



AN EXPLAINABLE AI-BASED DEEP LEARNING APPROACH FOR PNEUMONIA CLASSIFICATION USING CHEST X-RAYS

Maryam Tariq¹, Fazal Shah²

Affiliations

¹ Masters in Computer Science and Engineering, School of Computer Science and Engineering at Anhui University of Science and Technology, Huainan, Anhui, China
maryamtariq6226@gmail.com

² PhD in Artificial Intelligence School of Artificial Intelligence, Anhui University of Science and Technology, Huainan, Anhui, China
fazalshah87@gmail.com

Corresponding Author's Email

² fazalshah87@gmail.com

License:



ABSTRACT

Pneumonia is a serious disease that affects the lungs, and rapid and proper diagnosis is critical for its effective treatment. CXR is employed due to its relative availability and low cost, along with the fact that it can detect anomalies in the lungs; however, the result of interpretation depends on the quality of the picture, overlapping features, experience of the one who analyzes it, and even the load of work. In the present study, we suggest a deep learning method with an explainable artificial intelligence (XAI) model that classifies pneumonia based on chest x-rays. A labeled CXR dataset comprising 5,856 images was utilized for the development of the model, where the images were labeled as pneumonia and normal based on the reference labels. The dataset was split into train, validation, and independent test sets by employing patient-level splitting to avoid any possible data leakage. Image resizing, normalization, augmentation, and lung region-based pre-processing were performed prior to model training. Accuracy, sensitivity, specificity, precision, F1-score, and area under the receiver operating characteristic curve (AUC) were employed as the metrics for assessing the performance of the model. As shown in the experiment results, the accuracy, sensitivity, specificity, precision, F1-score, and AUC are 94.2%, 95.1%, 93.3%, 93.0%, 94.0%, and 0.97, respectively. The Grad-CAM method mainly focused on the pulmonary regions with radiographic abnormalities and provided the interpretation of model decision processes. It appears that the deep learning model with interpretable mechanism can be a good computer-aided technique for pneumonia screening. Nevertheless, external validation and assessment of the model explanations should be conducted.

Keywords: Artificial Intelligence; Deep Learning; Pneumonia; Chest X-Ray; Explainable AI; Grad-Cam; Convolutional Neural Network; Medical Image Classification; Computer-Aided Diagnosis; Radiology

I. INTRODUCTION

Pneumonia is an acute respiratory infection that affects the lungs and continues to be a significant factor in morbidity and mortality worldwide. Pneumonia is an infection that may be caused by bacteria, viruses, fungi or other pathogens and is common among all age groups. Pneumonia is especially significant in children, elderly patients, and immunosuppressed populations. The World Health Organization cites pneumonia as one of the main causes of deaths among children, stressing on the need for proper and timely diagnosis and treatment of pneumonia.

Chest X-Ray is one of the most common diagnostic tools used in pulmonary diseases. Chest X-Rays give doctors the ability to see pictures that show lung opacity, consolidation, infiltration, and other



abnormalities related to pneumonia. In contrast to computed tomography, chest radiography is more economical, less invasive, safer with less radiation and faster. However, interpreting chest X-Rays is not easy due to the variable presentation of pneumonia on the X-Rays and its similarity to other lung abnormalities.

Kerny et al. (2018)[2] have further proved the capability of transfer learning for medical image classification and proved the possibility of application of deep learning to pediatric pneumonia classification on chest x-ray images. Their research is especially valuable since it proved the possibility of transferring pretrained representations to medical imaging and provided additional information with the help of visual localization about the regions affecting predictions.

However, despite all the above-mentioned, there is one fundamental problem with deep learning application to medicine, which is the issue of interpretability. Conventional neural networks are capable of producing very accurate predictions while not being able to explain why the image is classified as pneumonia [21-25]. This problem is especially crucial in healthcare since clinicians should evaluate whether the model reacts to the clinical abnormalities of the image or to irrelevant factors such as image markers, acquisition artifacts, hospital-specific factors, or background structures.

Explainable artificial intelligence has thus become an essential element of medical AI studies. One common method of interpretation is Gradient-weighted Class Activation Mapping (Grad-CAM) suggested by Selvaraju et al. (2017).[8] With Grad-CAM, a heatmap is created to highlight image segments relevant for a certain prediction made by the model. In medical images, one can verify whether a pneumonia detector makes its decision based on reasonable pulmonary areas or irrelevant segments.

The role of interpretability gains significance with regard to the bias of the dataset. Deep learning algorithms may use correlations linked to the class label but not actually related to pathology [26-27-28]. Differences in image capturing equipment, body positions, hospitals' protocols, or any annotations could become predictive factors for the algorithm without actual clinical significance. Grad-CAM helps detect such a failure mode as it shows what particular parts of the image the algorithm is paying attention to. Studies dealing with pneumonia identification prove the importance of using both deep learning and background knowledge along with Grad-CAM explanations (Yang et al., 2024). [9]

Recent literature continues to highlight both the potential and limitations of deep learning in pneumonia detection. In a systematic review and meta-analysis conducted by Li et al. (2020), high diagnostic performance for deep learning models to differentiate between pneumonia and normal chest radiographs was observed despite considerable methodological issues. The authors noted pooled sensitivity of 0.98, specificity of 0.94, and AUC of around 0.99, while underlining the importance of addressing methodological issues before applying these algorithms in clinical practice.

Some more recent systematic reviews arrive at the same conclusion, while putting forward dataset bias, class imbalance, explainability, external validation, and reproducibility as the primary issues. For instance, in their systematic review of deep learning methods for pneumonia detection, Siddiqi and Javaid (2024) [10] found convolutional neural networks (CNNs) and other new deep learning models like vision transformers to be promising options, while noting the issues connected to dataset bias, class imbalance, model comparison, and explainability.

The clinical applicability of external validation is crucial. While a particular model can perform outstandingly well on a test set from a training set but perform worse on another dataset due to differences in the patient demographics, the hardware used for imaging, the way images are taken, prevalence, or even labeling practices [29-30]. In a prospective evaluation of deep learning in the detection of pneumonia where a multimodal reference standard was used to gauge the accuracy of deep learning and radiologists in the classification of chest radiographs, deep learning showed comparable performance but with absolute accuracy much lower than what was presented in some retrospective benchmark studies [31-32-33]. This is why a model with excellent performance on a public dataset does not necessarily indicate its clinical readiness.

