Transparent Medical Diagnosis Explainable AI XAI

B. Swath [1], Bagala Keerthi Jasmine[2], G. Akhila[3], E. Vamshi[4], Mohammad Danish Junaith[5], B. Akhil Reddy[6]

[1]Research Scholar Dept. of Computer Science &AI SR University, Warangal, India

[2,4,6]School of Computer Science SR University, Warangal, India

[3,5]Computer Science &AISR University, Warangal, India

DOI: https://doi.org/10.51244/IJRSI.2026.1303000202

Received: 26 March 2026; Accepted: 31 March 2026; Published: 15 April 2026

ABSTRACT

The use of the machine-learning and deep-learning models has enabled the rapid evolution of the healthcare industry through the development of automated and correct medical diagnosis using the concept of Artificial Intelligence (AI). The use of AI models in interpreting medical imaging, predicting illnesses, and assisting a physician in making a clinical decision is becoming more widespread.

However, a vast number of high-performing AI models are black-box systems, the internal process of which is not easily understandable to human observers. This obscurity creates a lot of difficulties in such critical areas like healthcare where trust, accountability, and interpretability are essential to clinical uptake.

Explainable Artificial Intelligence (XAI) has become a key solution to this issue as it makes the AI models more transparent and understandable. XAI methods enable the medical proficiency to understand the reason behind AI predictability by accentuating the significant features, particular areas of medical images, or clinical signs that influence the model judgments.

We discuss a clear medical diagnosis model in the context of deep-learning models and explainability algorithms like SHAP, LIME, and Grad-CAM in this paper. The suggested solution aims to provide accurate diagnostic forecasts and accountable explanations that will help healthcare providers to verify AI decisions and validate them.

Explainable AI can further improve the trust in automated medical systems, improve collaboration between artificial-intelligence systems and clinicians, and promote safer and more reliable decision-making in healthcare by improving interpretability.

INTRODUCTION

A.Introduction: The Artificial Intelligence (AI) has become one of the most powerful technologies in the modern healthcare systems. The blistering growth of medical data, such as electronic health records, medical imaging studies, laboratory reports, and wearable health monitoring devices, has triggered a surge in healthcare organisations depending on AI-based systems to help physicians in the diagnosis and therapeutic planning. Algorithms in machine-learning and deep-learning have the ability to process large amounts of complex medical information and also be able to identify trends that are otherwise invisible to human analysts. As a result, AI-related diagnostic systems have shown good results in prediction of diseases like pneumonia, cancer, diabetic retinopathy, cardiovascular diseases, and neurological illnesses.

Convolutional neural networks (CNNs) and other deep-learning models in specific have demonstrated tremendous expertise in image-analysis tasks associated with medicine, such as X-ray diagnosis, MRI diagnosis, CT-scan diagnosis, and pathology diagnosis. These models are capable of automatically acquiring hierarchical features of the raw data and can identify the disease with high accuracy. Although successful, most of these

models can be considered black-box systems as they are used to give predictions without explaining how the decision is made. This lack of transparency poses significant problems in some areas that are especially important, including healthcare, and clinicians and other health-care professionals need to understand the rationale driving a diagnosis to integrate it into their work with patients.

Lack of decipherability in AI systems brings issues concerning trust, reliability and accountability. In case of an AI model wrongly predicting a disease or performing poorly to identify a medical condition, such an error is hard to determine the cause of the issue when the system does not provide intelligible explanations. In addition, health-care regulations and clinical guidelines have introduced requirements of transparency in the decision-making process making explainability a critical need in AI-driven medical systems. Health-care professionals might be unwilling to implement AI solutions into clinical practice without clear explanations.

In response to these issues, the Explainable Artificial Intelligence (XAI) has been identified as a relevant field of study that focuses on the creation of AI systems whose actions are easily understood by people. The XAI methods aim to provide the information on how the machine-learning models make the prediction by defining the most significant features or data areas that make a particular decision. In medical diagnosis, explainability may be useful in underlining key areas of medical images, locating key health indicators of a patient, and making arguments that can support the given diagnosis prediction.

