (FP) – Type I error: Occurs when the model incorrectly predicts a positive class when the actual class is negative. For example, when a healthy patient is incorrectly predicted to have a disease.
- False negative (FN) – Type-II error: Occurs when the model incorrectly predicts a negative class when the actual class is positive. For example, when a patient with a disease is incorrectly predicted as healthy.
- Accuracy: Accuracy is the ratio of all correct classifications, both positives and negative. It is calculated as: Accuracy $= \left(\frac{TP+TN}{TP+FP+FN+TN}\right) \times 100\%$.
- Precision: Precision measures how many of the predicted positives are correct. It is calculated as Precision $= \left(\frac{TP}{TP+FP}\right) \times 100\%$.
- Specificity: Specificity, or the true negative rate, quantifies how well a model identifies negative cases. Specificity complements sensitivity and is crucial for evaluating model performance on negative cases. It is calculated as Specificity $= \left(\frac{TN}{TN+FP}\right) \times 100\%$.
- Recall (Sensitivity): The true positive rate (TPR) is the ratio of all actual positives that were accurately identified as positives. It is also called recall. It is calculated as Recall $= \left(\frac{TP}{TP+FN}\right) \times 100\%$.
- F1 Score: The F1 score is a performance metric that brings together precision and recall into one values. It does this by calculating their harmonic mean. It is calculated as F1-Score $= \left(2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}\right) \times 100\%$.

3.2.3. ROC Curve. The receiver operating characteristic (ROC) curve is a graphical tool for evaluating binary classifier performance across different thresholds. The plot illustrates now the true positive rate (TPR) and false positive rate (FPR) vary at different decision thresholds; here, TPR is plotted on the y-axis and FPR on the x-axis.

Table 3 provides the hyperparameters of the proposed model with detailed descriptions.

3.3. Analysis. Table 4 summarizes the quantitative performance evaluation of the five deep-learning models, InceptionV3, DenseNet, EfficientNet, VGG-16, and ConvNeXt, with respect