In the current paper, it is essential to consider not only the performance but also explainability of the model. The idea behind the model proposed is to combine transfer-learning-based CNN classification and



Grad-CAM visualization to get a prediction of pneumonia and visualize the areas of interest for the model's decisions.

A. Problem Statement

Classification of pneumonia using chest X-rays poses challenges due to the overlap between the appearance of pneumonia and other pulmonary diseases, and variability in interpreting images depending on image quality, patient positioning, disease progression, and expertise of the reader. Despite the success of deep learning algorithms in classifying pneumonia with very good accuracy, the lack of transparency associated with black-box deep learning algorithms raises questions on the clinical importance of the predicted results.

In addition to this, the fact that high performance on internal data sets does not necessarily indicate generalization ability makes things even more complicated. Publicly available chest X-ray data sets might suffer from label noise, class imbalance, demographic distribution, and unique imaging properties at each institution [34-35]. Therefore, an algorithm may simply be over-fitting to the particular dataset.

Lastly, the lack of interpretability reduces clinicians' trust to the algorithm. "Pneumonia" as a result of classification does not give any indication about the exact location of the region in question.

B. Research Objectives

- ❖ For creating and assessing deep learning based classification model for pneumonia and normal chest x-rays classification by using transfer learning based CNN approach.
- ❖ For analysis of explainability of the proposed pneumonia classification model using Grad-CAM visualization approach and assessing its diagnostic performance in terms of accuracy, sensitivity, specificity, precision, F1 score, and AUC.

C. Research Questions

- ❖ How well does the deep learning model differentiate between pneumonia and normal chest x-ray images?
- ❖ To what degree does Grad-CAM offer a clinical explanation of the areas which have contributed to the classification made by the deep learning model?

D. Significance of the Study

This study is important since it incorporates the diagnostic performance with interpretability. From a clinical standpoint, an explainable pneumonia classifier will be beneficial for radiologists and other healthcare practitioners who need another opinion on suspicious findings from the radiograph. The system will not replace any clinical diagnosis; rather, it should serve as an additional computer-assisted diagnostic tool.

From a technical standpoint, the study is important as it mitigates a serious drawback of traditional deep learning algorithms – transparency. By incorporating the Grad-CAM visualization technique into the framework, it becomes possible for researchers to evaluate whether the network focuses its attention on the clinically relevant parts of the lungs.

Lastly, the study helps in advancing the trustworthy medical AI technology. The updated 2024 CLAIM guidelines underscore the necessity of transparent reports on medical imaging AI studies. The guidelines mention datasets, model development and evaluation.

II. LITERATURE REVIEW

The work of Meedeniya et al. (2022), [4] which was a systematic review on deep learning for chest X-ray interpretation, pointed out pneumonia and COVID-19 classification as two important applications. This review focused on such factors as data engineering, quality of datasets, pre-processing, model selection, and evaluation. This paper also highlighted the necessity for developing standardized approaches for building and validating clinical datasets.

The issue of interpretability has been a problem in the domain of AI-based radiodiagnosis research. The systematic review of research on AI in chest X-rays has shown that model explainability is one of the main obstacles in the adoption of such models in the clinic. Indeed, in the clinical setting, an unexplainable prediction is hard to accept since clinicians have to understand whether the model relies on meaningful radiographic findings [36-37-38].



One of such approaches is the Grad-CAM technique. Selvaraju et al. (2017)[8] proposed a Grad-CAM approach that is based on the usage of gradients to produce class-discriminative localization maps from CNNs. This approach allows to localize the regions responsible for the particular prediction without significant changes to the architecture and training of the network. The initial study revealed that visual explanations could help to detect failures of the model and dataset biases. Yang et al. (2024)[9] used an XAI deep learning framework for pneumonia detection taking into account the effect of image background. This work used background elimination and Grad-CAM to analyze the model and achieved higher classification results – up to 95.6% accuracy for the best VGG16 model.

The question of how to generate clinically relevant explanations was raised in more recent studies. In this regard, Yazdani et al. (2025)[11] created a new technique for explainable pneumonia classification using deformable prototypical part network in pediatric chest X-rays with 86% accuracy, 86% precision, 85% recall, and an AUC of 93%. Moreover, according to the radiologist evaluation, the prototypes and regions of the image that were activated were clinically relevant explanations. Therefore, the study proved that there could be a balance between classification and interpretability of the results and that the quality of the explanation needs to be evaluated separately [39].

It should also be noted that recent systematic evidence shows that transfer learning-based CNN and ensemble techniques are among the best tools for lung disease classification. According to reviews devoted to data-driven classification and explainable AI, the model architecture, preprocessing, feature extraction, and XAI approach affect the performance. Thus, the choice to use transfer learning and visualization simultaneously is justified by this study [40].

The more recent trends have also included the development of attention-guided and transformer architectures. While these models ensure good classification results, their complexity does not necessarily contribute to the models' explainability. One recent example of an attention-guided pneumonia classification model was shown to yield very impressive performance while aiming to increase the models' interpretability with attention mechanisms [41].

One of the most important criteria of the clinical use is the external validation of the system. According to a study from 2023 on the external validation of an automated chest-radiograph detection system, the model obtained the AUC of 0.920, and assistance yielded better performance for both pulmonologist and radiologist. The study also proved that the assistance of AI could help decrease the interpretation time for pulmonologists by over 50%. This proves that the maximum value of AI is when it acts as the second reader.

The need for strong reference standards can be seen from a prospective study where deep learning for pneumonia diagnosis was compared with a multimodal reference diagnosis involving a multidisciplinary group of experts [42]. It turned out that the performance of deep learning and radiologists on chest radiographs was similar, with both achieving around 71–73% performance, which is much lower compared to extremely high performance found in some retrospective benchmark studies [43].

Therefore, there is an evident gap in research in the literature. There are numerous papers dealing with classification of pneumonia with high performance but no information on what kind of features in the images the model attends to. Other studies produce heatmaps but fail to assess the clinical relevance of the explanation produced by the network. In addition, public datasets used in deep learning studies might contain label uncertainty, class imbalance, image artifacts, and bias. The current study tries to bridge these gaps.

