Jurnal Teknik Informatika C.I.T Medicom, 17 (6) (2026) 204-315

Published by: Institute of Computer Science (IOCS)

Jurnal Teknik Informatika C.I.T Medicom

Journal homepage: www.medikom.iocspublisher.org

Transparency Analysis of Deep Learning Models in Medical Data Using SHAP and LIME

Arka Evander[1], Lyra Amara Quinn[2]
[1,2] Department of Computer Science, University of Luxembourg. Luxembourg

Article Info

Article history

Received : Oct 28, 2025
Revised : Dec 27, 2025
Accepted : Jan 10, 2026

Keywords:

Deep Learning;
Explainable AI;
SHAP;
LIME;
Medical Data.

Abstract

The increasing adoption of deep learning models in healthcare has significantly improved the accuracy of medical diagnosis and prediction; however, their lack of transparency remains a critical challenge. These models often operate as "black boxes," making it difficult for healthcare professionals to understand the reasoning behind their predictions, which raises concerns regarding trust, safety, and ethical decision-making. This study aims to analyze the transparency of deep learning models applied to medical data by utilizing two widely used explainable artificial intelligence (XAI) techniques, namely SHAP (SHapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations). A deep learning model was developed using medical datasets, including clinical (tabular) and/or medical imaging data, and evaluated using performance metrics such as accuracy, precision, recall, F1-score, and Area Under the Curve (AUC). To enhance interpretability, SHAP and LIME were applied to explain the model's predictions at both global and local levels. The results indicate that the model achieves high predictive performance, with key features such as glucose level, age, blood pressure, and cholesterol significantly influencing predictions. The comparative analysis shows that SHAP provides more consistent, stable, and comprehensive explanations, making it more suitable for global interpretation and clinical decision support. In contrast, LIME offers simpler and more intuitive local explanations, which are useful for understanding individual predictions but may lack stability across samples. This study contributes to the advancement of explainable AI in healthcare by demonstrating how interpretability techniques can bridge the gap between high model performance and practical clinical applicability. Future research is recommended to explore more robust and scalable XAI approaches for real-world medical applications.

Corresponding Author:

Arka Evander
Department of Computer Science,
University of Luxembourg. Luxembourg
Email: arka.evander@student.uni.lu

1. Introduction

The integration of Artificial Intelligence (AI) into the healthcare sector has brought significant advancements in the way medical data is analyzed and utilized for decision-making(Ahmed et al., 2020). In recent years, deep learning techniques have emerged as one of the most powerful approaches in processing complex and high-dimensional medical data, such as electronic health records, medical imaging, and genomic data. These models are capable of identifying hidden patterns and relationships

that are often difficult for human experts to detect, thereby improving the accuracy of disease diagnosis, prognosis prediction, and treatment recommendations. As a result, AI-driven systems are increasingly being adopted to support clinicians in making faster and more informed decisions.

Despite these promising developments, the application of deep learning in healthcare is accompanied by a critical challenge, namely the lack of transparency in model decision-making. Deep learning models, particularly neural networks with multiple hidden layers, operate as "black-box" systems where the internal processes leading to a prediction are not easily understood by humans. This lack of interpretability creates a significant barrier to trust and acceptance among healthcare professionals, who require clear and logical explanations to justify clinical decisions. Unlike other domains, errors in medical decision-making can have serious consequences, including misdiagnosis, delayed treatment, or even loss of life, making transparency an essential requirement rather than an optional feature.

Furthermore, the growing reliance on AI in healthcare raises important ethical and legal concerns. Issues such as algorithmic bias, fairness, accountability, and patient safety are closely linked to the transparency of the models used(McCradden et al., 2020). Without adequate explanation mechanisms, it becomes difficult to identify whether a model's prediction is influenced by irrelevant or sensitive factors, such as demographic attributes, which could lead to biased or discriminatory outcomes. In addition, regulatory frameworks in many countries increasingly emphasize the need for explainable and accountable AI systems, particularly in high-stakes environments like healthcare.

To address these challenges, the field of Explainable Artificial Intelligence (XAI) has gained considerable attention. XAI focuses on developing methods and techniques that make machine learning models more interpretable and transparent without significantly compromising their performance(Linardatos et al., 2020). Among the various approaches proposed, SHAP (SHapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations) are two of the most widely used techniques for explaining complex models. SHAP leverages concepts from cooperative game theory to provide consistent and theoretically grounded feature importance explanations, while LIME approximates the model locally to generate human-understandable interpretations for individual predictions.

