ITJ | The International Tax Journal

A LIME-based approach for explainable AI in healthcare

Tejashree Moharekar [1], Dr. Amol B. Patil [2], and **Dr. Vidyullata S. Jadhav [3]**

[1] Assistant Professor, Yashwantrao Chavan School of Rural Development, Shivaji University, Kolhapur, India, Email: moharekartejashree@gmail.com
[2] V. P. Institute of Management Studies & Research, Sangli, India, Email: abpatil@vpimsr.edu.in
[3] V. P. Institute of Management Studies & Research, Sangli, India, Email: vsjadhav@vpimsr.edu.in

Abstract---The current study focuses on the use of LIME in diabetes prediction as a multiclass classification problem, where patients are classified into three categories: No Diabetes, Pre-Diabetes, and Diabetes. The analysis demonstrates how LIME elucidates the importance of features such as Body Mass Index (BMI), age, physical health, and lifestyle factors in determining risk categories. By providing transparent explanations for predictions, LIME enhances trust in AI systems and supports medical practitioners in interpreting model outputs. Challenges in applying LIME to multiclass healthcare datasets, such as computational overhead and explanation reliability, are also discussed. This research underscores the role of LIME in enabling ethical, interpretable, and effective AI solutions for diabetes prediction.

Keywords---Explainable Artificial Intelligence (XAI), AI, LIME, Diabetes Prediction, Multiclass Classification, Machine Learning.

Introduction

Timely prediction and intervention play a pivotal role in alleviating the burden of diabetes, particularly in identifying those at risk of progressing to Pre-Diabetes or Diabetes, and distinguishing them from individuals with no diabetes. Machine learning (ML) models have emerged as powerful tools in healthcare, enabling the prediction and classification of diabetes risk by analyzing various patient data, including Body Mass Index (BMI), age, physical health status, and lifestyle habits (Alghamdi et al., 2021). Explainable Artificial Intelligence, often known as XAI, is a sort of artificial intelligence that is capable of explaining to humans the reasoning behind a decision or a prediction that it has made. When it comes

How to Cite:
Moharekar, T., Dr. Amol B. Patil, & Dr. Vidyullata S. Jadhav. (2025). A LIME-based approach for explainable AI in healthcare. *The International Tax Journal*, *52*(5), 1732–1741. Retrieved from https://internationaltaxjournal.online/index.php/itj/article/view/186

The International tax journal ISSN: 0097-7314 E-ISSN: 3066-2370 © 2025

Submitted: 27 July 2025 | Revised: 09 August 2025 | Accepted: 13 September 2025

to crucial duties like security, healthcare, or money, the objective of XAI is to make artificial intelligence systems more open, trustworthy, and accountable (Vidyullata S Jadhav, 2023).

Multiclass classification in diabetes prediction involves categorizing patients into three distinct groups: No Diabetes, Pre-Diabetes, and Diabetes. This method helps identify individuals who could benefit from preventative measures or need immediate clinical intervention. Although advanced machine learning models such as random forests, gradient-boosted trees, and neural networks demonstrate high accuracy in performing these tasks, their lack of interpretability creates concerns regarding their reliability and acceptance in clinical settings (Zhou et al., 2022).

LIME has proven effective in interpreting complex machine learning models across various domains, including healthcare, computer vision, and natural language processing. It enhances confidence in these models by providing clearer insights and improving understanding, while also helping to uncover potential biases or errors in the decision-making process (Tallaswapna, 2024).

The authors of the paper aim to make a meaningful contribution to the expanding field of Explainable AI (XAI) in healthcare, offering valuable insights for researchers, practitioners, and decision-makers within the healthcare sector. In conclusion, they also analyze the effectiveness of various XAI methods when applied to medical healthcare systems (Shahab S Band, 2023).

The review highlights the widespread use of local explanation techniques, notably SHAP and LIME, with SHAP emerging as the preferred method due to its stability and mathematical guarantees. However, it identifies a significant gap in how XAI results are assessed, pointing out that many studies depend on anecdotal evidence or expert opinions rather than solid quantitative metrics. This emphasizes the pressing need for standardized evaluation frameworks to ensure the reliability and effectiveness of XAI applications in practice. (Saarela, 2024).

As artificial intelligence (AI) techniques continue to evolve, becoming more computationally efficient and integrated into our daily lives, there is an increasing demand to unravel the complexities within black-box AI models. Popular machine learning and deep learning methods, while powerful, often lack transparency. This calls for a deeper understanding and more detailed explanations to clarify the inner workings and decision-making processes of these models (Mrutyunjaya Panda, 2023).