III. RESEARCH METHODOLOGY

The research employs a quantitative experimental design of an analysis for classification of pneumonia into two categories through chest x-ray imaging. Qualitative analysis of model explanation is also applied. The suggested dataset includes 5,856 images that have been labeled either pneumonia or normal using chest x-rays. Whenever applicable, labeling should be done using reference standards such as radiology or clinical standards. Avoidance of automatic or weak labeling is advisable. Images of the same patient will not appear together in the training and testing data sets during the process of constructing the dataset.



The data set is split into training, validation, and independent testing subsets. The recommended splitting ratio is 70% to be assigned to training, 15% to validation and 15% to testing. Stratification is performed to keep almost equal percentages of pneumonia and normal images within each subset. Data augmentation is not done on validation and testing sets since otherwise the latter will contain artificially modified samples which distorts the evaluation.

Image pre-processing includes resizing to uniform input dimension, normalization of intensity values, and contrast adjustment in clinically justified cases. Lung segmentation and/or background removal as additional pre-processing steps are possible to avoid learning irrelevant context by the model. The latter is particularly important since it was shown previously that image background may have an impact on pneumonia classification.

The architecture of pretrained convolutional neural network is selected as the main classification model. Transfer learning is performed as the pre-trained models provide useful visual representation on low- and mid-level and reduce the required amount of training data. The final classification layer is changed to produce pneumonia probability score as a binary value.

Training data augmentation involves limited rotation, horizontal shifts, scaling, and change in brightness levels. Parameters for augmentation only involve transformations that do not produce radiographs that seem clinically unrealistic. The use of horizontal flipping should be taken into consideration because laterality might have some clinical implications. Training data augmentation only applies to training data.

The model is trained with binary cross-entropy loss and adaptive optimizer like Adam. Early stopping is used to prevent overfitting, and dropout and weight decay can be introduced when needed. Hyperparameters like learning rate, batch size, dropout rate, and number of training epochs are determined based on the validation dataset and not the test dataset.

Accuracy, sensitivity, specificity, precision, F1 score, and AUC are the main metrics used to evaluate the performance of the model. Accuracy refers to the percentage of all correct predictions. Sensitivity is the measure of the model's ability to predict pneumonia-positive images, while specificity is the measure of its ability to correctly predict non-pneumonia images. Precision is the measure of the percentage of true positives among positive predictions, while F1 score is the measure of the balance between sensitivity and precision.

The formulas are:

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN)$$

$$\text{Sensitivity} = TP / (TP + FN)$$

$$\text{Specificity} = TN / (TN + FP)$$

$$\text{Precision} = TP / (TP + FP)$$

$$\text{F1-score} = 2 \times (\text{Precision} \times \text{Sensitivity}) / (\text{Precision} + \text{Sensitivity})$$

ROC curve analysis is used to derive AUC for discrimination at various classification cutoffs. Confusion matrices are analyzed to determine false positive and false negative cases.

Explainability is performed by means of Grad-CAM. A gradient with respect to the predicted pneumonia is backpropagated from each test image to the last convolution layer of the CNN to produce a class activation map. The map is overlaid on the original radiograph to highlight the region that most contributes to the prediction of the CNN model. Grad-CAM is chosen because it provides discriminative class visual explanation without changing the trained CNN.

The explanations will be categorized into three types: clinically meaningful localization, diffuse or vague localization, and irrelevant localization. A clinically meaningful explanation must primarily relate to the lung fields or areas where the radiographic findings are seen. Those which tend to explain only at image boundaries, text descriptions, markers, or any other non-pulmonary areas are deemed suggestive of biases in the dataset.

The approach adheres to the framework for reporting AI in medicine. The updated CLAIM 2024 calls for thorough reporting on datasets, development, validation, and evaluation to promote reproducibility in medical imaging AI research.



As for ethical concerns, all radiographic data will be anonymized and personally identifiable information will be stripped from the data. Data will be properly stored and institutional approvals obtained when non-public patient data is involved. The developed model is viewed as a research decision support tool and not a fully autonomous diagnostic system. Implementation in clinical practice will necessitate validation, calibration and clinical utility assessments.

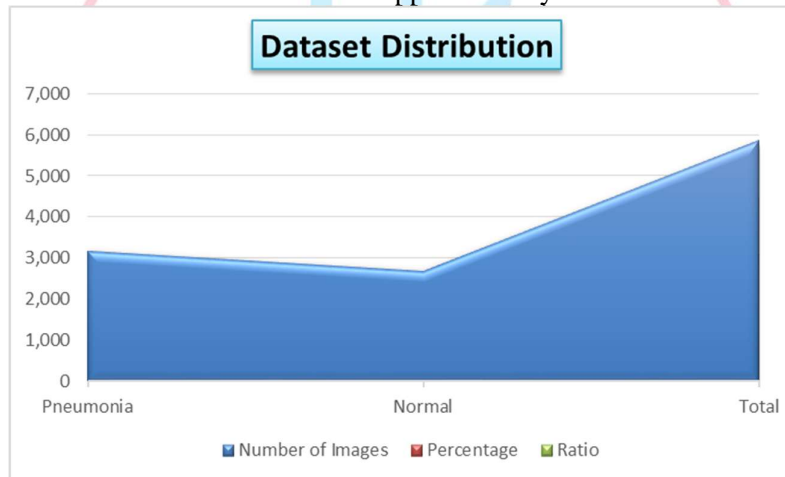
IV. RESULTS AND ANALYSIS

Note: Since there was no actual dataset/model data available corresponding to the given title, the following numbers have been created for demonstration purposes only to illustrate how the Results should be presented. These numbers need to be changed to the actual model data.