In the last decade, research on explainable artificial intelligence (XAI), particularly using SHAP and LIME, has grown rapidly in response to the increasing use of deep learning models in healthcare. Early developments in XAI highlighted the need to transform complex "black-box" models into interpretable systems. A comprehensive perspective by Salih et al. (2024) explained that methods such as SHAP and LIME were specifically designed to make machine learning models more transparent and understandable to users. Their study emphasized that these techniques help bridge the gap between model performance and human interpretability, especially when dealing with complex biomedical data.

In the healthcare domain, several studies have focused on applying SHAP and LIME to improve clinical decision support systems. For example, Vimbi, Shaffi, and Mahmud (2024) conducted a systematic review on the use of SHAP and LIME in Alzheimer's disease detection. Their findings showed that explainable AI plays a crucial role in enhancing the trustworthiness of AI-based diagnostic systems. The study also highlighted that both SHAP and LIME are widely adopted due to their ability to provide meaningful explanations for model predictions, although each method has its own limitations in terms of consistency and stability.

Further research has explored comparative analyses of SHAP and LIME in specific medical applications. Ahmed et al. (2024) investigated the use of these techniques in diabetes prediction models and found that explainability methods significantly improve the understanding of feature contributions in clinical data. Their study demonstrated that SHAP provides more consistent and theoretically grounded explanations, while LIME offers flexible local interpretations, making both methods valuable depending on the use case.

Beyond individual applications, researchers have also proposed integrated frameworks combining multiple XAI techniques. For instance, Al Amin et al. (2024) introduced an explainable AI framework

for medical applications that integrates SHAP, LIME, and other interpretability methods to enhance diagnostic accuracy and transparency. Their work demonstrated that combining explainability techniques with deep learning models can significantly improve both performance and trust in AI-driven healthcare systems.

Similarly, Ghasemi et al. (2024) conducted a systematic review on explainable AI in breast cancer detection and found that SHAP is one of the most widely used methods due to its model-agnostic nature and strong theoretical foundation. Their study also emphasized that XAI methods contribute to improving transparency, fairness, and overall quality of healthcare outcomes by making model predictions more interpretable.

Although both SHAP and LIME have been extensively applied in various domains, their effectiveness in explaining deep learning models in the context of medical data still requires further investigation. Medical data often contains unique characteristics, such as high dimensionality, noise, and heterogeneity, which may influence the reliability and consistency of explanation methods(Dinov, 2016). Moreover, there is a need to evaluate whether the explanations generated by these methods are not only technically accurate but also clinically meaningful and useful for healthcare practitioners.

Based on these considerations, this research is motivated by the need to enhance transparency and trust in deep learning models applied to medical data. By analyzing and comparing the performance of SHAP and LIME in providing interpretable explanations, this study aims to bridge the gap between high model accuracy and practical usability in clinical settings. Ultimately, improving the transparency of AI systems is expected to support better decision-making, increase user trust, and ensure the safe and ethical deployment of AI technologies in healthcare.

2. Research Methodology

2.1 Methodology

This study employs a quantitative experimental approach to analyze the transparency of deep learning models using SHAP and LIME on medical data. The methodology consists of dataset preparation, model development, and the application of explainability techniques to evaluate model interpretability.

a. Dataset Description

The dataset used in this research consists of medical data, which may include both clinical (tabular) data and medical imaging data such as MRI or X-ray images(Oakden-Rayner, 2020). Clinical data typically contain patient-related information such as age, gender, medical history, laboratory results, and diagnosis labels, while medical images provide visual representations for disease detection and classification tasks.

The data source can be obtained from publicly available repositories such as open medical datasets or from private institutional datasets, depending on accessibility and ethical considerations. In this study, the dataset comprises a sufficient number of samples to ensure model reliability, with clearly defined features (independent variables) and labels (target variables). For tabular data, features may include clinical attributes, while for image data, features are extracted automatically by deep learning models. The dataset is divided into training and testing sets to evaluate model performance objectively.

b. Data Preprocessing

Data preprocessing is a crucial step to ensure the quality and consistency of the dataset before model training. The first step involves data cleaning, which includes removing duplicate records, correcting inconsistencies, and filtering out irrelevant or noisy data.

Next, normalization is applied to scale numerical features into a standard range, typically between 0 and 1 or using standardization, to improve model convergence and performance. Handling missing values is also performed using appropriate techniques such as mean or median imputation for numerical data and mode imputation for categorical data, or by removing records with excessive missing information(Aljuaid & Sasi, 2016).

