Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 12 , December - 2025

Opening the Black Box: Explainable AI in COVID-19 Medical Imaging - A Narrative Review

Greeshma K V
[1] Research Scholar, Department of Computer Science,
PSGR Krishnammal College for Women, Coimbatore,
Tamil Nadu, India.

Dr. J. Viji Gripsy
[2] Associate Professor, Department of Computer Science,
PSGR Krishnammal College for Women, Coimbatore,
Tamil Nadu, India.

Abstract - The global spread of COVID-19 has intensified the need for rapid and reliable diagnostic tools, particularly through chest X-rays and CT imaging. Artificial Intelligence (AI), and deep learning in particular, has demonstrated strong performance in detecting COVID-19, yet its opaque decision-making processes limit clinical trust and adoption. Explainable AI (XAI) provides a pathway to address this challenge by revealing the reasoning behind model predictions. This systematic review synthesizes current research on XAI applied to COVID-19 medical image analysis, evaluating widely used techniques such as Local Interpretable Model-agnostic Explanations (LIME), Gradient-weighted Class Activation Mapping (Grad-CAM), SHapley Additive Explanations (SHAP), and related methods. The study examines their effectiveness in improving interpretability, identifies key limitations, and discusses barriers to clinical integration, including usability, bias, and regulatory considerations. By highlighting both methodological strengths and practical challenges, this review underscores the role of XAI in enhancing transparency, fostering clinician confidence, and supporting safer AI-assisted diagnostics. Future directions are outlined to advance user-centric XAI approaches, multimodal imaging explainability, and integration into healthcare workflows, ultimately enabling more trustworthy and clinically meaningful AI applications in pandemic response and beyond.

Keywords: Explainable Artificial Intelligence (XAI), COVID-19, X-ray, Medical Image Analysis, Deep Learning, Interpretability.

1 INTRODUCTION

The outbreak of COVID-19 created an urgent demand for rapid and reliable diagnostic methods to support overwhelmed healthcare systems. While chest X-rays and CT scans quickly became essential tools for detecting lung abnormalities associated with the disease, the sheer scale of imaging data highlighted the limitations of traditional diagnostic approaches. Artificial Intelligence (AI), particularly deep learning, emerged as a promising solution by offering automated analysis and high accuracy in classifying COVID-19 cases.

Despite these advances, a critical barrier remains: the lack of transparency in AI decision-making. Deep learning models often function as “black boxes,” producing predictions without revealing the reasoning behind them. For clinicians, this opacity undermines trust and hinders adoption in practice, as medical decisions require clear justification to ensure patient safety. Explainable Artificial Intelligence (XAI) addresses this challenge by providing interpretable insights into how models reach their conclusions. By integrating XAI into medical image analysis, AI systems can become more transparent, fostering confidence among healthcare professionals and enabling responsible use in clinical workflows.

2 BACKGROUND STUDY

2.1 Deep Learning in Medical Imaging

Deep learning is a branch of AI that relies on artificial neural networks with multiple layers to learn complex patterns from large datasets. In medical imaging, convolutional neural networks (CNNs) are particularly effective because they can capture spatial hierarchies of features within images. For COVID-19 diagnosis, CNNs trained on chest X-rays and CT scans can detect subtle abnormalities such as ground-glass opacities or lung consolidations, enabling accurate classification of infected versus healthy cases.

Advantages of deep learning in COVID-19 imaging include:

- **Automation:** Reduces radiologists’ workload by processing large volumes of scans efficiently.
- **High accuracy:** Achieves diagnostic performance comparable to or exceeding human experts in certain tasks.
- **Disease monitoring:** Facilitates longitudinal analysis of patient scans to track progression.

Limitations: The primary drawback is interpretability. While deep learning models excel at pattern recognition, they rarely provide explanations for their outputs, making it difficult for clinicians to validate or trust the results.

2.2 Explainable Artificial Intelligence (XAI)

XAI encompasses methods designed to make AI systems more transparent and interpretable. Techniques such as Local Interpretable Model-agnostic Explanations (LIME), SHapley Additive Explanations (SHAP), and Gradient-weighted Class Activation Mapping (Grad-CAM) highlight the features or regions of an image that influence a model's prediction. By offering visual or quantitative explanations, XAI enables clinicians to understand the rationale behind AI-assisted diagnoses, identify potential biases, and ensure accountability in medical decision-making.