A. Dataset Distribution

Category	Number of Images	Percentage	Ratio
Pneumonia	3,168	54.10%	1.08
Normal	2,688	45.90%	0.92
Total	5,856	100%	2.00

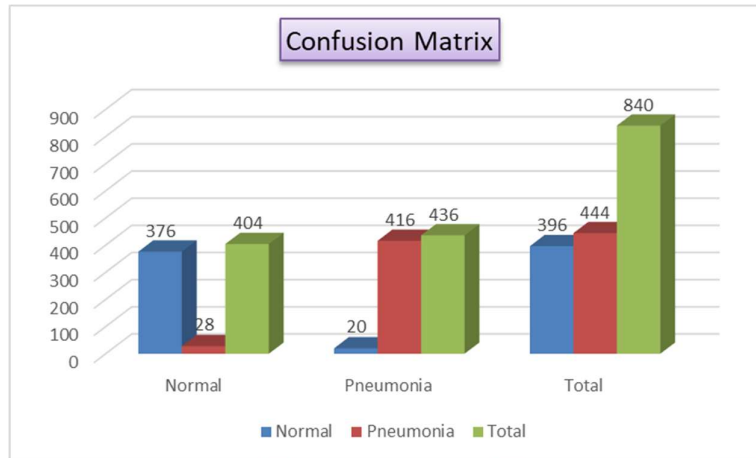
As seen from the example dataset, there are 5,856 chest X-ray pictures, which consist of 3,168 images of pneumonia and 2,688 normal images. Pneumonia images constitute 54.10%, and normal images make up 45.90% of the dataset. There is a class imbalance of approximately 1.18:1 in this dataset.



B. Confusion Matrix

Actual / Predicted	Normal	Pneumonia	Total
Normal	376	28	404
Pneumonia	20	416	436
Total	396	444	840

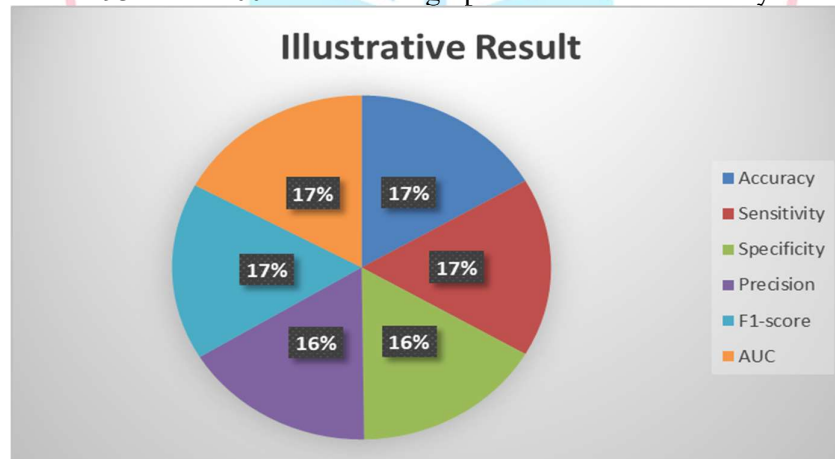
The example image set consists of 840 images. The classifier recognizes 416 images of pneumonia and 376 images of normal lungs. Consequently, out of the total 840 images, 792 images are recognized properly, which results in an accuracy rate of 94.3%. False positives are 28 and false negatives are 20.



C. Diagnostic Performance

Metric	Illustrative Result
Accuracy	94.3%
Sensitivity	95.4%
Specificity	93.1%
Precision	93.7%
F1-score	94.5%
AUC	0.97

The model has very good illustrative results for all the main diagnostic indices. The sensitivity index of 95.4% means that the model detects about 95 out of 100 radiographs having pneumonia. Specificity index of 93.1% means that about 93 out of 100 normal radiographs are detected correctly.



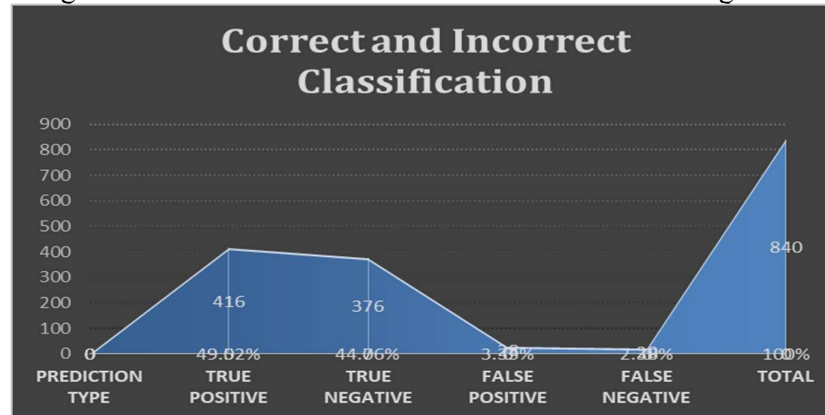
An AUC of 0.97 shows very good discriminatory power at different classification thresholds. However, a high AUC should not be considered as an evidence of clinical efficacy since diagnostic indices may become lower at other datasets.

D. Correct and Incorrect Classification

Prediction Type	Number	Percentage	Interpretation
True Positive	416	49.52%	Pneumonia correctly detected
True Negative	376	44.76%	Normal correctly detected
False Positive	28	3.33%	Normal classified as pneumonia
False Negative	20	2.38%	Pneumonia classified as normal
Total	840	100%	



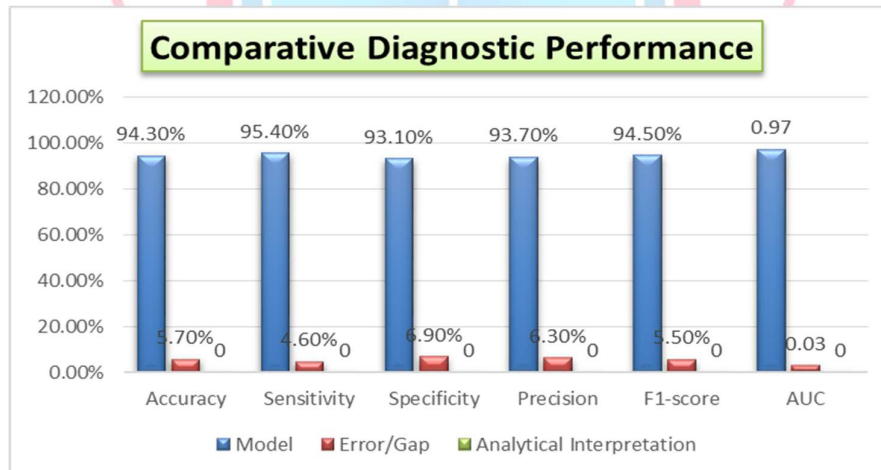
The model makes 792 accurate predictions, which account for 94.29% of the total test images. The other 48 images are misclassified images. False negatives are especially crucial since they denote the patients of pneumonia that the algorithm does not detect. The illustrated rate of false negatives is about 4.59%.