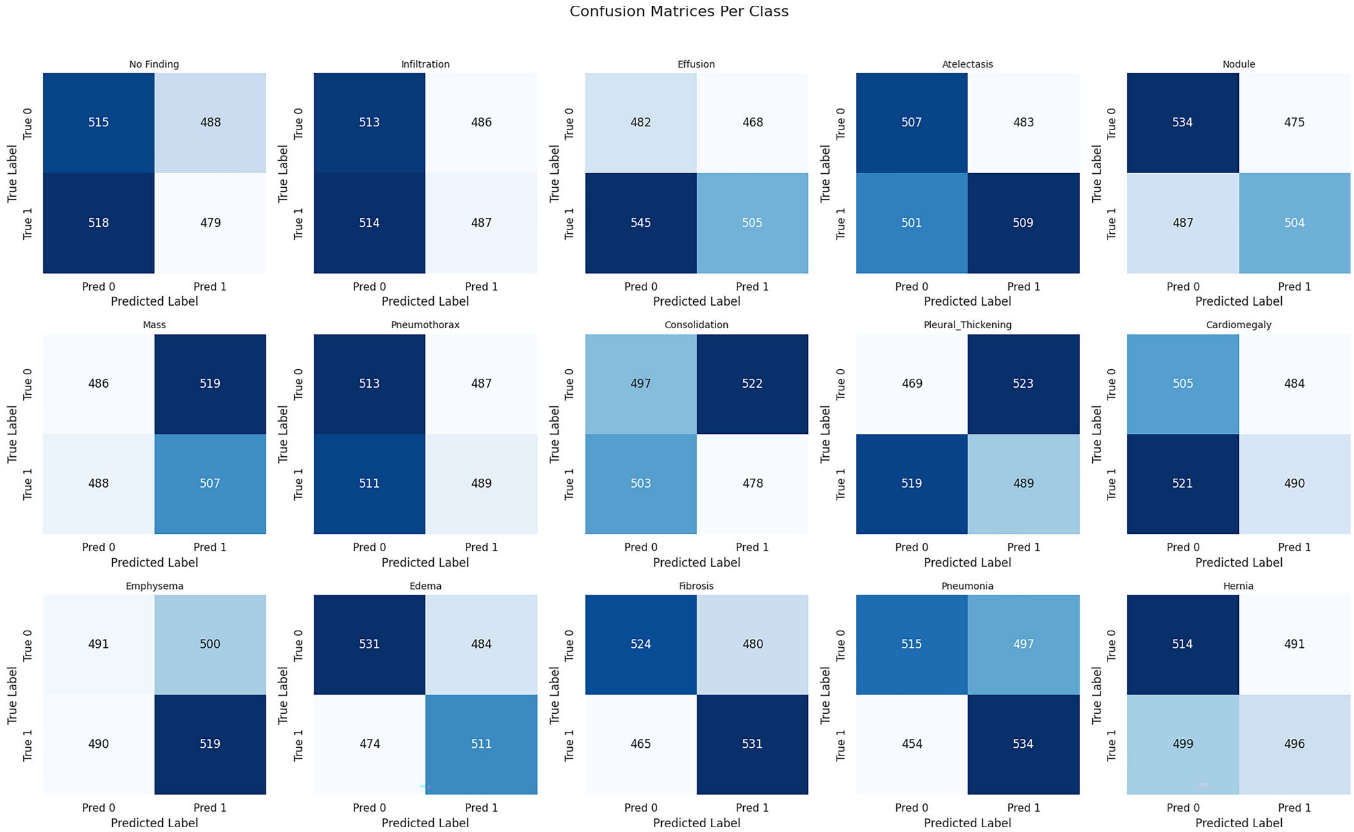


Figure 6. Confusion matrix for VGG-16

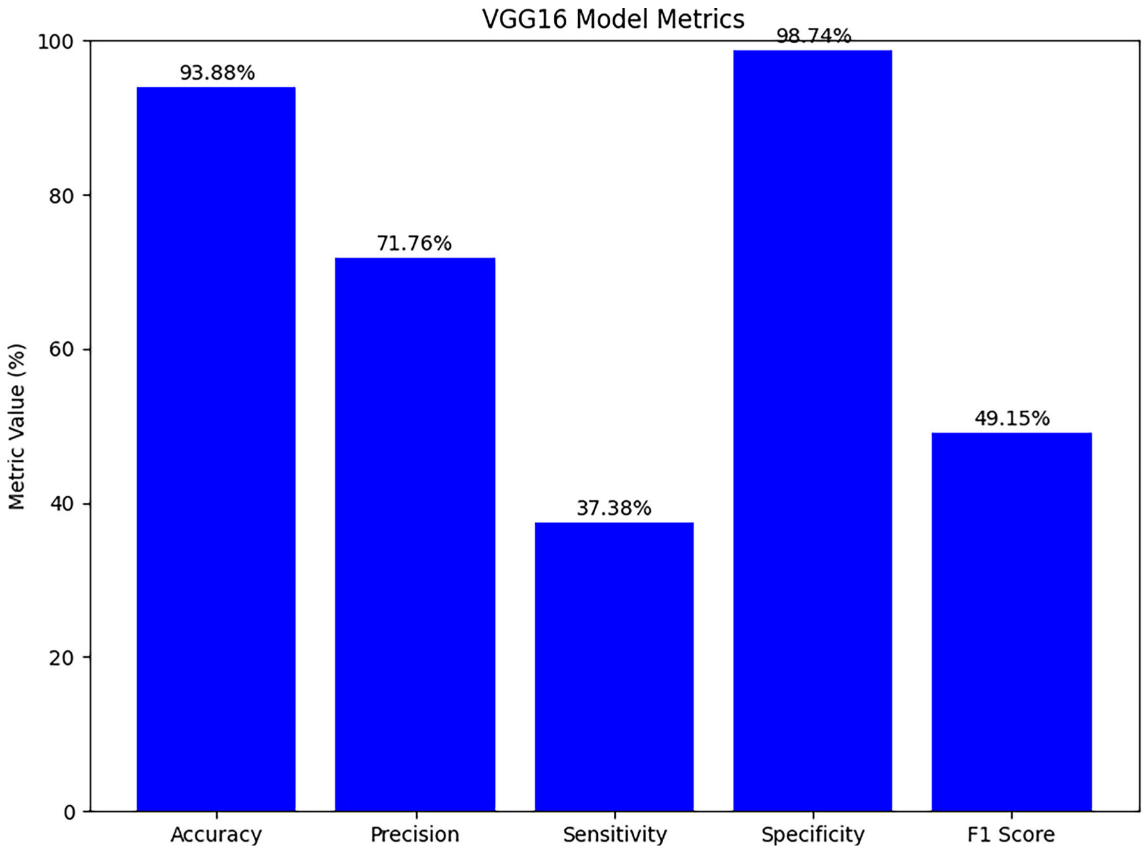


Figure 7. Evaluation metrics for VGG-16

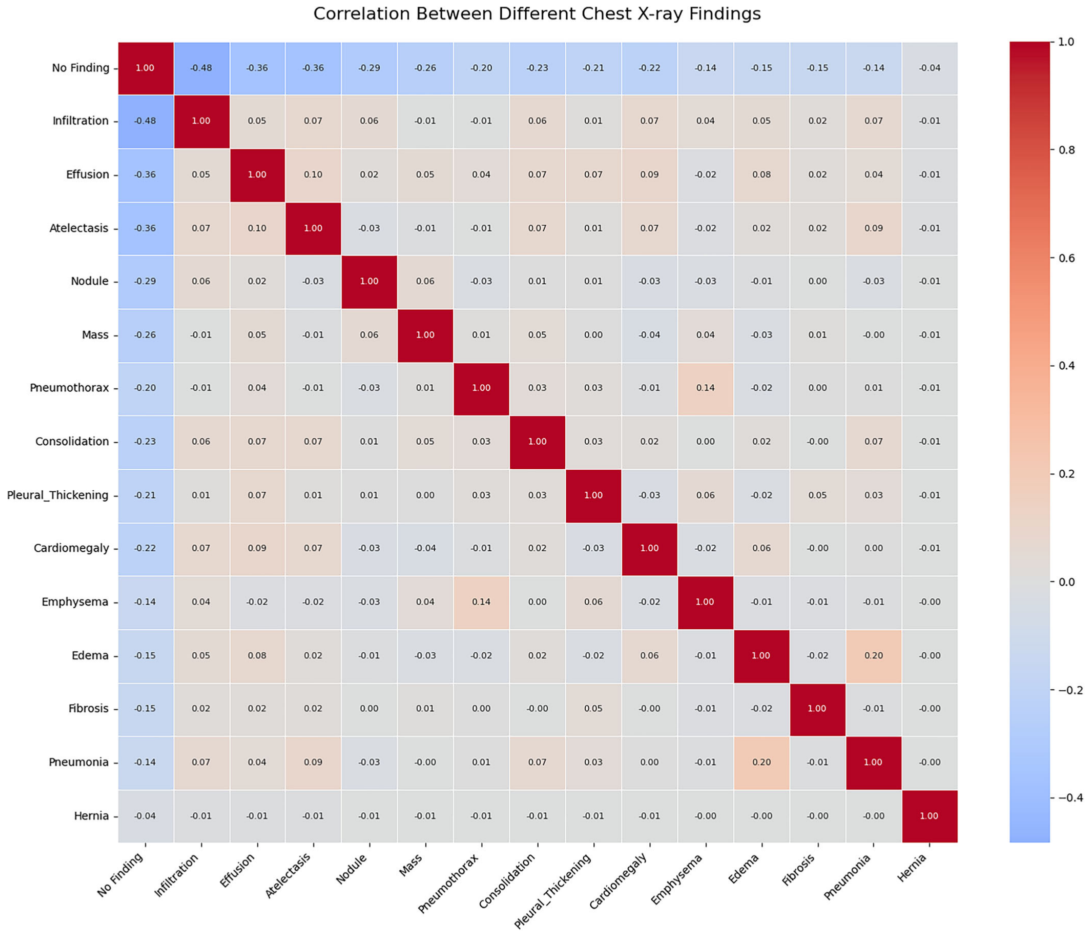


Figure 8. Correlation matrix for VGG-16

to precision, sensitivity, specificity, F1 score, accuracy, and validation accuracy. The table reports the distinct capabilities of each deep learning model: as shown, VGG-16 achieves the highest performance in accuracy (93.88%), precision (71.76%), and specificity (98.74%) but has the lowest sensitivity (37.38%), indicating a tendency to miss positive cases. EfficientNet demonstrates superior performance in sensitivity (48.67%) and F1 score (53.01%) for detecting positive cases but at the expense of lower precision and specificity. DenseNet excels in specificity (98.83%) and precision while showing the lowest sensitivity (33.50%), reflecting its conservative classification approach. ConvNeXt and InceptionV3 provide balanced and reliable results across metrics. The optimal model selection depends on task requirements, VGG-16 for overall accuracy, EfficientNet for recall optimization, and others for specific performance trade-offs.