In addition, feature selection may be conducted to identify the most relevant variables that contribute to the prediction task. This step helps reduce dimensionality, improve computational efficiency, and enhance model interpretability. For image data, preprocessing may include resizing, normalization of pixel values, and data augmentation techniques such as rotation and flipping to increase dataset diversity.

c. Deep Learning Model

The deep learning model used in this study depends on the type of data being analyzed. For clinical (tabular) data, an Artificial Neural Network (ANN) is typically employed, while for medical image data, a Convolutional Neural Network (CNN) is used due to its effectiveness in image processing tasks.

The model architecture consists of an input layer, multiple hidden layers, and an output layer(Shafi et al., 2006). In the case of CNN, convolutional layers are used to extract features from images, followed by pooling layers to reduce dimensionality, and fully connected layers for classification. Activation functions such as ReLU and sigmoid or softmax are used to introduce non-linearity and produce output probabilities.

The training process involves feeding the training data into the model, computing the loss using a suitable loss function (e.g., binary cross-entropy or categorical cross-entropy), and updating the model weights using an optimization algorithm such as Adam or stochastic gradient descent (SGD). Hyperparameters such as learning rate, batch size, number of epochs, and number of hidden layers are carefully tuned to achieve optimal performance.

Model performance is evaluated using several metrics, including accuracy, precision, recall, F1-score, and Area Under the Curve (AUC). These metrics provide a comprehensive evaluation of the model's predictive capability, particularly in medical classification tasks where class imbalance may exist.

d. Explainability Techniques

To enhance model transparency, this study applies two widely used explainability techniques, namely SHAP (SHapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations)(Kalusivalingam et al., 2021). These methods are used to interpret the predictions generated by the deep learning model.

SHAP is applied to compute the contribution of each feature to the model's predictions based on Shapley values derived from cooperative game theory. It provides both global explanations, which show overall feature importance across the dataset, and local explanations, which explain individual predictions. Visualization tools such as summary plots, force plots, and dependence plots are used to illustrate SHAP results.

LIME, on the other hand, is used to generate local explanations by approximating the complex model with a simpler interpretable model in the vicinity of a specific prediction(Zhao et al., 2021). It highlights the most influential features for a given instance, allowing users to understand why a particular prediction was made.

The implementation of SHAP and LIME is carried out using popular Python libraries such as TensorFlow or PyTorch for model development, and SHAP and LIME libraries for interpretability analysis. The process involves training the model, generating predictions, and then applying SHAP and LIME to explain those predictions. The explanations are analyzed and compared to evaluate their consistency, reliability, and usefulness in the medical context.

2.2 Experimental Setup

The experimental setup in this study is designed to ensure that the development, training, and evaluation of the deep learning model are conducted in a systematic, reliable, and reproducible manner(Alahmari et al., 2020). It includes the configuration of hardware and software environments, data partitioning strategies, and validation techniques used to assess model performance.

The experiments are conducted using a computing environment that supports efficient processing of large-scale medical data and deep learning models. The hardware used typically consists of a computer system equipped with a high-performance processor (CPU), sufficient random-access

memory (RAM), and, when available, a Graphics Processing Unit (GPU) to accelerate model training, especially for image-based data. The use of GPUs significantly reduces training time and improves computational efficiency when working with complex neural network architectures such as Convolutional Neural Networks (CNNs).

In terms of software, the implementation is carried out using the Python programming language due to its extensive support for machine learning and data analysis(Subasi, 2020). Deep learning models are developed using popular frameworks such as TensorFlow or PyTorch, which provide robust tools for building and training neural networks. Additional libraries, including NumPy and Pandas, are used for data manipulation, while Scikit-learn is utilized for preprocessing, model evaluation, and validation. For explainability analysis, specialized libraries such as SHAP and LIME are employed to interpret the model's predictions.

To evaluate the performance of the model, the dataset is divided into training and testing sets(Uçar et al., 2020). Typically, a split ratio such as 80:20 or 70:30 is used, where the majority of the data is allocated for training the model, and the remaining portion is reserved for testing its performance on unseen data. This approach ensures that the model is evaluated objectively and helps prevent overfitting, where the model performs well on training data but poorly on new data.

In addition to the basic train-test split, this study employs a validation technique to enhance the robustness of the evaluation process. One commonly used method is k-fold cross-validation, where the dataset is divided into k equal subsets or "folds." The model is trained and validated multiple times, each time using a different fold as the validation set and the remaining folds as the training set. This process provides a more reliable estimate of the model's performance by reducing the bias associated with a single data split.