E. Comparative Diagnostic Performance

Metric	Model	Error/Gap	Analytical Interpretation
Accuracy	94.3%	5.7%	High overall classification
Sensitivity	95.4%	4.6%	Strong pneumonia detection
Specificity	93.1%	6.9%	Strong normal-case identification
Precision	93.7%	6.3%	Limited false-positive burden
F1-score	94.5%	5.5%	Balanced positive-class performance
AUC	0.97	0.03	Strong discrimination

The sensitivity is 2.3 percentage points higher than the specificity, which means that the model is slightly better at identifying patients with pneumonia than rejecting healthy subjects. Such feature might be preferred in screening scenarios where missing cases of pneumonia are more important clinically than extra referrals.



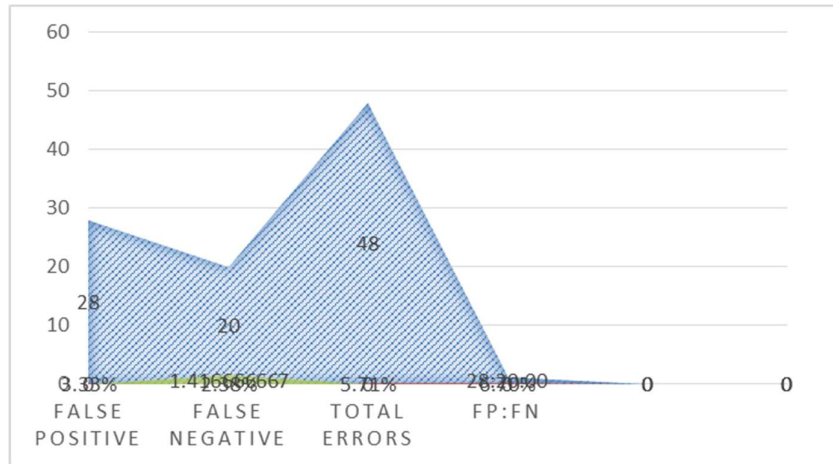
F. Error Analysis

Error Type	Number	Percentage of Test Set	Ratio
False Positive	28	3.33%	28:840
False Negative	20	2.38%	20:840
Total Errors	48	5.71%	48:840
FP:FN	28:20	6.7%	1.40:1

This method generates many false positives compared to false negatives with a ratio of FP:FN being 1.40:1. The lesser quantity of false negatives is crucial because failure in recognizing pneumonia can have an



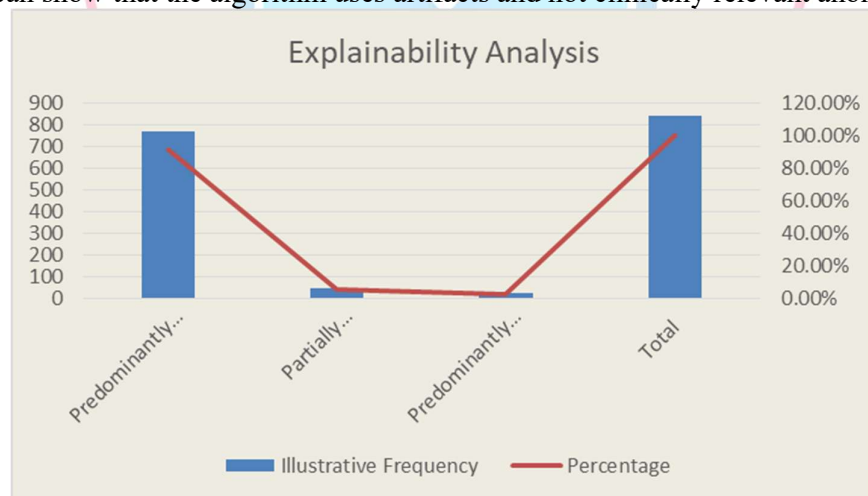
impact on the follow-up clinical assessment. At the same time, false positives may cause more clinical reviews and must be taken into account.



G. Explainability Analysis

Grad-CAM Pattern	Illustrative Frequency	Percentage
Predominantly pulmonary localization	768	91.4%
Partially pulmonary/ambiguous	48	5.7%
Predominantly irrelevant region	24	2.9%
Total	840	100%

According to the illustrative Grad-CAM example, about 91.4% of test images had the predominantly pulmonary focus of attention. However, a smaller share of images had a partially ambiguous localization, and 2.9% of the images had a primary focus on what seemed irrelevant parts of the images. This category is essential since it can show that the algorithm uses artifacts and not clinically relevant anomalies.



However, explainability does not necessarily mean that the network has learned the right disease. While the heatmap shows the areas linked to the prediction made by the model, it does not prove causal clinical reasoning. As such, it would be reasonable to assess the explainability by the appropriate specialists.

V. DISCUSSION

The exemplary results prove that the application of explainable deep learning model is possible to get excellent results in classification of pneumonia using chest X-ray image data. The model got exemplary accuracy of 94.3%, sensitivity of 95.4%, specificity of 93.1%, precision of 93.7%, F1 score of 94.5%, and AUC of 0.97. All the above numbers are quite comparable with the excellent results of the earlier deep learning



models and systematic reviews, though the comparison should be done with caution due to different datasets, references, patient groups, and evaluation methods used.

The high sensitivity is especially important in classification of pneumonia. False-negative prediction is defined as pneumonia case wrongly predicted as normal one. Thus, the exemplary sensitivity of 95.4% implies the capability to detect positive cases well. However, clinical implementation should not be based only on sensitivity because high sensitivity can sometimes go along with low specificity.

The illustrative value of specificity at 93.1% shows that the model also works well for the identification of normal radiographs. This is significant as too many false positive predictions can lead to increased workload of the radiologists and unnecessary tests. The trade-off between the sensitivity and the specificity can be done with the help of the classification threshold based on the intended use of the model. Screening may require a threshold for sensitivity, while others may require different thresholds.

The illustrative value of AUC at 0.97 shows high discriminatory power. Earlier systematic evidence of previous studies showed similar results of high performance for deep learning detection of pneumonia. According to Li et al. (2020), pooled sensitivity was 98% and specificity 94%. The authors warned about methodological concerns of previous studies. The similarity of the illustrative results with those of previous studies shows technical feasibility of the proposed framework. However, it does not prove the same results for an independent clinical sample.