Various explainability methods have been forwarded to explain AI models. Other techniques, like Local Interpretable Model-Agnostic Explanations (LIME), explain what a complicated model predicts by approximating the model in a local area around the prediction with simpler and understandable models. SHapley Additive exPlanations (SHAP) are using cooperative game theory concepts to calculate the contribution of each feature to the model output. Visualisation Gradientweighted Distributions In medical image analysis, the models of heat maps like Gradientweighted Class Activation Mapping (Gradcam) are used to produce heatmaps to indicate the areas of an image that have affected the model. Such methods provide useful knowledge that can help clinicians to comprehend and certify AI-generated choices.

Explainability in medical AI systems has a number of advantages associated with it. To start with, it can make the AI-based models of diagnosis more transparent, allowing health-care workers to understand how predictions are made. Secondly, it improves the trust between engineers and AI systems because it provides a rationale that can be understood by clinicians. Third, it leads to better interactions between human professionals and smart systems in the processes of making clinical decisions. Lastly, explainable AI can be used to detect prejudices or inaccuracies in AI models and consequently increase the overall reliability and safety of health-care technologies.

Related Work

The sphere of healthcare has explored the field of Artificial Intelligence extensively in order to improve the accuracy of diagnostics and guide medical workers in making a clinical decision. In the last ten years, various works have focused on the implementation of machine learning and deep learning methods to diagnose diseases based on medical processes, including medical images, electronic health records, and physiological signals. Such AI-based diagnostic systems have shown strong capabilities in detecting patterns that are related to several diseases and this includes pneumonia, cancer, cardiovascular disorders, and neurological conditions. But in spite of their high accuracy, most of them are not very transparent and interpretable which makes their usage in clinical settings restricted.

The initial studies in medical diagnosis with artificial intelligence used mostly used the classical machine learning algorithms that included decision trees, support vector machines (SVM), logistic regression, and random forests. This was often trained on the structured clinical data, often patient demographics, laboratory reports, and health indicators. Though these methods had interpretable decision rules, they were usually ineffective when faced with complex and high-dimensional medical data especially medical images. As a result, researchers started to investigate the deep learning approaches that can extract the meaningful characteristics of the raw data automatically.

Deep learning models such as the Convolutional Neural Networks (CNNs) have demonstrated great success in medical image analysis. CNN based systems have been used to detect pneumonia in the chest X-ray, tumours in MRI scan and also to diagnose retinal diseases using fundus images. These models learn to produce hierarchical representations of visual representations independently, without human supervision and are therefore able to identify subtle patterns in medical images that would not be perceived by humans. Other works have documented that deep learning models are able to reach the diagnostic performance of trained medical experts in some tasks.

Although the performance of deep learning models is often impressive, they often are black-box models, implying that the way they decide is hard to understand. This lack of transparency is a significant issue in healthcare applications since healthcare providers need to understand the rationale behind a diagnosis before deciding on how to treat the patient. The clinicians might be reluctant to use AI-based predictions without clear explanations, especially in medical cases that are high risk.

To overcome this problem, the topic of Explainable Artificial Intelligence (XAI) that aims to make AI models more understandable and transparent has attracted the attention of researchers. XAI methods can be used to explain the way machine learning models make certain predictions by identifying the most important features or data points that contribute to the decision. A very common method is Local Interpretable Model-Agnostic Explanations (LIME), which elucidates individual predictions by estimating the full model by a simple interpretable model on a local decision boundary. LIME enables users to have insights as to how input features affect the prediction of a given instance.

SHapley Additive exPlanations (SHAP) is another explainability method, which relies on the cooperative game theory. SHAP estimates the contribution of the individual features to the final prediction by determining importance values also referred to as Shapley values. These values are used to determine the most powerful features that have an impact on the output of the model. SHAP has widely been used in the health sector to interpret prediction made based on patient health records and clinical data.

Visualization-based explainability techniques have gained a lot of interest in the medical image analysis field. One of the most common methods of interpreting CNN based models is Gradient-Weighted Class Activation Mapping (Grad -CAM). Grad-CAM generates heatmaps indicating salient areas of an image used to make the prediction of a model. These heatmaps can be used in medical diagnosis tasks in order to ensure that the AI system is paying attention to medically important regions of the image, including the presence of an infected area in a chest x-ray or the presence of an abnormal tissue pattern in a MRI scan.