3.4. Proposed Model VGG-16.

3.4.1. Confusion Matrix. The confusion matrices in Figure 6 provide a comprehensive visualization of the model's performance across 15 thoracic disease classes. Each matrix represents classifications as true positives, false positives, true negatives, or false negatives. The diagonal entries indicate correct classifications, whereas off-diagonal entries indicate misclassifications. The model demonstrates well-balanced performance across classes such as "No Finding," "Infiltration," "Effusion," and "Cardiomegaly." However, classes such as "Pleural Thickening" and "Edema" exhibit higher rates of false positives and false negatives, indicating challenges in recognizing subtle radiographic features. Clinically overlapping conditions such as "Pneumonia" and "Consolidation" show increased confusion because of visual similarities on chest X-rays. These findings highlight model limitations and suggest potential improvements, including class-specific calibration or attention mechanisms. The matrix visualization facilitates comprehensive error analysis, identifying biases, class imbalance effects, and adjustable parameters in medical image classification models.

3.4.2. Evaluation Metrics. The VGG-16 model (Figure 7) yielded a high specificity (98.74%) and accuracy (93.88%) while exhibiting a low sensitivity (37.38%), indicating weak positive case detection. The moderate precision (71.76%) and low F1 score (49.15%) suggest negative class dominance because of class imbalance. Performance improvements could be achieved through class rebalancing or threshold adjustments.

3.4.3. Correlation Matrix. The correlation matrix in Figure 8 predominantly shows weak correlations (< 0.5) between most disease labels, with the strongest correlation observed between "No Finding" and "Infiltration" (0.48). Moderate correlations such as "Edema–Pneumonia" (0.20) and "Effusion–Cardiomegaly" (0.09) indicate infrequent overlap. "Hernia" shows minimal correlation with other conditions,