Explainable Artificial Intelligence (XAI) techniques aim to tackle the challenge of understanding complex machine learning models by offering insights into their decision-making processes. One such framework, Local Interpretable Model-Agnostic Explanations (LIME), provides human-readable explanations for individual predictions. LIME works by approximating the behavior of a machine learning model in a localized manner, creating simpler surrogate models that are easier to interpret, while still capturing the key factors driving the model's predictions. This approach allows users to gain a clearer understanding of how specific features contribute to a given prediction. (Ribeiro et al., 2016).

LIME approximates the behavior of complex models locally by fitting an interpretable surrogate model around the prediction of interest, enabling medical practitioners to identify the most significant features influencing a given classification. LIME has been successfully applied in medical diagnostics to interpret model predictions regarding disease classification or diagnosis. For example, in the prediction of diseases such as cancer, diabetes, or heart disease, LIME provides insights into which features (e.g., patient age, blood pressure, cholesterol levels) contributed most to the model's diagnosis.

In a study by Caruana et al. (2015), LIME was used to explain predictions made by machine learning models for pneumonia risk, improving the interpretability of model outputs and increasing the confidence of clinicians in using automated tools for diagnosis. By using LIME to explain these

predictions, healthcare professionals can better understand how factors like patient age, medical history, and comorbidities affect outcomes.

Ribeiro et al. (2016) demonstrated the use of LIME for explaining the survival predictions of patients with heart disease, where the local explanations helped clinicians identify key factors influencing survival probabilities and led to better clinical decision-making. LIME is also employed to provide interpretability in systems that suggest personalized treatment plans.

In a study by Chen et al. (2018), LIME was used to explain the rationale behind treatment recommendations made by machine learning models, helping healthcare professionals understand how specific patient features influenced treatment choices. This transparency fosters collaboration between clinicians and AI systems, ensuring that the recommendations align with the patient's unique needs and clinical context. Clinical Decision Support Systems (CDSS) are increasingly powered by machine learning models. LIME has been incorporated into these systems to offer interpretable explanations for predictions related to disease risk, medication dosages, and treatment options. By providing detailed explanations for decisions made by the model, LIME enhances trust in these automated systems.

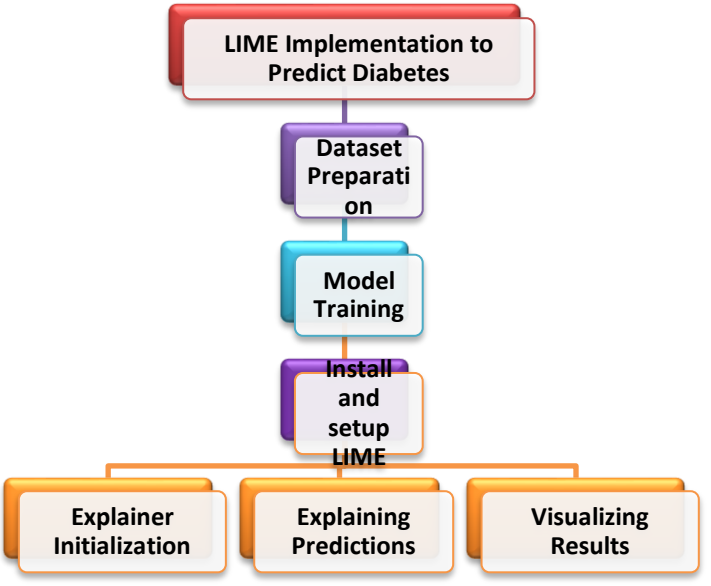
Doshi-Velez and Kim (2017) discussed how LIME could be applied in clinical settings, helping physicians understand the rationale behind model-generated decisions and improving the adoption of AI-driven CDSS in healthcare environments. By uncovering these biases, LIME helps ensure that healthcare models are fairer and more equitable.

Ribeiro et al. (2016) highlighted how LIME could be used to detect and mitigate bias in predictive models, ensuring that the models do not disproportionately disadvantage certain demographic groups. LIME enables better human-AI collaboration by allowing clinicians to understand and trust AI-based decisions. It serves as a tool for clinicians to verify model predictions while combining them with their own expertise to make better-informed decisions for patient care (Ribeiro et al., 2016).