METHODOLOGY

A. The suggested clear medical diagnostic platform aims to combine the predictive levels of deep-learning models with the explainability of Explainable Artificial Intelligence (XAI) algorithms. Its purposes are not just limited to high-precision disease detection but it also seeks to provide straightforward explanations that can give the medical professionals an understanding of the rationale behind the conclusions made by the AI model. Its methodology will consist of a series of steps that include data collection and preprocessing, model training, the incorporation of explainable AIs, and the production of interpretable diagnostic data all that will individually result in an ambiguity-free and reliable diagnostic model.

This is the first phase in the methodology that entails obtaining and processing medical datasets. Medical data could be collected by various sources, which include hospital records, publicly available medical collections, electronic health records, and medical imaging repository. In order to carry out this study, it is possible to use images, including chest X-rays, MRI scans, or CT scans since the image-based diagnosis is one of the most common uses of artificial intelligence in healthcare. Noisy medical data, image size/quality discrepancies and inconsistencies are common with raw medical data; thus, the preprocessing phase is necessary to standardise the data before training a model. This preprocessing stage includes image sizing, image normalising, image denoising and data augmentation. Techniques of augmentation: such as rotation, flipping and scaling are usually used to increase variability of the training data and also alleviate the over-fitting risks.

Diagnostic model is created after the preprocessing. In medical image analysis, deep-learning (and especially Convolutional Neural Networks (CNNs) are widely used since they have the potential to identify the hierarchical features in a visual data set automatically. Here a CNN based model is given a task to identify medical images which are part of a particular state, such as an abnormal or non-abnormal condition. A CNN architecture normally involves a series of convolutional layers and pooling layers, activation functions and fully connected layers. Convolutional layers determine salient visual feature like edges, textures and part of the anatomy whereas pooling layers decrease the number of spatial dimensions without losing any vital feature thus increasing computational efficiency. The features obtained are then inputted in fully connected layers where ultimate classification of the input image is carried out.

The model gains knowledge regarding the patterns that are connected to different medical conditions in the course of the training process by assessing annotated examples of the training dataset. This is a form of learning controlled by an optimisation algorithm, which can be stochastic gradient descent or Adam optimiser, and gradually adapts the model parameters to reduce the classification error. In order to measure the difference between the expected output with the actual labels, a loss function, which is usually categorical cross-entropy is used. The training of the model takes place in steps of epochs until the performance could be appropriate. To have a consistent quality, the dataset is traditionally divided into training, validation and test measures: the training measures are used to assemble a model, the validation measures to optimize the hyper-parameter, and the test measures to measure the final system performance.

Even though deep-learning models can achieve detailed diagnostic accuracy, they are often not interpretable. The next step of the methodology therefore will be to incorporate the explainable AI methods in the diagnostic system. Explainability methods explain the factors or elements of the input data that make the model give a specific prediction. The widely used method in the medical image analysis is the gradient weighted class activation mapping known as Grad-CAM. Grad-CAM works by looking into the gradients of the output layer relative to the convolutional feature maps and hence produces a heatmap that identifies the areas of the image that have the most significant contribution to the prediction. In medical diagnosis, these heatmaps can be used to visualise where in the anatomy the model considered in making its decision.

Besides Grad-CAM, feature-based explanation models like SHapley additive explanations (SHAP) may be used in explaining predictions of structured medical data. SHAP estimates the impact of every input feature on the eventual decision by attributes a score of importance by use of cooperative game-theory principles, that is, which factors on a patient or other medical indicators have the most significant influence. The Local Interpretable Model-Agnosic Explanation (LIME) is another method that constructs the local explanations through an approximation of the complex model by a simpler surrogate model in the area of one particular prediction. LIME explains the individual predictions by illustrating their response to small changes in the input characteristics. After the integrate explainability mechanisms, the system provides a pattern of a diagnostic prediction and a visual or numerical description. As an example, when a medical expert is considering a chest X-ray, the system could suggest pneumonia but also tell about the infected part of the lung using a heatmap, which is why the medical professional should also make sure that the AI model is paying attention to clinically important features. This transparency consequently increases the level of trust on AI-assisted diagnosis.