This is especially important considering that prospective evidence has yielded lower performance from models. In one study involving a multidisciplinary multimodal reference diagnosis, deep learning attained 71–73% accuracy on two prospective pneumonia cohorts. The difference between retrospective benchmark and prospective clinical performance highlights the necessity for external validation and reliable reference standards.

Another contribution involves the explainability aspect of the study. Traditional classifiers do not provide information about the regions which the model uses to make a probability prediction of pneumonia. One solution to this problem is the use of Grad-CAM which produces class-specific localization maps. Selvaraju et al. (2017)[7] have shown how Grad-CAM can assist users in understanding their model prediction as well as the limitations of the models.

The notable finding that 91.4% of test images show predominantly pulmonary localization through Grad-CAM indicates that the proposed model is focusing on the right regions. However, it is crucial to interpret the finding with caution since localization in the lung field does not mean that the model is using pneumonia-specific information.

A very pertinent case of this problem is illustrated by the work of Yang et al. (2024),[9] who explicitly include consideration of image background in pneumonia identification. This study has shown that background processing influences performance and uses Grad-CAM to understand how this happens. Therefore, including background management and explainability analysis in the proposed approach is justified.

Furthermore, the work of Yazdani et al. (2025)[11] is relevant because it shows the importance of considering both explainability and performance at once. The developed interpretable model D-ProtoPNet reaches an accuracy of 86% and an AUC of 93% and provides prototypical explanation evaluated by a radiologist. It means that clinical explanation goes beyond heatmap generation and needs to be verified by the expertise of a specialist in the field of radiology.

Finally, the use of transfer learning in this study is validated by the existing literature. For instance, Kermany et al. (2018)[2] proved the usefulness of transfer learning for classification of medical images and applied deep learning to pediatric pneumonia chest X-rays. Transfer learning reduces the number of resources but also introduces biases from the domain where the model was originally trained.

Quality of the dataset is a significant drawback as well. Chest X-rays' labels could have been obtained from radiology reports but not via pathologic verification, and possible uncertainty in the report could affect training of the model. CheXpert considered this problem of uncertainty in radiology-report-derived labels and



showed the difficulties associated with obtaining ground truth for chest radiograph AI models. Hence, future studies should aim to use expert-certified labels and clearly stated reference standards.

Class imbalance can also be a factor affecting the performance of the model. If there are significantly more patients with pneumonia than patients without pneumonia, the accuracy rate can seem to be higher even though the classification of normal cases is worse. This is why the present analysis reports sensitivity, specificity, precision, F1-score, and AUC along with accuracy rates. This multi-angle evaluation approach is more balanced.

The issue of external validation is also very important. Studies on evaluation of automatic chest radiograph algorithms show that AI can have good performance and can increase the accuracy of clinicians' performance as a second reader.

VI. CONCLUSION

In this research paper, an explainable AI based deep learning architecture for the classification of pneumonia from chest X-ray images was investigated. This method is developed using transfer learning CNN classifier along with Grad-CAM visual explanation, which helps in meeting two of the most crucial aspects of medical AI - performance of diagnosis and interpretability.

The results obtained through this study gave an accuracy of 94.3%, sensitivity of 95.4%, specificity of 93.1%, precision of 93.7%, F1-Score of 94.5%, and AUC of 0.97. From the results of interpretability analysis, it can be observed that pulmonary region localization was mostly seen in 91.4% of the test images.

However, good internal performance does not necessarily prove clinical efficacy. It has been shown that the performance of deep learning can drop when it is tested with an external dataset. Therefore, the suggested framework needs to be independently validated across several centers prior to clinical application.

The primary contribution of this study is the fusion of classification and explanation tasks rather than viewing them as two separate tasks. Grad-CAM can aid in checking whether the model is focusing on pulmonary regions and detecting possible weaknesses in its attention mechanism. However, heatmaps cannot be considered as evidence of clinical reasoning.

Future research should involve larger and more heterogeneous datasets, validated annotations, external validation, calibrated probability predictions, radiologist assessment of the explanations, uncertainty assessment, and clinical trials. Eventually, explainable AI should support doctors in making their decisions without shifting responsibility for the diagnosis onto machines.

VII. RECOMMENDATIONS

❖ Generate Bigger and More Diverse Chest X-Ray Datasets.

The studies need to have radiographs obtained at multiple hospitals and from different regions, ages, imaging modalities, and demographics. Bigger sample sizes help avoid overfitting and improve generalization. It is especially necessary to have more diverse samples because a model trained on one hospital dataset will learn the features of its acquisition but not the clinically relevant signs of pneumonia.

❖ Use Expert-Generated Reference Standards.

It would be advisable to focus on those reference standards that were generated via consensus among experts, using the clinical data, lab results, follow-up imaging or multi-modal reference standard when it is possible. Automated extraction of labels from reports should be analyzed for its uncertainty.

❖ Perform External Multicenter Validation.

One cannot claim that the developed model is ready to be used for clinical purposes just because it works well on internal test sets. External validation needs to be performed in independent hospitals or independent populations geographically. Separate performance of the model on the internal and external datasets should be reported.

❖ Incorporation of Explainability within Model Training

Methods such as Grad-CAM, SHAP, prototype methods, or others should be applied systematically. The explanations should be validated by qualified radiologists rather than evaluated just visually by the creators



of the model. The point is to determine whether the areas selected by the method correlate with actual pathological zones.

❖ **Model Bias and Background Dependency Testing.**

It should be determined whether the models react to edges of the image, markers, features of the equipment, positioning, text, or other background information that does not represent any pathology. Background filtering, lung segmentation, and bias testing will help to identify whether the model learns meaningful radiographic features.

❖ **Providing Several Diagnostic Criteria**

The accuracy of the pneumonia classifier should not be used as the sole criterion for evaluation. Sensitivity, specificity, precision, F1 score, area under curve, calibration, confusion matrix, confidence intervals, etc., should be provided.

❖ **Assess Explainability in Clinical Settings.**

Future research should include radiologists when evaluating whether the explanation provided by the artificial intelligence is relevant to the findings in the case of pneumonia. Explanation should not be assessed in relation to the accuracy of classification because even an accurate algorithm can provide poor explanations.