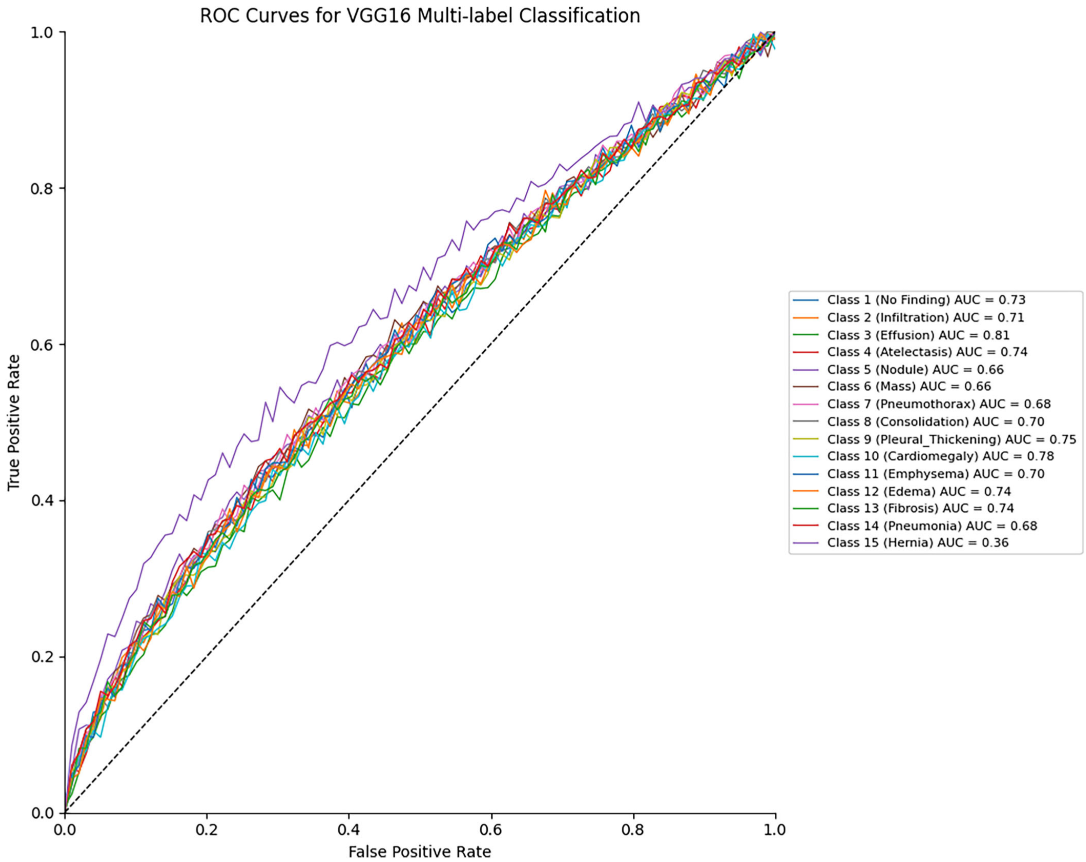
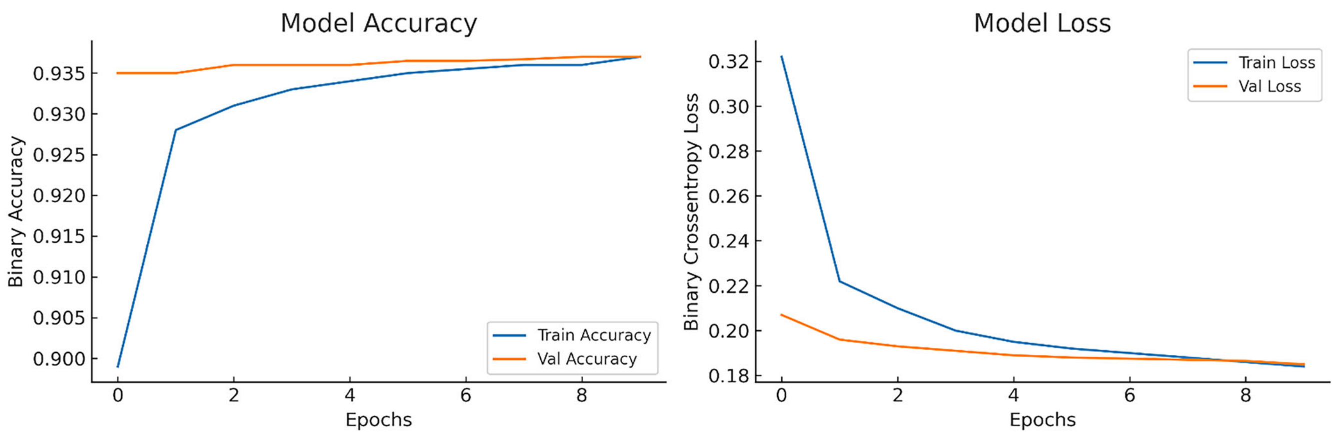


Figure 9. ROC curve for VGG-16

confirming the independence of most labels and supporting the necessity for multilabel classification approaches.