3 XAI Techniques

3.1 Model-Agnostic Techniques

These techniques can be applied to any AI model, regardless of its underlying architecture. They work by analyzing the model's behavior and identifying the features in the input data that have the most significant influence on the output prediction. Here are two common examples:

- **LIME (Local Interpretable Model-agnostic Explanations):** LIME approximates the local behavior of a complex model by training simpler, interpretable models like linear models or decision trees around a specific prediction [2]. This allows for the creation of explanations that are specific to a particular instance, highlighting the features in the input data (e.g., a chest X-ray) that most contributed to the model's prediction (e.g., presence of COVID-19).

- **SHAP (SHapley Additive exPlanations):** SHAP assigns a contribution score to each feature in the input data, indicating its relative importance in influencing the model's prediction [3]. This allows clinicians to understand how different features (e.g., lung opacity patterns) interact and contribute to the final diagnosis.

3.2 Model-Specific Techniques

These techniques are tailored to the specific architecture and characteristics of a particular AI model, such as deep learning models commonly used for COVID-19 image analysis. They leverage the model's internal structure and workings to explain its predictions. Here are two examples:

- **Grad-CAM (Gradient-weighted Class Activation Mapping):** Grad-CAM [4] focuses on analysing the gradients of the target class (e.g., COVID-19 positive) with respect to the final convolutional layer of a deep learning model. By visualizing these gradients, Grad-CAM highlights the regions of an image (e.g., specific areas of the lungs) that contribute most to the model's prediction.

- **Integrated Gradients:** This technique builds upon the concept of gradients but takes a more holistic approach. It integrates the gradients of the target class with respect to each layer of the model, ultimately providing a heatmap that highlights the image features that influence the prediction throughout the entire neural network.

These are just a few examples, and the field of XAI is constantly evolving. Choosing the most appropriate XAI technique depends on the specific AI model, the desired level of explainability, and the intended audience (clinicians vs. researchers).

By utilizing XAI techniques, healthcare professionals can gain valuable insights into how AI models arrive at their diagnoses. This transparency fosters trust in AI-assisted diagnostics and allows clinicians to:

- Understand the rationale behind the AI's prediction
- Identify potential biases in the model
- Improve the overall reliability and robustness of AI-powered diagnostic tools

Ultimately, XAI empowers humans to work alongside AI, leveraging its capabilities while maintaining control and accountability in the realm of medical decision-making.

4 LITERATURE REVIEW

Sarp proposes an XAI-enhanced transfer learning model for COVID-19 detection in chest X-rays. This method leverages pre-trained models for efficiency and XAI techniques for interpretability, achieving a high F1-score with ResNet [5]. While promising, future research should explore the specific XAI method used and validate the model's performance on larger and more diverse datasets for real-world clinical application. Ong [6] addresses the challenge of interpretability in AI-based COVID-19 diagnosis using chest X-rays. They analyze how LIME and SHAP, XAI techniques, can explain the predictions of a SqueezeNet model for classifying pneumonia, COVID-19, and normal lung images. While achieving an accuracy of 84.34%, the study's key contribution lies in demonstrating that LIME and SHAP successfully highlight the image regions critical for the model's decisions. This fosters transparency and interpretability, which are crucial for trust in AI-powered diagnostics. Future research could explore a wider range of XAI methods and assess their impact on clinician acceptance and real-world application.

Khan [7] proposes a deep learning framework with explainable AI (XAI) for COVID-19 classification using chest X-rays. To address limitations of manual diagnosis, their approach combines transfer learning for feature extraction with Grad-CAM visualization for interpretability. The framework incorporates pre-processing with hybrid contrast enhancement, improved feature fusion using iCCA, and classification with an ELM. This combination achieves high accuracy on public datasets, highlighting its potential for accurate and interpretable COVID-19 diagnosis.Bhandari et al. [8] developed a deep learning model to accurately classify Chest X-ray (CXR) images into four categories: COVID-19, Pneumonia, Tuberculosis (TB), and Normal. They integrated three major XAI techniques—Grad-CAM, LIME, and SHAP—to interpret the model's high-accuracy predictions (94.31%), aiming to provide clinicians with persuasive and coherent explanations for the categorization of these pulmonary diseases.

The paper by Ghnemat et al. [9] introduces an explainable AI (XAI) model focused on enhancing the interpretability of deep learning-based medical image classification, particularly on chest X-ray datasets for COVID-19 detection. The model leverages image segmentation to provide a clearer understanding of the features guiding the AI's decision-making process, ultimately offering a more transparent and practical diagnostic tool with a reported testing accuracy of 90.6%. Mallick et al. [10] provide a comprehensive systematic review of transfer learning (TL) techniques applied to classify COVID-19 cases using chest X-ray (CXR) images. The study provides a thorough evaluation of TL techniques, highlighting their potential for improving diagnostic accuracy while addressing critical challenges in data and methodology.