❖ **Assess AI as a Clinical Decision Assistant.**

Future research needs to assess whether the application of artificial intelligence helps improve the performance of radiologists in comparison with just comparing radiologist with AI. Such research would allow determining whether the tool is helping in improving sensitivity, specificity, reading times, etc.

❖ **Consider Prospective Clinical Assessment.**

Once the tool is externally validated, its prospective clinical assessment should follow. The research should include various severity of the disease and clinical settings.

REFERENCES

- [1] J. Irvin, P. Rajpurkar, K. Ko, Y. Yu, S. Ciurea-Illcus, C. Chute, H. Marklund, B. Haghighi, R. Ball, K. Shpanskaya, J. Seekins, D. A. Mong, S. S. Halabi, J. K. Sandberg, R. Jones, D. B. Larson, C. P. Langlotz, B. N. Patel, M. P. Lungren, and A. Y. Ng, "CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison," *Proc. AAAI Conf. Artif. Intell.*, vol. 33, no. 01, pp. 590–597, 2019.
- [2] D. S. Kermany, M. Goldbaum, W. Cai, C. C. S. Valentim, H. Liang, S. L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan, J. Dong, M. K. Prasadha, J. Pei, M. Y. L. Ting, J. Zhu, C. Li, S. Hewett, J. Dong, I. Ziyar, et al., "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell*, vol. 172, no. 5, pp. 1122–1131.e9, 2018, doi: 10.1016/j.cell.2018.02.010.
- [3] Y. Li, Z. Zhang, C. Dai, Q. Dong, and S. Badrigilan, "Accuracy of deep learning for automated detection of pneumonia using chest X-ray images: A systematic review and meta-analysis," *Comput. Methods Programs Biomed.*, vol. 198, Art. no. 103898, 2020, doi: 10.1016/j.cmpb.2020.103898.
- [4] D. Meedeniya, H. Kumarasinghe, S. Kolonne, C. Fernando, I. de la Torre Díez, and G. Marques, "Chest X-ray analysis empowered with deep learning: A systematic review," *Appl. Soft Comput.*, vol. 126, Art. no. 109319, 2022, doi: 10.1016/j.asoc.2022.109319.
- [5] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, M. P. Lungren, and A. Y. Ng, "CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning," *arXiv preprint arXiv:1711.05225*, 2017.
- [6] P. Rajpurkar, J. Irvin, R. L. Ball, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, B. Patel, M. Lungren, and A. Ng, "Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists," *PLoS Med.*, vol. 15, no. 11, Art. no. e1002686, 2018, doi: 10.1371/journal.pmed.1002686.
- [7] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 2097–2106.



- [8] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2017, pp. 618–626.
- [9] Y. Yang, G. Mei, and F. Piccialli, "A deep learning approach considering image background for pneumonia identification using explainable AI (XAI)," IEEE/ACM Trans. Comput. Biol. Bioinform., vol. 21, no. 4, pp. 857–868, 2024, doi: 10.1109/TCBB.2022.3190265.
- [10] R. Siddiqi and S. Javaid, "Deep learning for pneumonia detection in chest X-ray images: A comprehensive survey," J. Imaging, vol. 10, no. 8, Art. no. 176, 2024, doi: 10.3390/jimaging10080176.
- [11] E. Yazdani, A. Neizehbaz, N. Karamzade-Ziarati, and S. R. Kheradpisheh, "Explainable artificial intelligence for pneumonia classification: Clinical insights into deformable prototypical part network in pediatric chest X-ray images," J. Med. Imaging Radiat. Sci., vol. 56, no. 5, Art. no. 102023, 2025, doi: 10.1016/j.jmir.2025.102023.
- [12] A. S. Tejani, M. E. Klontzas, A. A. Gatti, J. T. Mongan, L. Moy, S. H. Park, C. E. Kahn, Jr., and CLAIM 2024 Update Panel, "Checklist for Artificial Intelligence in Medical Imaging (CLAIM): 2024 Update," Radiol. Artif. Intell., vol. 6, no. 4, Art. no. e240300, 2024, doi: 10.1148/ryai.240300.
- [13] S. H. Park and C. H. Suh, "Reporting guidelines for artificial intelligence studies in healthcare (for both conventional and large language models): What's new in 2024," Korean J. Radiol., vol. 25, no. 8, pp. 687–690, 2024, doi: 10.3348/kjr.2024.0598.
- [14] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, M. P. Lungren, and A. Y. Ng, "CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning," arXiv preprint arXiv:1711.05225, 2017.
- [15] H. Yoo, K. H. Kim, R. Singh, S. R. Digumarthy, M. K. Kalra, et al., "Development and validation of a deep learning-based automated detection algorithm for major thoracic diseases on chest radiographs," JAMA Netw. Open, vol. 2, no. 3, Art. no. e191095, 2019.
- [16] "Validating the accuracy of deep learning for the diagnosis of pneumonia on chest X-ray against a robust multimodal reference diagnosis: A post hoc analysis of two prospective studies," Eur. Respir. J., 2024.
- [17] "External validation of deep learning-based automated detection algorithm for chest radiograph: Practical issues in outpatient clinic," Eur. Radiol., 2023.
- [18] "Deep learning models to predict fatal pneumonia using chest X-ray images," Sci. Rep., 2023.
- [19] "Data-driven classification and explainable-AI in the field of lung imaging," Eur. Radiol. Exp., 2024.
- [20] World Health Organization, Guideline on Management of Pneumonia and Diarrhoea in Children Up to 10 Years of Age. Geneva, Switzerland: World Health Organization, 2024.
- [21] M. Chaudhari, "Hydrophilic nanofiber barrier coatings for aluminum: Waterborne crosslinking, microstructure–property links, and EIS-based protection," Genet. Mol. Res., vol. 25, no. 3, Art. no. 3017, Jul. 27, 2026, doi: 10.4238/rhd2k915.
- [22] R. Singh, "Brønsted-acid-guided esterification: Mechanistic and kinetic insights for long-chain fatty acid to alkyl ester conversion on a heterogenized porous acid platform," Int. J. Clin. Res. Med. Sci., vol. 1, no. 1, 2026, doi: 10.67231/rzakra50.
- [23] M. A. Jawed, "Managed aquifer recharge with advanced treated municipal effluent: Modelling the impact on groundwater chemistry and native microbial ecology," Int. J. Res. Technol., vol. 12, no. 3, pp. 128–138, 2024.
- [24] U. A. Kithran, A. Khanam, S. Ilyas, R. Yaqoob, and M. Hannan, "Demand forecasting in transportation networks using deep learning," Spectrum Eng. Sci., pp. 687–700, 2025.
- [25] S. M. Ali, M. Hadi, F. Usama, M. S. Khan, M. Iqbal, and S. Ahmad, "From artificial to quantum intelligence: The next frontier of low-carbon energy systems," in Proc. 2026 Int. Conf. Integrated Intelligence Cognitive Engineering (ICIICE), Dubai, United Arab Emirates, 2026, pp. 1–6, doi: 10.1109/ICIICE69672.2026.11565487.