3.4.4. ROC Curve. The ROC curve of the VGG-16 model (Figure 9) shows area under the curve (AUC) values ranging from 0.36 to 0.81 across 15 classes. The model achieved the highest performance for "Effusion" (0.81) and "Cardiomegaly" (0.78) but showed the lowest performance for "Hernia" (0.36). Most classes demonstrate moderate AUC values between 0.73 and 0.74. Poor-performing classes could benefit from data augmentation or class rebalancing strategies.

Figures 10(a) and 10(b) present the training performance of the VGG-16 deep learning model through accuracy and loss curves. These curves illustrate how model accuracy and loss evolve across training epochs. Such performance plots are essential for evaluating model efficiency, revealing potential underfitting or overfitting symptoms, and providing insights into the learning process. The close alignment of training and validation curves indicates effective training and good generalization capability. Monitoring accuracy and loss metrics is crucial for performance assessment and implementing appropriate adjustments to enhance model effectiveness.



(a) Accuracy graph for VGG16

(b) Loss graph for VGG16

Figure 10. Accuracy and loss graphs for the model architecture: a accuracy graph for VGG-16, b loss graph for VGG-16

Table 5. *Comparison of the proposed model with existing works*

Authors	Year	Techniques	Dataset	Accuracy (%)
Magalhães et al.[7]	2024	Multimodal transformer (Swin Transformer + GPT-2), bilingual datasets	NIH Chest X-ray, IU X-ray	ROUGE-L≈ 0.404–0.748, METEOR ≈ 0.393–0.741
Dansana et al.[8]	2023	Tuned VGG-19, Inception V2, decision tree	X-ray and CT scan images (360 images)	Best 91% (VGG-19)
Yang et al.[9]	2023	Learned knowledge base, multimodal alignment	IU-Xray, MIMIC-CXR	79.5% (MIMIC-CXR)
Pang et al.[10]	2023	Hierarchical RNN, attention-based frameworks, reinforcement learning	IU X-ray, MIMIC-CXR, COVID-19, spinal and skin image sets	Up to 0.829
Liu et al.[11]	2021	Contrastive attention (aggregate and differentiate attention)	IU X-ray, MIMIC-CXR	F1 score = 0.303
Our proposed model	2025	CNN-based deep learning model (VGG-16)	NIH Chest X-ray (10,000 images)	93.88%

4. CONCLUSIONS

This paper presents an investigation into the feasibility of developing a deep learning system for the automated classification of chest X-ray illnesses and report generation. Preprocessing techniques of normalization and augmentation were used to enhance image quality for improved model learning. Different CNN models were tested, and the best accuracy of 93.88% for multilabel disease classification of thoracic diseases was obtained using the novel VGG-16-based model. This approach holds immense potential to assist radiologists in reducing diagnostic workload, enhancing turnaround time, and improving consistency, particularly in low-resource, high-demand clinical settings. Furthermore, by supporting digital diagnostics and reducing reliance on printed reports and film-based imaging, this work contributes to a cleaner and more environmentally sustainable healthcare environment. Future research should include the use of transformer-based language models for autonomous report generation, larger and more diverse dataset evaluations, and the integration of advanced interpretability and optimization techniques. In addition, future research will formally assess environmental effects and investigate more environmentally sustainable AI solutions regarding balancing diagnostic performance with computational efficiency.

AFFILIATIONS AND AUTHOR DETAILS

Undergraduate Author

Md. Siam Ahmad – *Department of Computer Science and Engineering, Jamalpur Science and Technology University, Jamalpur 2012, Bangladesh;* 0009-0006-4350-9255
Email: mdsiamahmad1010@gmail.com

Author

Md. Faruk Hossen – *Department of Computer Science and Engineering, Jamalpur Science and Technology University, Jamalpur 2012, Bangladesh;* 0009-0003-9792-0187
Email: farukhossen1401@gmail.com

Corresponding Author

Mohammad Hasan – *Research Mentor, Department of Computer Science and Engineering, Jamalpur Science and Technology University, Jamalpur 2012, Bangladesh;* 0000-0002-1972-3239
Email: hasan.cse@jstu.ac.bd

ACKNOWLEDGEMENTS

We are extremely grateful to all the individuals who assisted us throughout this research

REFERENCES

(1) Kasban, H., El-Bendary, M., Salama, D., et al. (2015). A comparative study of medical imaging techniques. International Journal of Information Science and Intelligent System, 4(2):37–58.