5 CHALLENGES AND CONSIDERATIONS OF INTEGRATING XAI

While Explainable Artificial Intelligence (XAI) holds immense promise for enhancing trust and understanding of AI-powered diagnostics in COVID-19, seamlessly integrating it into clinical practice presents several hurdles. Here are some key challenges to address:

- **User-Centric Design for Clinicians:** Many XAI techniques involve complex algorithms. The challenge lies in presenting explanations in a user-friendly and interpretable way for doctors with varying levels of technical expertise. Overly technical jargon can overwhelm clinicians and hinder workflow. XAI tools need to be intuitive and provide clear, concise summaries of the reasoning behind AI predictions. User interface design should prioritize simplicity and clarity, with interactive visualizations and context-specific explanations. Additionally, training programs for healthcare professionals can bridge the knowledge gap and facilitate effective XAI utilization.

- **Mitigating Bias in Explanations:** AI models are susceptible to inheriting biases from the data they are trained on. XAI techniques themselves can also introduce biases if not carefully designed and implemented. For example, a model trained on data with an overrepresentation of a particular demographic group might generate biased explanations that reflect those pre-existing biases. It's vital to ensure XAI methods are fair and unbiased in their explanations. Rigorous validation and testing for fairness and robustness are critical to mitigate bias and ensure equitable healthcare outcomes.

- **Balancing Computational Efficiency and Explanation Depth:** Some XAI techniques can be computationally expensive, potentially slowing down the diagnostic process. Clinicians need explanations delivered promptly to avoid hindering patient care. Finding a balance between the comprehensiveness of explanations and computational efficiency is crucial. Research efforts should focus on developing more efficient XAI algorithms and optimizing existing techniques for faster explanation generation.

- **Navigating the Explainability vs. Accuracy Trade-off:** There can be a trade-off between interpretability and performance. While more interpretable models may provide clearer explanations, they might sacrifice some accuracy compared to complex black-box models. Striking a balance is essential to ensure that XAI-enabled diagnostic tools are both reliable and transparent. Further research might explore techniques that improve the interpretability of high-performing models without sacrificing significant accuracy.

- **Regulatory Landscape and Establishing Trust:** Integrating XAI into clinical practice might necessitate new regulatory frameworks. Regulatory bodies may need to establish standards for the types of explanations required and the level of detail needed to ensure transparency and trust in AI-assisted diagnosis. Additionally, building trust with clinicians and patients regarding the use of XAI in healthcare is crucial. Educational initiatives can help address concerns and promote wider adoption.

Beyond Challenges: The Broader Benefits of XAI

While addressing these challenges is crucial, it's important to recognize the broader benefits of XAI:

- *Identifying Errors or Biases in AI Models:* XAI insights can help identify potential errors or biases within the AI model itself. By understanding the model's reasoning, healthcare professionals can flag inconsistencies and work with data scientists to improve the model's performance and overall reliability.

- *Enhancing Communication Between Doctors and Patients:* XAI explanations can be used to communicate the AI's rationale for its diagnosis to patients, promoting better understanding and shared decision-making. This fosters trust in the overall diagnostic process and empowers patients to actively participate in their healthcare.

By addressing the challenges and harnessing the full potential of XAI, we can pave the way for its successful integration into routine healthcare delivery. This will ultimately lead to more transparent, trustworthy, and efficient AI-powered diagnostics, not just for COVID-19 but for a wide range of medical conditions.

6 FUTURE DIRECTIONS

The intersection of XAI and AI-powered diagnostics for COVID-19 holds immense promise. Here, we explore some key areas for future research to further enhance XAI capabilities and facilitate its seamless integration into clinical practice:

- **Development of User-Centric XAI Techniques:** Future research should focus on developing XAI methods specifically tailored for clinicians with varying levels of technical expertise. Interactive visualizations, context-specific explanations, and natural language generation techniques can be explored to create user-friendly interfaces that present complex information in a clear and actionable way.

- **Explainability of Ensemble Learning Models:** Many deep learning models for COVID-19 diagnosis leverage ensemble learning approaches, combining predictions from multiple models. Current XAI techniques primarily focus on explaining individual models. Future research should explore methods for explaining the collective reasoning behind ensemble predictions, providing a more holistic understanding of the diagnostic process.

- **Explainable AI for Emerging Modalities:** As AI expands beyond chest X-rays and CT scans to encompass other modalities like lung ultrasounds, XAI techniques need to adapt and evolve. Research should address the specific explainability challenges associated with different medical imaging modalities.