In the context of diabetes prediction, LIME plays a pivotal role by helping clinicians understand why a patient has been classified into a specific risk category. For instance, LIME can explain whether a combination of high BMI and sedentary lifestyle contributes to a Pre-Diabetes classification or whether genetic predisposition and poor physical health indicate a high risk of Diabetes. By providing these detailed explanations, LIME fosters trust in AI-driven healthcare systems and bridges the interpretability gap between AI models and domain experts. This paper focuses on applying LIME to diabetes prediction as a multiclass classification problem, emphasizing its ability to:

- Illuminate feature contributions for individual predictions across No Diabetes, Pre-Diabetes, and Diabetes categories.
- Enhance transparency in predictive modelling, enabling healthcare providers to make informed decisions.

Steps for Lime Implementation



The dataset includes a variety of health, lifestyle, and demographic features designed to predict and analyze diabetes status. The target variable, Diabetes_012, categorizes individuals into three groups: 0 for no diabetes, 1 for pre-diabetes, and 2 for diabetes. Before using LIME, the dataset needs to be pre-processed:

- Handle missing values (if any).
- Encode categorical features (e.g., Sex, Age, Education) using techniques like one-hot encoding or label encoding.
- Normalize or scale continuous features (e.g., BMI, MentHlth, PhysHlth) to ensure they are on the same scale. Several health-related variables provide insights into the individual's medical history and conditions.

Here's the dataset explained in a tabular format:

Feature	Type	Description	Values
Diabetes_012	Categorical	Diabetes status of the individual.	0: No Diabetes, 1: Pre-Diabetes, 2: Diabetes
HighBP	Binary	Whether the individual has high blood pressure.	0: No, 1: Yes
HighChol	Binary	Whether the individual has high cholesterol.	0: No, 1: Yes
CholCheck	Binary	Whether the individual had a cholesterol check in the past five years.	0: No, 1: Yes
BMI	Continuous	Body Mass Index, a measure of body fat based on height and weight.	Numerical values
Smoker	Binary	Whether the individual is a smoker.	0: No, 1: Yes
Stroke	Binary	Whether the individual has ever had a stroke.	0: No, 1: Yes
HeartDiseaseorAttack	Binary	History of heart disease or heart attack.	0: No, 1: Yes
PhysActivity	Binary	Whether the individual engages in regular physical activity.	0: No, 1: Yes
Fruits	Binary	Frequency of fruit consumption.	0: No, 1: Yes
Veggies	Binary	Frequency of vegetable consumption.	0: No, 1: Yes
HvyAlcoholConsump	Binary	Whether the individual engages in heavy alcohol consumption.	0: No, 1: Yes
AnyHealthcare	Binary	Whether the individual has access to any form of healthcare.	0: No, 1: Yes
NoDocbcCost	Binary	Whether the cost prevented the individual from seeing a doctor.	0: No, 1: Yes
GenHlth	Categorical	Self-reported general health on a scale of 1–5.	1: Excellent, 2: Very Good, 3: Good, 4: Fair, 5: Poor
MentHlth	Continuous	Number of days with poor mental health in the past 30 days.	Numerical values (0–30)
PhysHlth	Continuous	Number of days with poor physical health in the past 30 days.	Numerical values (0–30)
DiffWalk	Binary	Whether the individual has difficulty walking.	0: No, 1: Yes
Sex	Binary	Gender of the individual.	0: Female, 1: Male
Age	Categorical	Age of the individual grouped into categories.	1: 18–24, 2: 25–29, 3: 30–34, ..., 13: 80 or older
Education	Categorical	Level of education attained by the individual.	1: Never attended school, 2: Elementary, ..., 6: College graduate
Income	Categorical	Income categories representing annual household income.	1: Less than $10,000, ..., 8: $75,000 or more

Source: Compiled by Researcher

A Random Forest classifier was chosen for building the model on the diabetes dataset due to its strong predictive performance, robustness, and ability to handle complex relationships in data. Random forests are ensemble models that consist of multiple decision trees and are known for their ability to generalize well, even when the dataset has a mix of categorical and numerical features. They are less prone to overfitting compared to individual decision trees, making them ideal for this kind of healthcare-related dataset, where complexity and noisy data are often present. In terms of interpretability, while random forests generally perform well, they are considered "black-box" models because it's difficult to understand the decision-making process behind their predictions.

```
Classification Report:
              precision    recall  f1-score   support

         0.0       0.86      0.97      0.91     42795
         1.0       0.00      0.00      0.00       944
         2.0       0.47      0.20      0.28      6997

    accuracy                           0.84     50736
   macro avg       0.44      0.39      0.40     50736
weighted avg       0.79      0.84      0.81     50736
```