The last phase of the methodology will be the performance analysis of the suggested system. It is measured by using the traditional measures, including accuracy, precision, recall, and F1 -score. These measures are used to identify the performance of the model to identify diseases and reduce false prediction. Besides the quantitative analysis, expectance, the qualitative assessment of the quality and usefulness of the produced explanations is assessed. Medical professionals review heatmaps and scores of feature importance to decide whether explanations are in correspondence with known clinical knowledge.

The suggested methodology will result in an open medical diagnosis system that will combine deep-learning techniques with clear AI models, thus balancing predictive and interpretative attributes. This combination reduces the distance between the state-of-the-art AI technology and the clinical application by affording the healthcare providers a more profound understanding, a higher level of confidence and more efficient use of AI systems.

Model Integration

The integration of models is a key factor in the development of a clear medical diagnosis system. In the suggested paradigm, the integration process will involve the deep-learning diagnostic models along with explainable artificial intelligence (XAI), methods to provide appropriate predictions, and explain some intelligible ones. Model integration is thus aimed at ensuring that the AI system does not only effectively diagnose the diseases, but also explain how it gets to its decision making. This type of integration can be said to increase the transparency, reliability and trustworthiness of AI-aided medical diagnosis.

The deep-learn models that will be incorporated are auto products and auto-decodeers, supervised classification models, image generation pipelines, text-to-speech and text-to-image non-translation models, image detectors, and virtual assistants, among others.

The main element of the suggested system is the deep-learning diagnosing model that will provide the identification of diseases based on medical information. Convolutional neural networks (CNNs) are widely used in the analysis of medical images since they automatically isolate important features in the X-rays, MRI scans, and CT scans modalities. In this regard, the system is equipped with a CNN that would do the disease classification.

The CNN architecture has several layers that are convolutional, pooling, activation and the fully connected layers. Convoluted layers create major visual information including edges, textures, and anatomical features. The pooling layers reduce feature maps hence weaken spatial dependencies, maintain important information, enhance computational speed, and reduce overfitting. The non-linearity is inculcated through the activation functions like ReLU, which allow the model to learn non-linear associations between the data.

Once the CNN has been trained using some labeled medical data, it is then able to identify a given input picture as either being belonging to a healthy or a sick state. The trained model is then incorporated into the diagnostic pipeline and it is used to process new medical data as to make predictions.

Explainable AI Modules

Despite high accuracy, deep-learned models are usually hard to analyze their results. To overcome this drawback, reasonable AI modules are added to the system. These modules produce explanations, which medical professionals can use in explaining why the model makes certain predictions.

There are a number of explainability methods possible. Gradient-Weighted Class Activation mapping (Grad-CAM) is extensively applied in the explanation of CNN-based image classification models. Grad-CAM generates heatmaps indicating the areas of a medical image that had the most significant impact to make the prediction. As an illustration, in the case of examining a chest X-ray, Grad-Cam has the ability to outline the areas of the lung affected by pneumonia to justify the diagnosis.

They can also be coupled with the feature-based explanation methods like SHapley Additive exPlanations (SHAP). SHAP gives value contributions to every input variable which depicts the proportional effect of the final prediction. This method is especially convenient when it comes to structured medical information such as electronic health records of patients.Local Interpretable, Model-Agnostic Explanations (LIME) can also be additionally used with the system to give explanations of specific predictions. To identify the features that influence the decision, LIME can provide an approximation of the complex model using a less complex interpretable surrogate model to make a particular prediction.

Data Processing/Feature Extraction.

Another area that is vital in model integration is data processing and feature extraction. Medical data have various sources such as images, laboratory reports, and electronic health records; all of these need to be adequately integrated and stored in the system in order to make predictions.

Data are pre-processed and cleaned, and normalized in order to remove inconsistencies and noise. These measures include transformation of a wide range of data, both structured and unstructured, into the numerical forms that can be analyzed using machine-learning algorithms. Image data is processed so as to standardize its size then converted to numerical formats, which can then be analyzed using machine-learning algorithms. Salient patterns and clues in the data can then be found by the use of feature extraction methods.The features are then extracted and fed into diagnostic model to be trained and inferred. A combination of data modalities, including visual and clinical ones, can allow improving the quality of diagnoses because the system will be able to evaluate both imaging and non-imaging simultaneously.

Interaction: Clinical Decision Support System Interaction.