- **Integration with Clinical Workflows:** For successful adoption, XAI needs to integrate seamlessly with existing clinical workflows. This could involve embedding XAI explanations directly within Electronic Health Record (EHR) systems or developing standalone applications that provide clear and concise explanations alongside AI-generated diagnoses.

- **Mitigating Bias in XAI Methods:** As with AI models themselves, XAI techniques can inherit biases from the data and algorithms used. Future research should focus on developing robust fairness assessment methods for XAI to ensure explanations are unbiased and reflect the true decision-making process of the underlying AI model.

- **Standardization and Regulatory Frameworks:** As XAI adoption in healthcare grows, standardized approaches for explanation formats and levels of detail are necessary. Collaboration between researchers, clinicians, and regulatory bodies is crucial to establish clear guidelines and ensure transparency and trust in AI-assisted diagnostics.

- **Explainable AI for Treatment Recommendation:** While the current focus lies on explaining diagnosis, XAI can be extended to explain AI-driven treatment recommendations. This would empower clinicians to understand the rationale behind suggested treatment plans, leading to more informed and patient-centred care.

7 CONCLUSION

This review highlights the pivotal role of explainable artificial intelligence (XAI) in strengthening the transparency and reliability of AI systems applied to COVID-19 medical imaging. Evidence from existing studies shows that approaches such as LIME, SHAP, and Grad-CAM can help overcome the “black-box” problem of deep learning models by clarifying how predictions are made. These explanations not only build trust but also give clinicians the ability to detect errors, validate outcomes, and refine models for better diagnostic accuracy. In doing so, XAI contributes to greater confidence among both healthcare providers and patients.

Looking ahead, future work should focus on embedding clinical context into explanations so that they are meaningful and actionable for practitioners. Advancing multimodal approaches that combine imaging with other patient data, incorporating methods to quantify uncertainty, and developing interactive explanation tools will be essential steps. Addressing these directions will ensure that XAI evolves into a practical companion for medical decision-making, enabling safer, more transparent, and more effective diagnostic practices not only for COVID-19 but also for a broad spectrum of medical conditions.

○ Ethics declarations

Conflict of Interest

There are no conflicts of interest in the content of this article.

Ethical Approval

Not required.

Funding Information

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

REFERENCES

[1] Van der Velden, B. H., Kuijf, H. J., Gilhuijs, K. G., & Viergever, M. A. (2022). Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. Medical Image Analysis, 79, 102470.

[2] Mishra, S., Sturm, B. L., & Dixon, S. (2017, October). Local interpretable model-agnostic explanations for music content analysis. In ISMIR (Vol. 53, pp. 537-543).

[3] Mangalathu, S., Hwang, S. H., & Jeon, J. S. (2020). Failure mode and effects analysis of RC members based on machine-learning-based SHapley Additive exPlanations (SHAP) approach. Engineering Structures, 219, 110927.

[4] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE international conference on computer vision (pp. 618-626).

[5] Sarp, S., Catak, F. O., Kuzlu, M., Cali, U., Kusetogullari, H., Zhao, Y., ... & Guler, O. (2023). An XAI approach for COVID-19 detection using transfer learning with X-ray images. Heliyon, 9(4).

[6] Ong, J. H., Goh, K. M., & Lim, L. L. (2021, September). Comparative analysis of explainable artificial intelligence for COVID-19 diagnosis on CXR image. In 2021 IEEE International Conference on Signal and Image Processing Applications (ICSIPA) (pp. 185-190). IEEE.

[7] Khan, M. A., Azhar, M., Ibrar, K., Alqahtani, A., Alsubai, S., Binbusayyis, A., ... & Chang, B. (2022). COVID-19 classification from chest X-ray images: a framework of deep explainable artificial intelligence. Computational Intelligence and Neuroscience, 2022(1), 4254631.

[8] Bhandari, M., Shahi, T. B., Siku, B., & Neupane, A. (2022). Explanatory classification of CXR images into COVID-19, Pneumonia and Tuberculosis using deep learning and XAI. Computers in Biology and Medicine, 150, 106156.

[9] Ghnemat, R., Alodibat, S., & Abu Al-Haija, Q. (2023). Explainable artificial intelligence (XAI) for deep learning based medical imaging classification. *Journal of Imaging*, *9*(9), 177.

[10] Mallick, D., Singh, A., Ng, E. Y. K., & Arora, V. (2024). Classifying chest x-rays for COVID-19 through transfer learning: a systematic review. Multimedia Tools and Applications, 1-60.

[11] Greeshma K V, Dr. J. Viji Gripsy. (2023). HOG-Based Machine Learning Models for Classifying COVID-19 in Chest X-Ray Images. Journal of Computational Analysis and Applications (JoCAAA), 31(4), 768–774.