Agrawal *et al. Diagnostic Pathology* (2025) 20:105
https://doi.org/10.1186/s13000-025-01686-3

Diagnostic Pathology

RESEARCH Open Access

Fostering trust and interpretability: integrating explainable AI (XAI) with machine learning for enhanced disease prediction and decision transparency

Renuka Agrawal[1*], Tawishi Gupta[1], Shaurya Gupta[1], Sakshi Chauhan[1], Prisha Patel[1] and Safa Hamdare[2]

Abstract

Medical healthcare has advanced substantially due to advancements in Artificial Intelligence (AI) techniques for early disease detection alongside support for clinical decisions. However, a gap exists in widespread adoption of results of these algorithms by public due to black box nature of models. The undisclosed nature of these systems creates fundamental obstacles within medical sectors that handle crucial cases because medical practitioners needs to understand the reasoning behind the outcome of a particular disease. A hybrid Machine Learning (ML) framework integrating Explainable AI (XAI) strategies that will improve both predictive performance and interpretability is explored in proposed work. The system leverages Decision Trees, Naive Bayes, Random Forests and XGBoost algorithms to predict the medical condition risks of Diabetes, Anaemia, Thalassemia, Heart Disease, Thrombocytopenia within its framework. SHAP (SHapley Additive exPlanations) together with LIME (Local Interpretable Model-agnostic Explanations) adds functionality to the proposed system by displaying important features contributing to each prediction. The framework upholds an accuracy of 99.2% besides the ability to provide understandable explanations for interpretation of model outputs. The performance combined with interpretability from the framework enables clinical practitioners to make decisions through an understanding of AI-generated outputs thereby reducing distrust in AI-driven healthcare.

Highlights

- The proposed manuscript suggests the development of a hybrid Model with machine learning Techniques and includes interpretability behind outcomes: ML-XAI framework for predicting diseases.
- Five Diseases are considered in the proposed work: Diabetes, Anaemia, Thalassemia, Heart Disease, and Thrombocytopenia.
- Black Box nature of Most Machine Learning models leads to distrust and acceptance about results especially in medical domain.
- The proposed model solves this issue by diagnosing disease with ML models and providing attributes which are responsible for disease being diagnosed by models.

*Correspondence:
Renuka Agrawal
renuka.agrawal@sitpune.edu.in

Full list of author information is available at the end of the article

Keywords Explainable artificial intelligence (XAI), Machine learning (ML), Healthcare prediction, Local interpretable model agnostic explanations (LIME), SHapley additive exPlanations (SHAP), XGBoost, Random forest

Introduction

In the evolving landscape of modern medicine, AI is emerging as a transformative force, reshaping the way healthcare is delivered. These models enable early disease detection and personalized recommendations, addressing challenges such as the shortage of medical practitioners, especially in urban settings [1, 2]. AI powered diagnostic tools analyze vast datasets from medical records, imaging, and patient histories, uncovering patterns that may not be readily apparent to human practitioners. This integration of AI into healthcare has paved the way for more efficient, accurate, and proactive care [3].

Despite their precision, traditional ML models, especially deep neural networks and other complicated models face a critical limitation: their "black box" nature. This lack of transparency is a major concern in healthcare, where understanding *why* a diagnosis has been made is as important as the decision itself. The opacity of predictions of ML models is a major concern in high-stakes domains like healthcare [4, 5]. While these models excel at making accurate predictions, they often fail to explain their decision-making process, hindering trust and acceptance. This is referred to as the black box problem, where the internal workings of the model are not visible or interpretable, making it difficult for clinicians to understand or validate the results. Explainability ensures clinicians can validate the model's logic and communicate findings effectively to patients, fostering trust and informed decision-making. Existing research has highlighted significant gaps in the explainability of ML models used for disease prediction and diagnosis [6, 7]. Many high-performance models prioritize accuracy over interoperability, leaving clinicians and patients in the dark about how decisions are made [8, 9]. Scholars have emphasized the need for models that balance predictive performance with interoperability [10, 11]. Studies have also noted the underutilization of XAI techniques, such as SHAP and LIME, in current diagnostic systems. This gap underscores the need for frameworks that integrate explainability into ML models without compromising accuracy, thereby promoting trust and actionable insights in clinical contexts. The demand for interpretability poses a major bottleneck for the wide adoption of AI/ML in healthcare when decision making for the well-being of patients requires reliability and accountability.