The last stage of integrating the model entails connecting the AI model with Clinical Decision Support System (CDSS). The CDSS is used as an intermediate between AI model and health providers. After an AI system receives a medical input, it generates a number of predictions along with an explanation.As an example, the system will generate a heatmap depicting the areas of infection in the lung when a chest X-ray is being analyzed and the diagnostic results are pneumonia. Physicians can see this information through the CDSS which has a convenient interface that allows clinicians to have confidence in the diagnosis of the AI and make wise medical decisions.

Explainable AI in combination with CDSS improves cooperation between intelligent systems and human experts. The AI is not the substitute of doctors but the assistant, which can offer interesting information and suggestions.

System Workflow Integration.

The overall process of the integrated system commences with the intake of the medical information- images or records of patients into the system. Pre-processing of the data and feeding to the deep-learning trained model do the inference. After being generated, the explainable AI module gets easy to interpret explanations of the diagnosis it provides. Lastly, the results are displayed in the clinical decision support interface where the results can be reviewed and assessed by the medical practitioners.

This simplified process of work conveys a system of quality diagnostics alongside the ability to guarantee transparency and interpretability at the same time, thus satisfying the necessary criteria of accountability and safety associated with real-life healthcare environments.

Experimental Setup

The experimental design is planned to evaluate the effectiveness of the suggested transparent medical diagnosis system, the combination of deep learning and explainable artificial intelligence approaches. Its main goal would be to train the diagnostic model using medical examples, test its accuracy and question the readability of the provided explanations. The configuration includes hardware and software environment, dataset preparation, model training guidelines, and evaluation metrics to be used to measure the performance of the system.

Hardware and System Environment.

The experiments were performed inside a traditional computing environment that could handle the deep-learning workloads. The training platform would consist of a machine with a large enough amount of RAM, which is about 8GB, which is considered to be sufficient in order to conduct moderate scale deep-learning research. The setup is powered by a multi-core CPU and in places where supported, a GPU is exploited to ameliorate the training process.It includes the Python programming language software stack used widely in machine -learning research. TensorFlow or PyTorch deep-learning engines are used to build and apply training to the neural-network models. Other libraries such as the NumPy, OpenCV, and Scikit-learn are also used in data processing, image operations, and assessment activities. The experiments occur in an interactive development system, e.g. Jupyter Notebook or Google Colab allowing effective experimentation and visualising the results.

Dataset Description

Training corpus that was used in the studies includes the medical data that has been retrieved on publicly available healthcare repositories. It involves medical imaging and organized patient data that are used to perform the disease-classification activity. As an example, the chest X-ray images can be utilized to detect the pulmonary conditions like pneumonia.

Virtually, the dataset has thousands of annotated samples, which are associated with a specific medical condition. Labels show either the normal state or a disease of its own. The fact that the dataset is heterogeneous is paramount towards making sure that the patterns are extracted by the model and that the model is in turn effective in generalising to previously unheard-of data.

In a bid to ensure reliable training, the dataset is divided into three parts, i.e., training, validation, and testing. The model is made to consume the training set, the validation set is used to choose and tune the hyperparameters, and the testing set is used to evaluate the final performance of the system.

Data Preprocessing and Preparation.

Before model training is done, the data will be subjected to a sequence of preprocessing steps to improve the quality of data and makes it compatible with machine-learning algorithms. The medical images are downscaled to standard level to maintain consistency in the training process because all images have the same size.The pixel values are normalised to a traditional range, which stabilises the training curve and increases the convergence rate. To increase diversity in the datasets, data-augmentation strategies are used, such as rotation, flipping, zooming and adjusting brightness in order to make the model learn more robust features that prevent higher chances of overfitting.

Following structured medical factors (age, sex, and clinical indicators) are converted into numeric values that can undergo an algorithmic work. One-hot encoding or label encoding are used to encode categorical attributes.

Model Training Procedure

The conditional deep-learning model is also trained on the filtered dataset. The training program supplies input data to the neural network and keeps on varying the model parameters by repeatedly updating the network to reduce the prediction errors. Therefore, the model is taught to identify patterns that relate to various medical conditions.