The classification report shows that the model performs well for the majority class (0.0), achieving high precision (0.86) and recall (0.97), resulting in a strong F1-score of 0.91. However, the model performs poorly for the minority class 1.0, with both precision and recall at 0.00, leading to an F1-score of 0.00, indicating that it fails to identify any instances of this class. For class 2.0, the model shows moderate performance, with a precision of 0.47 and a recall of 0.20, resulting in a low F1-score of 0.28. The random forest model achieves an accuracy of 0.84, which is largely driven by the correct identification of the majority class. The macro average scores, which treat all classes equally, show weak performance with an F1-score of 0.40, reflecting the imbalance between classes.

Initialize Lime

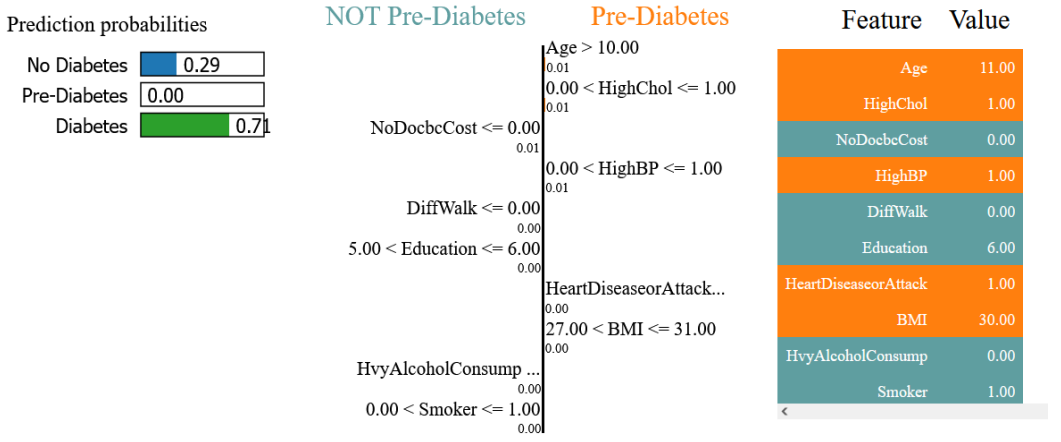
To initialize the LIME Tabular Explainer for a Random Forest model on the diabetes dataset, the process begins with importing necessary libraries, including LIME's LimeTabularExplainer and other machine learning tools from sklearn. The diabetes dataset is loaded and prepared, with features and labels extracted for training the model. A Random Forest classifier is trained on this data to make predictions, leveraging its strength in handling complex datasets.

The LIME explainer is then initialized, using the training data, model, and feature names to set up the explanation process. The explainer helps provide interpretable, local explanations for individual predictions made by the Random Forest model. By applying LIME, we can understand which features most influenced specific predictions, making the model more transparent and helping to build trust, particularly in fields like healthcare where interpretability is crucial. The explainer can be used to generate visual explanations of predictions, highlighting key features that contributed to the model's decision-making for each instance, ultimately improving the model's understandability and trustworthiness.

Result and Discussions

```
# Explain a specific instance
instance_index = 25 # Example: explaining the 25th test instance
explanation = explainer.explain_instance(
    data_row=X_test.iloc[instance_index].values,
    predict_fn=clf.predict_proba
```