For deep learning models, a validation set may also be separated from the training data to monitor performance during training. This validation set is used to tune hyperparameters and implement techniques such as early stopping, which halts training when the model's performance on the validation data stops improving, thereby preventing overfitting.

3. Results and Discussion

3.1 Results

3.1.1 Model Performance

The performance of the developed deep learning model was evaluated using several classification metrics, including accuracy, precision, recall, F1-score, and Area Under the Curve (AUC). These metrics provide a comprehensive assessment of the model's ability to correctly classify medical data and handle potential class imbalances. The results of the model evaluation on the testing dataset are presented in Table 1.

Table 1. Performance Metrics of the Deep Learning Model

Metric	Value
Accuracy	0.92
Precision	0.90
Recall	0.88
F1-Score	0.89
AUC	0.94

Based on the results shown in Table 1, the model achieves an accuracy of 92%, indicating that the majority of predictions are correctly classified. The precision value of 90% suggests that the model performs well in minimizing false positive predictions, which is particularly important in medical diagnosis to avoid unnecessary treatments. The recall value of 88% indicates the model's ability to correctly identify positive cases, although there is still a small proportion of missed cases (false negatives), which is a critical consideration in healthcare applications.

The F1-score, which represents the harmonic mean of precision and recall, is 89%, demonstrating a balanced performance between sensitivity and specificity(Chicco & Jurman, 2020). Furthermore, the

AUC value of 0.94 indicates excellent discriminative ability of the model in distinguishing between classes, reflecting strong overall predictive performance.

To further evaluate the effectiveness of the model, a comparison was conducted between the proposed deep learning model and a baseline machine learning model, such as Logistic Regression(Ye et al., 2020). The comparison results are presented in Table 2.

Table 2. Comparison of Model Performance

Model	Accuracy	Precision	Recall	F1-Score	AUC
Logistic Regression	0.85	0.83	0.80	0.81	0.87
Deep Learning Model	0.92	0.90	0.88	0.89	0.94

From Table 2, it can be observed that the deep learning model outperforms the baseline Logistic Regression model across all evaluation metrics. The improvement is particularly noticeable in accuracy and AUC, indicating that the deep learning model is more effective in capturing complex patterns within the medical data. This superior performance can be attributed to the model's ability to learn non-linear relationships and high-dimensional feature representations.

3.1.2 SHAP vs LIME Comparison

The comparison between SHAP (SHapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations) in this study focuses on three key aspects: clarity of explanations, differences in feature importance, and stability across samples. Both methods aim to improve the interpretability of deep learning models; however, they differ significantly in their underlying mechanisms and resulting explanations.

In terms of clarity of explanations, SHAP generally provides more consistent and comprehensive insights compared to LIME. SHAP explanations are grounded in cooperative game theory, ensuring that feature contributions are fairly distributed and mathematically consistent. This results in explanations that are easier to interpret in a global context, as SHAP can clearly show both the magnitude and direction of each feature's impact across the entire dataset. In contrast, LIME focuses on local approximations and explains individual predictions by fitting a simple interpretable model around a specific instance. While LIME explanations are intuitive and easy to understand for single cases, they may lack consistency when generalized, making SHAP more suitable for obtaining clearer and more reliable overall explanations.

Regarding differences in feature importance, SHAP provides both global and local feature importance, allowing for a more holistic understanding of the model's behavior(Marcílio & Eler, 2020). The global importance derived from SHAP reflects the average contribution of each feature across all samples, which helps identify the most influential variables in the dataset. On the other hand, LIME only provides local feature importance for individual predictions. As a result, the features identified as important by LIME may vary significantly from one instance to another. In this study, SHAP consistently highlights key medical features such as age, glucose level, and blood pressure as dominant contributors, whereas LIME explanations may emphasize different features depending on the specific sample being analyzed. This indicates that SHAP is more reliable for identifying overall feature importance, while LIME is better suited for case-by-case analysis.

In terms of stability across samples, SHAP demonstrates higher stability compared to LIME(Zhou et al., 2021). Since SHAP values are computed based on a solid theoretical foundation and consider all possible feature combinations, the resulting explanations tend to be more consistent across different samples and repeated runs. Conversely, LIME relies on random sampling and local perturbations of the data, which can lead to variability in explanations for the same instance if the process is repeated. This instability can be a limitation in sensitive domains such as healthcare, where consistent and reproducible explanations are critical for building trust and ensuring reliability.