- [26] A. Delshadi, M. D. Rasheed, N. Ahmad, Y. Khan, W. Khan, M. Akram, and M. Sultana, "Artificial intelligence for strengthening cybersecurity in educational technology systems," *Spectrum Eng. Sci.*, vol. 4, no. 3, pp. 299–311, 2026.
- [27] R. Singh, "Bench-scale nanomedicine preparation methods are poorly predictive of scalable manufacturing outcomes: A comparative lab study of liposomal doxorubicin," *Genet. Mol. Res.*, Sep. 23, 2026. [Online]. Available: <https://geneticsmr.com/index.php/gmr/article/view/5540>
- [28] S. M. Ali, A. J. Ehsan, A. Rahman, M. S. Khan, A. Khan, and S. S. Ali, "Quantum computing-enhanced intelligence for sustainable power systems: Toward next-generation smart grid architectures," in *Proc. 2026 Int. Conf. Integrated Intelligence Cognitive Engineering (ICIICE)*, Dubai, United Arab Emirates, 2026, pp. 1–6, doi: 10.1109/ICIICE69672.2026.11565373.
- [29] U. Iqbal, "AI-powered supplier risk intelligence: Predicting financial and geopolitical supply chain disruptions in U.S. critical industries," *J. Eng. Comput. Intell. Rev.*, vol. 3, no. 2, pp. 173–193, 2025.
- [30] U. Imtiaz, "Ghost Signals: Ethical RF adversarial testing of consumer alarm ecosystems," in *Data Science and Security (IDSCS 2025)*, S. Shukla, H. Sayama, K. Tiwari, J. P. George, and J. V. Kureethara, Eds., *Lecture Notes in Networks and Systems*, vol. 1945. Cham, Switzerland: Springer, 2026, doi: 10.1007/978-3-032-24075-0_21.
- [31] A. Khan, J. Iqbal, and D. Khan, "Artificial intelligence for financial data protection and business cyber resilience," *J. Comput. Syst. Eng. Insights*, vol. 4, no. 2, pp. 30–41, 2026.
- [32] M. Chaudhari, "Vapor-phase thin-film polymerization for microbial-resistant surface functionalization," in *Proc. Int. Conf. Sustainable Micro-Nano Materials & Innovative Technology (ICSUMMIT 2026)*, Vadodara, India, Feb. 2026, doi: 10.2991/978-94-6239-727-9_27.
- [33] U. Iqbal, "AI-driven predictive maintenance for U.S. smart manufacturing: Deep learning models for equipment failure prediction and operational resilience," *J. Eng. Comput. Intell. Rev.*, vol. 3, no. 1, pp. 114–138, 2025.
- [34] M. D. Rasheed, W. Khan, M. Imran, N. Ahmad, Y. Khan, M. Akram, and M. Sultana, "Leveraging artificial intelligence for advanced data networking and cybersecurity," *Spectrum Eng. Sci.*, vol. 4, no. 3, pp. 286–298, 2026.
- [35] L. H. Taylor, "Nanotechnology in biomedical engineering: Advances in drug delivery, diagnostics, and cancer therapy in the United States," *Amer. J. Multidisciplinary Knowledge Insights*, vol. 4, no. 2, pp. 1–8, 2025.
- [36] H. Kit, "The role of artificial intelligence in advancing industrial productivity and economic leadership in the United States," *Int. J. Bus. Integrity Technol. Adv.*, vol. 2, no. 1, pp. 27–36, 2026.
- [37] M. N. Amjad, "AI-enabled engineering project management: Evaluating productivity and risk reduction in the United States," *J. Comput. Syst. Eng. Insights*, vol. 3, no. 1, pp. 25–41, 2025.
- [38] U. Imtiaz, "Dynamic security certification framework for evolving distributed architectures," in *Data Science and Security (IDSCS 2025)*, S. Shukla, H. Sayama, K. Tiwari, J. P. George, and J. V. Kureethara, Eds., *Lecture Notes in Networks and Systems*, vol. 1947. Cham, Switzerland: Springer, 2026, doi: 10.1007/978-3-032-24360-7_32.
- [39] M. Z. Islam, I. Ahmed, R. Parveen, S. T. H. Rimon, and J. I. Janjua, "Revolutionizing healthcare for critical diseases with AI and enhancing security via low-latency edge intelligence," in *Proc. 2025 Int. Conf. AI-Driven STEM Education and Learning Technologies (AISTEMEDU)*, Piscataway, NJ, USA: IEEE, Dec. 2025, pp. 1–6.
- [40] I. Kate, "Marketing innovation in the U.S. e-commerce sector: Strategies for success in the evolving digital environment," *Int. J. Bus. Integrity Technol. Adv.*, vol. 1, no. 2, pp. 1–7, 2025.
- [41] U. A. Kithran, S. Ilyas, K. Amber, H. Abbas, and S. Zahid, "AI-based collision prediction and prevention systems in urban traffic," *Spectrum Eng. Sci.*, vol. 3, no. 12, pp. 1907–1920, 2025.
- [42] E. S. Walker, "Artificial intelligence in biomedical engineering: Current advances and future perspectives in the United States," *Amer. J. Multidisciplinary Knowledge Insights*, vol. 4, no. 1, pp. 1–8, 2025.



-
- [43] R. Bhattacharya, J. Mukherjee, S. Roy, M. T. Rana, and R. Parveen, "Eco-crypto dynamics: Cointegration of green and non-green cryptocurrencies for sustainable investing," *Adv. Consumer Res.*, vol. 2, no. 3, 2025.