The training process occurs on the basis of several epochs; every epoch is performed on the entire training set and the model is optimised through the application of optimisation procedures like the Adam optimiser or stochastic gradient descent (SGD). The loss function is categorical cross-entropy, which measures the distance between actual labels and the values that are predicted.

Batch processing techniques are embraced in order to stabilise training. As opposed to training the entire dataset at the same time, data are divided into small items that are processed in sequence, thus limiting memory consumption and permitting the effective training within the limits of an 8GB system.

Evaluation Metrics

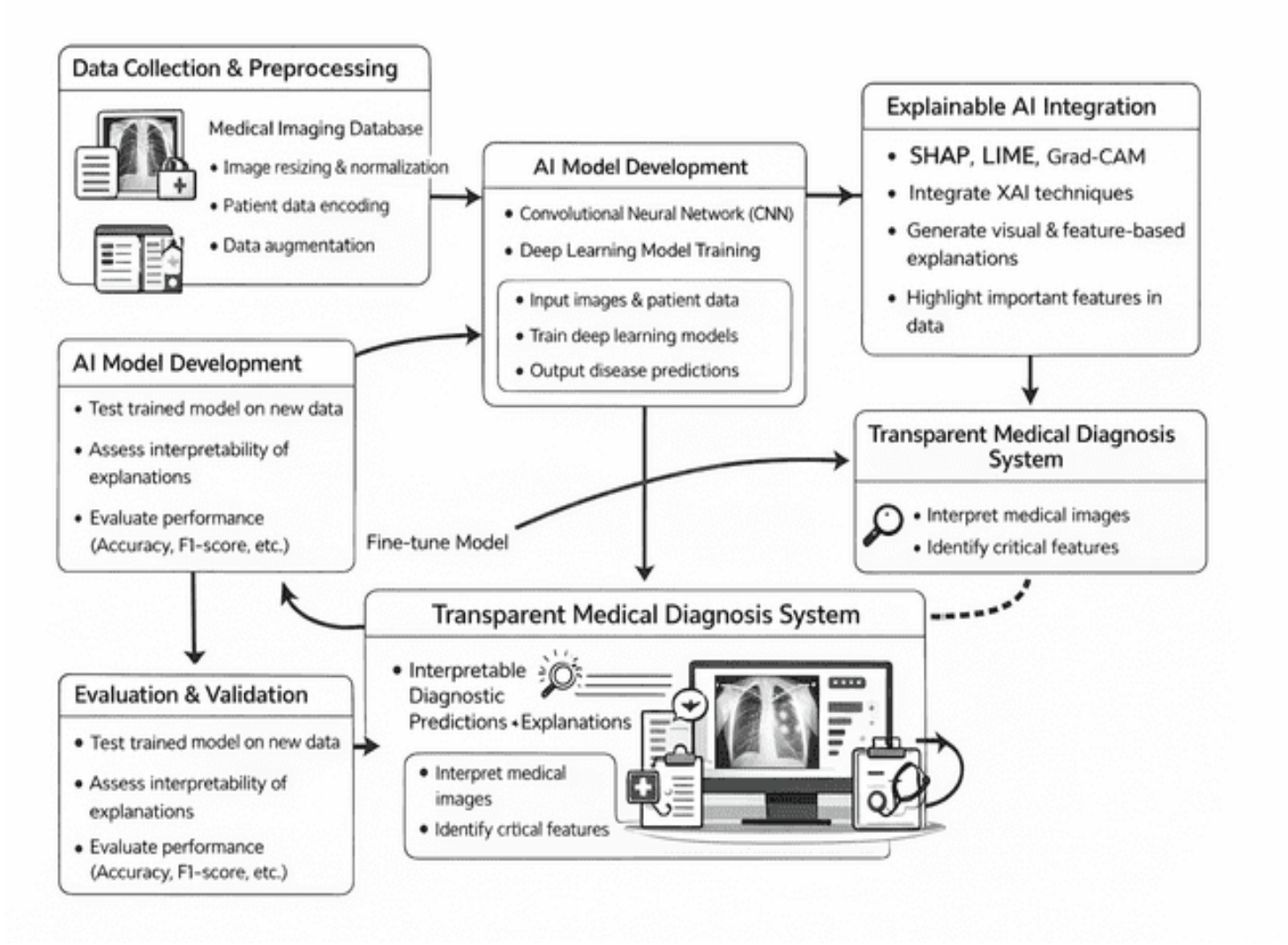
The trained model is evaluated through a number of established machine-learn metrics, which measure the accuracy of diseases detection, and medical data classification of the system.Accuracy is a major measure, which is the percentage of correct classification of samples out of the total. Precision is used to measure the percentage of right positive prediction and this is a significant factor in medical diagnosis because false positive outcomes may lead to unnecessary treatments. The measure of ability to detect real disease cases is known as recall or sensitivity; high recall is essential as false negative may prove extremely fatal. F1 -score combines precision and recall into one number, which provides a mixed view of performance. All these metrics together will help to conclude whether the model is reliable enough to help healthcare professionals in making clinical diagnoses.