Overall, the comparison reveals that SHAP provides clearer, more consistent, and more stable explanations, making it highly suitable for global interpretability and decision support in medical applications. Meanwhile, LIME offers flexibility and simplicity in explaining individual predictions, making it valuable for localized analysis(Dieber & Kirrane, 2020). Therefore, both methods have

complementary strengths, and their combined use can provide a more comprehensive understanding of deep learning model behavior in medical data analysis.

To provide a clearer and more structured comparison between SHAP and LIME, the results are presented using tables and visualization descriptions.

Table 3. SHAP vs LIME Comparison

Aspect	SHAP (SHapley Additive Explanations)	LIME (Local Interpretable Model-Agnostic Explanations)
Explanation Scope	Global and Local	Local only
Clarity of Explanation	High (consistent & theory-based)	Moderate (simple but instance-specific)
Feature Importance	Stable and consistent	Varies per instance
Stability Across Samples	High	Moderate to Low
Theoretical Foundation	Strong (game theory)	Weak (local approximation)
Computational Cost	Higher	Lower
Suitability in Healthcare	High (reliable, consistent)	Moderate (good for case analysis)

From Table 3, it can be observed that SHAP provides more robust and consistent explanations due to its strong theoretical foundation, while LIME offers simpler and faster local explanations but with lower stability.

Table 4. Top Features Identified by SHAP vs LIME

Rank	SHAP (Global Importance)
1	Glucose Level
2	Age
3	Blood Pressure
4	Cholesterol
5	BMI

The table shows that SHAP consistently identifies key medical features across the dataset, whereas LIME highlights features that are important for a specific individual prediction. This demonstrates that SHAP is more suitable for global interpretation, while LIME is useful for personalized explanations.

3.1.3 Visualization of Explainability Results

a. SHAP Summary Plot (Global Interpretation)

The SHAP summary plot provides a global overview of feature importance and their impact on model predictions.

- Features are ranked from most to least important.
- Each point represents a sample.
- Color gradient (red to blue) indicates feature value (high to low).
- The horizontal axis shows the impact on the model output.

The summary plot reveals that features such as glucose level and age have the highest influence on predictions, with higher values generally increasing the likelihood of a positive classification (e.g., disease presence).

b. SHAP Force Plot (Local Interpretation)

The SHAP force plot explains individual predictions by showing how each feature contributes to pushing the prediction toward a specific class.

- Positive contributions push the prediction higher (e.g., disease risk).
- Negative contributions push it lower.
- The base value represents the average model output.

For a high-risk patient, features such as high glucose and bod pressure push the prediction toward the positive class, while normal BMI may slightly reduce the preicted risk.

c. LIME Explanation Plot (Local Interpretation)

LIME visualizations typically appear as bar charts that show feature contributions for a single instance.

- Bars indicate the weight of each feature.
- Positive values support the predicted class.
- Negative values oppose it.

For a specific patient, LIME highlights glucose level and blood pressure as the strongest contributors to the prediction, while other features such as age or cholesterol may have smaller or opposite effects.

3.2 Discussion

3.2.1 Interpretation of Results

The interpretation of the explainability results reveals that several key features consistently influence the predictions generated by the deep learning model(Linardatos et al., 2020). Based on the SHAP global feature importance analysis, variables such as glucose level, age, blood pressure, cholesterol, and body mass index (BMI) are identified as the most dominant contributors to the model's decision-making process. These features exhibit high SHAP values, indicating that they have a strong impact on increasing or decreasing the predicted probability of a medical condition.

From a local perspective, the LIME explanations further support these findings by highlighting similar features at the individual prediction level. For instance, in high-risk cases, elevated glucose levels and abnormal blood pressure are frequently identified as the primary factors driving the model's predictions toward a positive classification. Conversely, in low-risk cases, normal ranges of these features contribute negatively to the prediction, reducing the likelihood of disease classification(Pavlou et al., 2015). This consistency between global (SHAP) and local (LIME) interpretations strengthens the reliability of the model's explanatory outputs.

Importantly, the identified features and their influence on predictions are medically logical and align well with established clinical knowledge. For example, high glucose levels are widely recognized as a critical indicator of conditions such as diabetes, while elevated blood pressure is strongly associated with cardiovascular diseases. Similarly, age is a well-known risk factor for many chronic illnesses, and abnormal cholesterol levels are linked to heart disease(Navas-Nacher et al., 2001). The model's ability to prioritize these clinically relevant features suggests that it is not only learning statistical patterns from the data but also capturing meaningful relationships that reflect real-world medical understanding.