1738



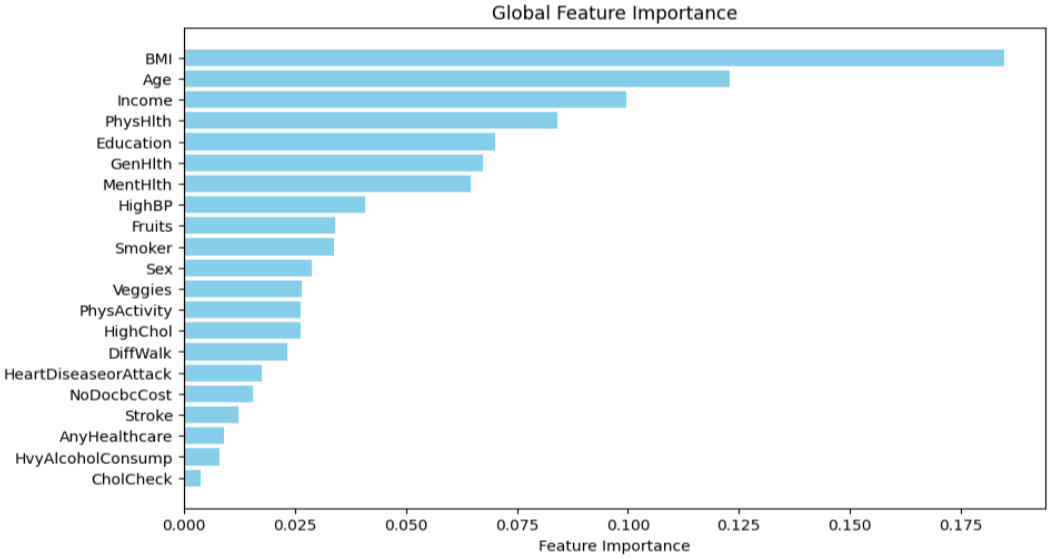
LIME visualization provides a breakdown of a predictive model's decision for determining the likelihood of diabetes. The prediction probabilities for 25th test instance indicates that the model strongly predicts the individual to have diabetes with a probability of 0.71, compared to 0.29 for "No Diabetes" and 0.00 for "Pre-Diabetes."

The middle panel highlights the key features contributing to the model's classification. For the "NOT Pre-Diabetes" class, factors like the absence of difficulty walking (DiffWalk = 0), higher education level (Education > 5), and no history of heavy alcohol consumption (HvyAlcoholConsump = 0) weigh against the prediction of pre-diabetes. Conversely, the "Pre-Diabetes" segment includes features such as age (Age > 10), high cholesterol (HighChol = 1), high blood pressure (HighBP = 1), and BMI values between 27 and 31 as relevant indicators.

The right panel outlines the feature values for this specific individual. Notably, the individual has risk factors such as Age = 11, HighChol = 1, HighBP = 1, BMI = 30, and smoking status (Smoker = 1). These features are critical contributors to the diabetes prediction.

The model leverages these inputs to identify that the individual is at high risk for diabetes, with lifestyle and health factors like high BMI, smoking, and existing medical conditions (high cholesterol and blood pressure) playing significant roles in the prediction.

Feature Importance



The bar chart displays the global feature importance derived from a machine learning model, indicating how each feature contributes to the overall prediction of the target variable. BMI is the most critical feature, suggesting that body mass index plays the most significant role in the model's decision-making process. Age ranks second, indicating that age is a strong predictor, likely due to the increased risk of diabetes and related conditions with age. Income and Physical Health (PhysHlth) follow, implying socioeconomic factors and self-reported physical health significantly affect predictions. Education and General Health (GenHlth) are moderately important, reflecting the impact of overall health awareness and lifestyle.

Mental Health (MentHlth) and High Blood Pressure (HighBP) also have considerable influence, highlighting the importance of mental well-being and hypertension in diabetes risk. Lifestyle factors like fruit consumption (Fruits), smoking status (Smoker), and vegetable intake (Veggies) also contribute but are less influential globally. High Cholesterol (HighChol) and Physical Activity (PhysActivity) show minor importance but are still relevant for model predictions. Variables like Healthcare Access (AnyHealthcare), Stroke, and Heavy Alcohol Consumption (HvyAlcoholConsump) have minimal importance, suggesting they are less predictive globally. The model places the most weight on BMI, Age, Income, and Physical Health, aligning with known risk factors for diabetes. While lifestyle and other health indicators are still important, they play a relatively smaller role in the overall prediction.

Challenges in applying Lime in Healthcare

Applying LIME to multiclass healthcare datasets presents several challenges. First, healthcare datasets often involve imbalanced classes, where certain conditions or outcomes are underrepresented. LIME's explanations may be less reliable for minority classes due to insufficient data representation, leading to potential biases in feature importance. Second, healthcare datasets typically include a mix of numerical, categorical, and often complex features (e.g., medical test results or patient histories), which require careful pre-processing to ensure meaningful explanations. Additionally, LIME generates local explanations for individual predictions, which can become computationally expensive and time-consuming in large multiclass datasets, especially when multiple instances need to be analysed. Furthermore, in multiclass settings, the explanations need to address multiple potential outcomes, which may lead to confusion or reduced interpretability for non-technical stakeholders, such as clinicians, when

comparing feature contributions across classes. Finally, ensuring that the generated explanations align with clinical knowledge and do not lead to incorrect interpretations is critical, as errors in healthcare predictions can have significant consequences.