(2) Litjens, G., et al. "A Survey on Deep Learning in Medical *Image* Analysis." *Medical Image Analysis*, vol. **42**, 2017, pp. 60–88.

(3) Rajpurkar, P., et al. "CheXNet: Radiologist-Level Pneumonia Detection on Chest X-rays with Deep Learning." arXiv preprint arXiv:1711.05225, 2017.

(4) Simonyan, K., and Zisserman, A. "Very Deep Convolutional Networks for Large-Scale Image Recognition." *International Conference on Learning Representations (ICLR)*, 2015.

(5) Wang, X., et al. "ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases." *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2097–2106.

(6) Health Care Without Harm. "Environmental Impact of Medical Imaging." *Health Care Without Harm Reports*, 2021.

(7) G. V. Magalhães, R. L. d. S. Santos, L. H. Vogado, A. C. de Paiva, and P. d. A. dos Santos Neto, "Xrayswingen: Automatic medical reporting for x-ray exams with multimodal model," *Heliyon*, vol. **10**, no. 7, 2024.

(8) D. Dansana, R. Kumar, A. Bhattacharjee, D. J. Hemanth, D. Gupta, A. Khanna, and O. Castillo, "Early diagnosis of covid-19-affected patients based on x-ray and computed tomography images using deep learning algorithm," *Soft* computing, pp. 1–9, 2023.

(9) Yang, S., Wu, X., Ge, S., Zheng, Z., Zhou, S. K., and Xiao, L. (2023). Radiology report generation with a learned knowledge base and multi-modal alignment. Medical *Image* Analysis, **86**:102798.

(10) Pang, T., Li, P., and Zhao, L. (2023). A survey on automatic generation of medical imaging reports based on deep learning. BioMedical Engineering OnLine, **22**(1):48.

(11) F. Liu, C. Yin, X. Wu, S. Ge, Y. Zou, P. Zhang, and X. Sun, "Contrastive attention for automatic chest x-ray report generation," arXiv preprint arXiv:2106.06965, 2021.

(12) K. El Asnaoui and Y. Chawki, "Using x-ray images and deep learning for automated detection of coronavirus disease," Journal of Biomolecular Structure and Dynamics, vol. **39**, no. 10, pp. 3615–3626, 2021.

(13) A. Zargari Khuzani, M. Heidari, and S. A. Shariati, "Covid-classifier: An automated machine learning model to assist in the diagnosis of covid-19 infection in chest x-ray images," Scientific Reports, vol. **11**, no. 1, p. 9887, 2021.

(14) P. Gifani, A. Shalbaf, and M. Vafaeezadeh, "Automated detection of covid-19 using ensemble of transfer learning with deep convolutional neural network based on ct scans," International journal of computer assisted radiology and surgery, vol. **16**, pp. 115–123, 2021.

(15) Y. Song, S. Zheng, L. Li, X. Zhang, X. Zhang, Z. Huang, J. Chen, R. Wang, H. Zhao, Y. Chong et al., "Deep learning enables accurate diagnosis of novel coronavirus (covid-19) with ct images," *IEEE/ACM* transactions on computational biology and bioinformatics, vol. **18**, no. 6, pp. 2775–2780, 2021.