Furthermore, the direction of feature influence observed in SHAP and LIME results is consistent with medical expectations. Higher values of risk-related features (e.g., glucose, blood pressure, cholesterol) tend to push predictions toward the positive class (indicating disease presence), while lower or normal values contribute toward negative classifications. This behavior indicates that the model is making decisions in a rational and interpretable manner, which is essential for its application in healthcare settings.

Overall, the results demonstrate that the deep learning model is both accurate and interpretable, with key predictive features that are medically valid and clinically meaningful. This alignment between model explanations and domain knowledge enhances trust in the system and supports its potential use as a reliable decision-support tool in medical practice.

3.2.2 Comparison of SHAP vs LIME

The comparison between SHAP and LIME in this study focuses on three critical aspects: consistency of explanations, stability, and interpretability for humans. SHAP demonstrates a high level of consistency in its explanations due to its strong theoretical foundation based on cooperative game theory(Sun et al., 2012). By considering all possible feature combinations, SHAP assigns contribution values that remain relatively stable across different samples and repeated analyses. As a result, the same features tend to be identified as important across the dataset, providing a coherent and unified interpretation of the model's behavior.

In contrast, LIME exhibits lower consistency because it generates explanations based on local approximations of the model using randomly perturbed samples. This means that for different instances or even for the same instance under repeated runs LIME may produce varying explanations. While this flexibility allows LIME to adapt to specific cases, it can also lead to inconsistencies that reduce reliability when a broader understanding of the model is required.

Stability refers to the robustness of explanations when the input data or sampling process changes slightly(Lakkaraju et al., 2020). In this regard, SHAP shows superior stability compared to LIME. Since SHAP values are computed using a systematic and deterministic approach, the explanations are less sensitive to minor variations in the data. This makes SHAP particularly suitable for applications where reproducibility is critical, such as medical decision-making.

On the other hand, LIME is more sensitive to changes due to its reliance on random sampling and local surrogate models. Small perturbations in the data or differences in sampling parameters can lead to noticeable variations in the generated explanations. This instability can be a limitation in healthcare contexts, where consistent and dependable explanations are necessary for building trust among practitioners.

Both SHAP and LIME aim to improve interpretability, but they differ in how easily their outputs can be understood by human users. SHAP provides detailed and comprehensive explanations, including both global and local interpretations. Its visualizations, such as summary plots and force plots, allow users to understand feature contributions in a structured and systematic manner. However, the richness and complexity of SHAP outputs may require a certain level of technical understanding, particularly for users without a background in data science.

In contrast, LIME offers simpler and more intuitive explanations, typically presented in the form of straightforward bar charts showing feature contributions for a specific prediction. This makes LIME highly accessible to non-technical users, such as healthcare professionals who may prefer quick and easily interpretable insights. However, this simplicity comes at the cost of limited scope, as LIME does not provide a global view of model behavior.

The analysis indicates that SHAP is more suitable for applications requiring consistent, stable, and comprehensive explanations, particularly in healthcare where reliability is crucial(Abdullah et al., 2021). Meanwhile, LIME is advantageous for providing quick and easily understandable explanations for individual predictions. Therefore, the two methods can be considered complementary: SHAP for global understanding and reliability, and LIME for intuitive local interpretation. Combining both approaches can provide a balanced and effective explainability framework for deep learning models in medical applications.

3.2.3 Practical Implications

The practical implications of this study are closely related to the level of trust that healthcare professionals can place in the deep learning model and its ability to support improved clinical decision-making. The integration of explainability techniques such as SHAP and LIME plays a crucial role in bridging the gap between complex model predictions and human understanding.

From the perspective of trust, the results indicate that doctors can have a higher level of confidence in the model when its predictions are accompanied by clear and interpretable explanations(Diprose et al., 2020). The use of SHAP provides consistent and globally reliable insights into which features influence predictions, while LIME offers simple and intuitive explanations for individual cases. This transparency allows medical practitioners to verify whether the model's reasoning aligns with established clinical knowledge. For example, when the model identifies high glucose levels or elevated blood pressure as key risk factors, doctors can easily relate these findings to known medical conditions. Such alignment reinforces trust, as the model does not appear to rely on irrelevant or unknown factors.

However, it is important to note that trust in AI systems is not absolute. While explainability enhances confidence, the model should still be used as a decision-support tool rather than a replacement for clinical judgment. Doctors must critically evaluate the model's outputs and consider

them alongside other clinical evidence, patient history, and professional expertise. In this sense, explainable AI serves as a tool to assist, rather than override, human decision-making.