Computationally, LIME generates local explanations by perturbing data and fitting interpretable models for individual predictions. In multiclass healthcare datasets, this process becomes increasingly resource-intensive due to the higher number of classes and the need to generate explanations for each possible outcome. This overhead can be prohibitive, especially when dealing with large datasets or when explanations are required for many instances.

Explanation reliability is another significant challenge. In multiclass scenarios, LIME provides class-specific explanations, which can lead to inconsistencies or ambiguities when feature contributions vary significantly across classes. This is particularly problematic in healthcare, where datasets often suffer from class imbalance, and the minority classes may lack sufficient data to produce accurate and reliable explanations. Additionally, the complexity of healthcare features, such as medical test results or clinical observations, can make it difficult for LIME's perturbation-based approach to capture nuanced relationships accurately, potentially leading to misleading explanations. These challenges underscore the importance of optimizing LIME's implementation and validating its outputs in the context of multiclass healthcare datasets to ensure meaningful and reliable insights.

Conclusion

In conclusion, using Random Forest for predicting diabetes provides a robust and interpretable model that works well for handling tabular datasets with mixed feature types. By integrating the LIME framework, the black-box nature of Random Forest is mitigated, offering clear, instance-level explanations for predictions. LIME allows us to understand the contribution of individual features, such as age, cholesterol levels, blood pressure, and smoking habits, toward the prediction of diabetes or non-diabetes. The combination of Random Forest's accuracy and LIME's explanatory power makes it an effective approach for predictive modeling and decision-making in diabetes risk assessment.

Acknowledgments

The authors express their sincere gratitude to V. P. Institute of Management Studies and Research, Sangli for providing the necessary support and research facilities for this study. The authors also acknowledge the use of the **Scopus® database** for accessing scholarly literature and conducting bibliometric analysis. The authors gratefully acknowledge the **UCI Machine Learning Repository** for providing open access to the dataset used in this study.

Author Contribution

Conceptualization, Methodology Dr. Amol B. Patil.; Dataset analysis,and Lime implementation, writing-reviewing and editing: Dr. Vidyullata S. Jadhav.

Data Availability

The dataset used in this study is publicly available from the UCI Machine Learning Repository and can be accessed at: https://archive.ics.uci.edu/.

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] Alghamdi, M., & Zaki, M. (2021). Predicting diabetes using machine learning techniques: A comprehensive review. Journal of Healthcare Engineering, 2021, 1-12. https://doi.org/10.1155/2021/8534669

[2] Caruana, R., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1721-1730. https://doi.org/10.1145/2783258.2788613

[3] Chen, X., Zhang, J., & Song, L. (2018). Towards interpretable machine learning for personalized treatment recommendations. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 459-468. https://doi.org/10.1145/3219819.3220083

[4] Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608. https://arxiv.org/abs/1702.08608

[5] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you? Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135-1144. https://doi.org/10.1145/2939672.2939778

[6] Zhou, J., Zhang, Z., & Huang, X. (2022). Interpretability in machine learning: A review and future directions for medical diagnosis. Journal of Medical Systems, 46(4), 43. https://doi.org/10.1007/s10916-022-01857-9

[7] Mrutyunjaya Panda, S. R. (2023). Explainable artificial intelligence for Healthcare applications using Random Forest Classifier with LIME and SHAP. In B. T. Seetha, Transparent, Interpretable and Explainable AI Systems. CRC Press.

[8] Saarela, M. &. (2024). Recent Applications of Explainable AI (XAI): A Systematic Literature Review. Applied Sciences. Applied Sciences.

[9] Shahab S Band, A. Y.-C.-W. (2023). Application of explainable artificial intelligence in medical health: A systematic review of interpretability methods. Informatics in Medicine Unlocked.

[10] Tallaswapna. (2024). LIME(Local Interpretable Model-Agnostic Explanations) in XAI with an example in Python. Retrieved from Medium.

[11] Vidyullata S Jadhav, T. T. (2023). UNPACKING EXPLAINABLE AI (XAI) IN EDUCATION: A COMPREHENSIVE REVIEW AND OUTLOOK. International Research Journal of Humanities and Interdisciplinary Studies .