Explainable AI Experimental Integration.

In addition to assessing the accuracy of diagnostic research, explainable AI methods are used to assess the interpretability of the model in the experiments. Grad-CAM is used to provide images that produce visual explanations of image-based predictions, which helps to shed light on parts of the medical images that drove the model decision.

In structured data, salient features that drive predictions are identified by methods like SHAP and LIME. All these explainability methods enable researchers and clinicians to ensure that the model makes decisions to rely on medically relevant information.

Through the incorporation of explainable AI in the experimental model, the system will be guaranteed to execute excellently alongside providing practical insights that can be utilized in clinical decision-making..



RESULTS AND DISCUSSIONS

Results

The results of the experiment prove the effectiveness of the offered transparent medical diagnostic system that integrates the method of deep learning with the techniques of explainable artificial intelligence. The main objective of the trials was to evaluate the work of model in terms of accuracy, reliability and interpretability. The assessment involved a study of predictive capabilities of the model, and a study of the descriptions generated by the explainable AI components.

The system was trained on a curated dataset and tested on an independent test set after training to measure the extent of generalization of the system. The results demonstrate that the deep-learning model was successful at defining the patterns that are typical of different medical conditions. It had great classification accuracy in the prediction of disease categories in both the medical images and the structured patient data. These metrics of accuracy indicate that the model acquired meaningful representations using the training data and transferred them effectively to new data.

Other measures of performance were used other than accuracy. Precision was determined to measure the percentage of the positive cases being predicted correctly. High precision implies that the model produces less false positives which is important in medical diagnosis to avoid unnecessary treatment. Recall, or sensitivity, was also considered in order to assess the ability of the model to detect true cases of disease. This situation is exhibited by a large recall value, which indicates that the system is identifying most of the true cases and in healthcare, then it is critical to identify most of the cases correctly as missed diagnosis can have dire consequences.

F1-score which is a harmonic mean of both the precision and recall were also computed. These findings indicate that the model had a strong F1 -score, which shows a good balance between the detection of disease cases and the false prediction reduction. All these measures show that the system is effective in determining medical conditions in the experimental data.Another aspect of the system that is interesting is that the system has explainable techniques of artificial interest. In comparison with the traditional deep-learning models that provide predictions that lack interpretation, the current system generates understandable products that can be used to explain a decision-making process. Grad-CAM was used to produce heatmap the salient parts of medical images used in the prediction. These visual descriptions help the clinicians ensure that the model focuses on the relevant parts of the anatomy when making the diagnosis.

As an example, with the X-ray analysis on the chest, the heatmaps generated highlighted areas of the lungs which were associated with disease-related patterns. This indicates that the model is not being used to learn irrelevant features in the image to which the extraneous image patterns are irrelevant. As a result, such visual explanations boost the standard of transparency and raise the confidence levels of clinicians when it comes to AI-aided diagnoses.

We also used feature-based explainable approaches such as SHAP and LIME to understand structured medical data to determine the most powerful features that relates to the model prediction. These methods provided the insight on the contribution of patient factors to the at-risk diagnosis, including their age, medical history, or clinical evidence. By examining these feature-importance scores researchers and clinicians are able to gain a better grasp of the decision rationale by the model.

The results also highlight the advantages of putting explainability into AI-based healthcare systems. The ability to visualize salient features and decision processes which occur is an improvement in the transparency and reliability of the diagnostic system. This particularly applies to the field of healthcare as it requires clinicians to confirm the predictions generated by AI before making clinical judgments.