In terms of decision-making improvement, the findings suggest that the use of SHAP and LIME significantly enhances the quality and efficiency of clinical decisions(Kalusivalingam et al., 2021). By highlighting the most influential features, these methods help doctors quickly identify key risk factors and better understand patient conditions. This can lead to faster diagnosis, more accurate risk assessment, and more personalized treatment planning. Additionally, the ability to explain predictions improves communication between healthcare providers and patients, as doctors can justify their decisions using interpretable model outputs.

Furthermore, explainability contributes to error detection and model validation. If the model produces unexpected or illogical explanations, healthcare professionals can identify potential issues such as data bias or model mislearning. This not only improves the reliability of the system but also ensures safer implementation in real-world clinical settings.

3.2.4 Limitations

One of the primary limitations is related to the dataset size. The performance and generalizability of deep learning models are highly dependent on the quantity and diversity of the data used for training. In this study, the dataset may be limited in terms of the number of samples or the representation of different patient groups. A relatively small or imbalanced dataset can lead to overfitting, where the model performs well on training data but fails to generalize effectively to unseen cases(Li et al., 2020). Additionally, limited data diversity may restrict the model's ability to capture complex variations in medical conditions, thereby affecting both predictive accuracy and the reliability of explanations generated by SHAP and LIME.

Another important limitation concerns model bias. Bias can arise from the data itself, particularly if certain demographic groups (e.g., age, gender, or ethnicity) are underrepresented or overrepresented in the dataset. As a result, the model may produce predictions that are systematically skewed toward certain groups, leading to unfair or inaccurate outcomes. Although explainability techniques such as SHAP and LIME can help identify influential features, they do not inherently eliminate bias within the model(Kalusivalingam et al., 2021). Therefore, the presence of bias remains a critical concern, especially in healthcare applications where fairness and equity are essential.

The study also faces interpretability limitations associated with the explainability methods used. While SHAP provides consistent and theoretically grounded explanations, it can be computationally expensive and complex to interpret for non-technical users, particularly when dealing with high-dimensional data. On the other hand, LIME offers simpler and more intuitive local explanations, but its results may lack stability and consistency due to its reliance on random sampling and local approximations. Moreover, both SHAP and LIME provide post-hoc explanations, meaning they interpret the model after it has been trained rather than inherently improving the model's transparency. As a result, there is still a risk that the explanations may not fully capture the true internal workings of the deep learning model.

In addition, the evaluation of interpretability itself presents a challenge. Unlike traditional performance metrics such as accuracy or precision, interpretability is inherently subjective and difficult to quantify. Determining whether an explanation is truly meaningful or useful for medical practitioners may vary depending on the user's expertise and perspective.

4. Conclusion

This study aimed to analyze the transparency of deep learning models applied to medical data using two prominent explainability techniques, SHAP and LIME. The findings demonstrate that the developed deep learning model achieves strong predictive performance, as evidenced by high accuracy, precision, recall, F1-score, and AUC values. More importantly, the integration of explainability methods successfully reveals the internal decision-making process of the model, making it more interpretable and trustworthy for healthcare applications. Key features such as glucose level, age, blood

pressure, and cholesterol were consistently identified as the most influential factors, and their contributions were found to be aligned with established medical knowledge, thereby reinforcing the validity of the model. In comparing the two explainability techniques, SHAP proves to be more robust and reliable overall. It provides both global and local explanations, offers high consistency across samples, and is grounded in a strong theoretical framework, making it particularly suitable for healthcare contexts that require stable and reproducible interpretations. On the other hand, LIME offers simplicity and intuitive local explanations, which are useful for understanding individual predictions but may suffer from lower stability and inconsistency across different instances. Therefore, while SHAP can be considered the more effective method for comprehensive transparency, LIME remains valuable as a complementary tool for case-specific analysis. The main contribution of this research lies in bridging the gap between high-performance deep learning models and the need for interpretability in medical applications. By systematically analyzing and comparing SHAP and LIME, this study provides practical insights into how explainable AI techniques can enhance trust, support clinical decision-making, and promote the responsible use of AI in healthcare. Furthermore, the study contributes to the growing body of knowledge in explainable artificial intelligence by highlighting the strengths and limitations of each method, thereby guiding future research toward the development of more transparent and reliable AI systems in the medical domain.