(16) https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.istockphoto.com%2Fphotos%2Fchestxray&psig=AOvVaw1IqfnROiYwraAxjTTzWaVg&ust=1754550925236000&source=images&cd=vfe&opi=89978449&ved=0CBIQjRxqFwoTCPiog9fR9Y4DFQAAAAAdAAAAABAE

(17) https://nohat.cc/f/logo-product-line-font-angle-databases-outline/5176033321943040-201902131427.html

(18) https://www.istockphoto.com/vector/csv-file-icon-gm1036168986-277364197

(19) https://letsdatascience.com/handling-missing-values/

(20) https://medium.com/@akashmishra77/box-plots-detect-and-remove-outliers-from-distribution-a124ee88cf3e

(21) https://www.freepik.com/icon/scalability_3024206

(22) https://www.gabormelli.com/RKB/VGG_Convolutional_Neural_Network

(23) https://www.g2.com/articles/pooling-layers

(24) https://datascience.stackexchange.com/questions/57005/why-there-is-no-exact-picture-of-softmax-activation-function

(25) https://ravinasingla94.wordpress.com/2016/03/26/neural-network/

In response to the research questions outlined in the introduction, the following answers are provided:

RQ1: How effective is the proposed AI-based chest X-ray interpretation model developed using the NIH Chest X-ray dataset in terms of diagnostic accuracy and consistency?

Answer: Our proposed VGG-16-based model performed exceptionally well, achieving 93.88% accuracy on chest X-ray classification. While its precision (71.76%) and specificity (98.74%) were particularly strong—indicating reliability in detecting and ruling out diseases—its sensitivity (37.38%) was comparatively lower, meaning it sometimes missed subtle cases. Despite this, the overall performance shows that the model is not only consistent but quite dependable for supporting clinical diagnosis.

RQ2: How does the system ensure reliable diagnostic reporting of chest pathologies?

Answer: We ensured reliability by using a robust deep learning pipeline: preprocessing the dataset thoroughly (including normalization, augmentation, and outlier handling), applying multilabel classification, and evaluating the model with a range of performance metrics (e.g., precision, F1 score, AUC). These measures helped us detect where the model excels and where improvements are needed, especially in handling rarer conditions.

RQ3: What are the primary advantages of using AI-based reporting for chest X-rays compared to conventional manual radiology reporting?

Answer: Traditional radiology reporting can be time-consuming and subject to interobserver variation. In contrast, our AI system provides fast, consistent, and reproducible reports—free from fatigue or bias. This not only supports radiologists but also reduces delays in diagnosis, especially where radiologists are scarce.

RQ4: How can this AI-driven reporting system enhance healthcare delivery, particularly in settings with limited radiology infrastructure?

Answer: In rural or under-resourced settings where trained radiologists may not be available, our model can act as a frontline diagnostic tool, offering initial interpretations that can assist general practitioners or community health workers. It helps bridge the gap in healthcare access, ensuring patients in remote areas receive quicker evaluations and referrals.

RQ5: How do the quality and reliability of AI-generated reports compare to those produced by expert radiologists?

Answer: While AI is not a replacement for human expertise, our results show that AI-generated reports using VGG-16 are comparable in consistency and accuracy for many common chest conditions. However, for complex or overlapping cases like "Pneumonia vs. Consolidation", radiologist oversight is still essential. The AI acts as a reliable assistant, not a final authority.

RQ6: What are the quantifiable environmental benefits achievable through transitioning from conventional paper-based radiology practices to AI-based digital reporting systems?

Answer: Although we didn't measure the exact reduction in carbon footprint, transitioning to digital, AI-powered diagnostics naturally eliminates the need for printed films, paper records, and associated storage. This contributes to greener medical practices by reducing waste, energy use, and environmental impact—aligning healthcare with sustainability goals.

RQ7: What ethical, legal, and clinical challenges must be addressed for the successful integration of AI-based diagnostic systems into routine clinical practice?

Answer: Like any medical technology, ethical issues around data privacy, bias, accountability, and trust must be carefully handled. Legally, systems like ours must comply with medical device regulations, and clinically, they must be validated across diverse populations. We acknowledge that real-world deployment will require collaboration with healthcare professionals, policymakers, and ethicists to ensure safe, fair, and effective use.