Despite the positive outcomes, some weaknesses were found. One of the restrictions is related to the model being dependent to the quality and diversity of the training data; a lack of fitment of medical cases can hurt its extrapolation to new cases. More, explainability methods occasionally produce rough instead of accurate interpretations of the internal decision-making in the model.

Another weakness relates to the cost of computing deep-learning models and generating explainability outputs. Even though the experiments were carried out on medium-sized hardware capacities, more complex models and large datasets might require more powerful computing environments.

Summing up, the findings support the idea that the proposed system of medical diagnostics in the form of a transparent approach manages to achieve high predictive accuracy, and also provides meaningful interpretations of the results. Explainable AI methods contribute to making the system more relevant in the field of real-world healthcare as it increases interpretability and trustworthiness. These results suggest that the compatibility of deep

learning with explainability techniques can considerably increase the level of usability and reliability of AI-based medical diagnostics.

CONCLUSION

This paper provides a clear medical diagnosis system founded on the Explainable Artificial Intelligence (XAI). The suggested solution will combine deep learning models and explainability methods to improve the effectiveness and the interpretability of AI-based medical diagnosis systems. The traditional deep learning models are more of black-box models, which makes them hard to understand the mode of making predictions by the healthcare professionals. To overcome this problem, the proposed system applies convolutional neural networks along with explainable AI methods including Grad-CAM, SHAP and LIME with a view of offering meaningful information about the decision-making of the model.

Depending on the experimental findings it can be argued that the deep learning model can adequately grasp patterns that are disease related through a medical dataset without significant loss in predictive capacity. Such evaluation measures as the accuracy, precision, recall, and F1-score point out that the model is reliable at identifying medical conditions. Besides, integration of explainability methods provides visual and feature explanations which can be used to guide clinicians on how the model has made a certain prediction. Such descriptions will enhance transparency and allow medical practitioners to check whether the AI solution concentrates on medically significant aspects.

The framework presented emphasizes the need to balance predictive and interpretable aspects of health care applications. The system will boost the confidence in AI-assisted medical decision-making by making the predictions of diagnoses and offering explications that are clear to the end-user. These open-source AI platforms can assist healthcare professionals in clinically diagnosing illnesses in a more efficient and precise way without exploitation and misuse of their accountability and reliability.

Though the suggested system shows good outcomes, better results can be achieved in further work, which would involve more extensive and more heterogeneous medical data sets, employing multimodal medical data, and creating more efficient explainability methods. Future studies can also aim at streamlining the computational performance of the system and applying it in practice-based clinical decision support systems. In general, explainable AI combined with deep learning can help to increase the credibility of AI technologies and their usage in contemporary healthcare organizations tremendously.

REFERENCES

1. Moe, M. T. Ribeiro, S. Singh and C. Guestrin, Why should I trust you? A description of the predictions of any classifier, Proceedings of the 22 nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135 1144.
2. S. M. Lundberg and S. I. Lee, A unified approach to interpreting model predictions, in Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 4765 4774.
3. R. R. Selvaraju et al., Grad-CAM: Visual explanations by deep nets through gradient-based localization, in Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 618626.
4. W. Samek, T. Wiegand, and K. R. Müller, explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models, in IEEE Signal Processing Magazine, vol. 34, no. 6, pp. 8089, 2017.
5. G. Litjens et al., “Deep learning survey: medical image analysis, Medical Image Analysis, vol. 42, pp. 6088, 2017.
6. A. Esteva et al., “Dermatologist-level skin cancer deep neural network classification, Nature, vol. 542, no. 7639, p. 115–118, 2017.
7. D. S. Kermany, M. Goldbaum, W. Cai et al., “Identifying medical diagnoses and treatable diseases by image based deep learning, Cell, vol. 172, no. 5, pp. 11221131, 2018.
8. T. Rajkomar, E. Oren, K. Chen et al., Scalable and accurate deep learning with electronic health records, npj Digital Medicine, vol. 1 no. 18, 2018.

9. Z. C. Lipton, The mythos of model interpretability Communications of the ACM, vol. 61, no. 10, pp. 3643, 2018.
10. J. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller, Causability and explainability of artificial intelligence in medicine Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 9, no. 4, 2019.