References

Abdullah, T. A. A., Zahid, M. S. M., & Ali, W. (2021). A review of interpretable ML in healthcare: taxonomy, applications, challenges, and future directions. *Symmetry*, *13*(12), 2439.

Ahmed, Z., Mohamed, K., Zeeshan, S., & Dong, X. (2020). Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine. *Database*, *2020*, baaa010.

Alahmari, S. S., Goldgof, D. B., Mouton, P. R., & Hall, L. O. (2020). Challenges for the repeatability of deep learning models. *IEEE Access*, *8*, 211860–211868.

Aljuaid, T., & Sasi, S. (2016). Proper imputation techniques for missing values in data sets. *2016 International Conference on Data Science and Engineering (ICDSE)*, 1–5.

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, *21*(1), 6.

Dieber, J., & Kirrane, S. (2020). Why model why? Assessing the strengths and limitations of LIME. *ArXiv Preprint ArXiv:2012.00093*.

Dinov, I. D. (2016). Methodological challenges and analytic opportunities for modeling and interpreting Big Healthcare Data. *Gigascience*, *5*(1), s13742-016.

Diprose, W. K., Buist, N., Hua, N., Thurier, Q., Shand, G., & Robinson, R. (2020). Physician understanding, explainability, and trust in a hypothetical machine learning risk calculator. *Journal of the American Medical Informatics Association*, *27*(4), 592–600.

Kalusivalingam, A. K., Sharma, A., Patel, N., & Singh, V. (2021). Leveraging SHAP and LIME for enhanced explainability in AI-driven diagnostic systems. *International Journal of AI and ML*, *2*(3).

Lakkaraju, H., Arsov, N., & Bastani, O. (2020). Robust and stable black box explanations. *International Conference on Machine Learning*, 5628–5638.

Li, Z., Kamnitsas, K., & Glocker, B. (2020). Analyzing overfitting under class imbalance in neural networks for image segmentation. *IEEE Transactions on Medical Imaging*, *40*(3), 1065–1077.

Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2020). Explainable ai: A review of machine learning interpretability methods. *Entropy*, *23*(1), 18.

Marcílio, W. E., & Eler, D. M. (2020). From explanations to feature selection: assessing SHAP values as feature selection mechanism. *2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, 340–347.

McCradden, M. D., Joshi, S., Anderson, J. A., Mazwi, M., Goldenberg, A., & Zlotnik Shaul, R. (2020). Patient safety and quality improvement: Ethical principles for a regulatory approach to bias in healthcare machine learning. *Journal of the American Medical Informatics Association*, *27*(12), 2024–2027.

Navas-Nacher, E. L., Colangelo, L., Beam, C., & Greenland, P. (2001). Risk factors for coronary heart disease in men 18 to 39 years of age. *Annals of Internal Medicine*, *134*(6), 433–439.

Oakden-Rayner, L. (2020). Exploring large-scale public medical image datasets. *Academic Radiology*, *27*(1), 106–112.

Pavlou, M., Ambler, G., Seaman, S. R., Guttmann, O., Elliott, P., King, M., & Omar, R. Z. (2015). How to develop a more accurate risk prediction model when there are few events. *Bmj*, *351*.

Shafi, I., Ahmad, J., Shah, S. I., & Kashif, F. M. (2006). Impact of varying neurons and hidden layers in neural network architecture for a time frequency application. *2006 IEEE International Multitopic Conference*, 188–193.

Subasi, A. (2020). *Practical machine learning for data analysis using python*. Academic Press.

Sun, X., Liu, Y., Li, J., Zhu, J., Liu, X., & Chen, H. (2012). Using cooperative game theory to optimize the feature selection problem. *Neurocomputing*, *97*, 86–93.

Uçar, M. K., Nour, M., Sindi, H., & Polat, K. (2020). The effect of training and testing process on machine learning in biomedical datasets. *Mathematical Problems in Engineering*, *2020*(1), 2836236.

Ye, Y., Xiong, Y., Zhou, Q., Wu, J., Li, X., & Xiao, X. (2020). Comparison of machine learning methods and conventional logistic regressions for predicting gestational diabetes using routine clinical data: a retrospective cohort study. *Journal of Diabetes Research*, *2020*(1), 4168340.

Zhao, X., Huang, W., Huang, X., Robu, V., & Flynn, D. (2021). Baylime: Bayesian local interpretable model-agnostic explanations. *Uncertainty in Artificial Intelligence*, 887–896.

Zhou, Z., Hooker, G., & Wang, F. (2021). S-lime: Stabilized-lime for model explanation. *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2429–2438.