The manuscript aims to address these challenges by developing a hybrid ML-XAI framework for predicting diseases along with providing reasoning for predictions. This is done by integrating XAI models with outcome of predictions from ML models. The proposed system combines the predictive accuracy of advanced ensemble models like Random Forests and XGBoost besides others. It incorporates XAI techniques such as LIME and SHAP to provide transparent and actionable insights into the decision-making process. Using health metrics derived from blood tests and lifestyle factors, the system predicts the risk of five diseases: Diabetes, Anemia, Thalassemia, Heart Disease, and Thrombocytopenia. The proposed work aims to integrate interpretable AI into clinical workflows, empowering practitioners to make informed decisions and enhancing patient understanding. This paper outlines the design, implementation, and evaluation of the system, demonstrating its practical application in bridging the gap between accuracy and interpretability in medical AI tools. The structure of this paper is as follows:

- Section Literature Review presents a comprehensive literature review, providing an overview of related work and identifying research gaps.
- Section Proposed system methodology outlines the design of the proposed system, detailing its data flow using system block diagram. It describes the methodology employed, encompassing data collection, pre-processing, analysis, handling of data imbalance, and the system's application. Also the section discusses the integration of the proposed system with XAI techniques.
- Section Results showcases the results, including performance evaluation, the effects of XAI integration, and the impact of dataset balancing on model accuracy.
- Section Conclusion concludes the study, summarizing key findings and offering directions for future research.

Table 1 below lists the acronyms used in the paper, showing the common terms referenced throughout the paper.

Literature review

XAI has become a pivotal tool in healthcare, offering transparency and trust in AI systems used for disease diagnosis. Biswas et al. [2] emphasized model-agnostic methods like LIME and SHAP, model-specific techniques such as CNN visualizations, and rule-based approaches for enhancing interpretability. Similarly, Band et al. [12] systematically reviewed XAI methods across healthcare datasets, highlighting their strengths in transparency but noting challenges like resource intensity and data biases.

Table 1 List of acronyms used

Sr.No.	Full Form	Acronym
1	Artificial Intelligence	AI
2	Electronic Healthcare Records	EHR
3	Machine Learning	ML
4	Explainable Artificial Intelligence	XAI
5	SHapley Additive exPlanations	SHAP
6	Local Interpretable Model-agnostic Explanations	LIME
7	eXtreme Gradient Boosting	XGBoost
8	Exploratory Data Analysis	EDA
9	Synthetic Minority Oversampling Technique	SMOTE
10	Convolutional Neural Network	CNN
11	Cardiovascular Disease	CVD
12	Naive Bayes	NB
13	Multilayer Perceptron	MLP
14	Random Forest	RF
16	Red Blood Cells	RBC
17	White Blood Cells	WBC

Guleria et al. [13] introduced an XAI frame work for cardiovascular disease prediction, achieving high accuracy with ensemble methods but facing limitations due to small datasets and resource demands. Amann et al. [14] provided a multidisciplinary perspective on XAI, advocating alignment with ethical principles but identifying unresolved legal challenges in prediction transparency. Similarly, Magesh et al. [15] employed LIME with transfer learning for Parkinson's disease detection, achieving notable accuracy but facing class imbalance and reliance on image quality Sheu and Pardeshi [16] stressed human interaction in XAI and demonstrated the effectiveness of Grad-CAM and SHAP while calling for consistent scoring systems for broader acceptance.

Several studies demonstrated the application of ML with XAI in specific diseases. For instance, Gabbay et al. [17] proposed an XAI-based model for COVID-19 severity prediction, showcasing real-time applicability despite dataset constraints. Dehghani and Yazdanparast [6] explored symptom associations in COVID-19 patients but lacked causal interpretation and symptom progression tracking. Efforts in multi-disease prediction have also been noticeable. Gaurav et al. [18] proposed a framework for disease prediction using real-life parameters, achieving promising results but with limited dataset adaptability.

After reviewing the relevant literature, it is evident that the integration of advanced technologies plays a critical role in enhancing the performance and interpretability of complex systems. LIME, as shown in the works of different researchers [19, 20] is a model-agnostic technique that improves the interpretability of ML models. In the proposed model, LIME will be utilized to provide local explanations for the model's predictions, allowing for a deeper understanding of the decision-making process. This will be particularly beneficial in applications where transparency is crucial, such as disease diagnosis or energy management, enabling users to comprehend the factors influencing the model's output. Table 2 shows the tabular representation of work done by researchers in

Table 2 Work done by different researchers

Ref No.	Disease Detected	Dataset Used	Technology/Approach	Key Outcomes
[6]	COVID 19	COVID-19 records from 2,875 patients in three hospitals. 34 symptoms, with key ones as apnea, cough, fever, and CVD.	Frequency-based feature selection with set thresholds. Apriori algorithm for symptom-outcome association.	Symptom Associations: Recovery: fever, apnea, cough. Death: apnea, weakness, CVD, ventilator use. No causality, limited provider trust, lacks symptom progression tracking.
[12]	Cardiovascular Disease	Cleveland Heart Disease dataset (303 instances, 14 features)	Used SVM, KNN, AdaBoost, Gaussian Naive Bayes for heart disease prediction. Applied XAI for feature selection and model weight optimization.	Achieved 82.5% accuracy with SVM. Enhanced interpretability for clinical decision-making. Small dataset and limited attributes reduce robustness. Reliance on a single dataset affects generalizability.
[13]	Parkinson's Disease	642 DaTSCAN SPECT images (430 PD, 212 non-PD	VGG16 CNN with transfer learning for classification. LIME for visual explanations of image influencing decisions.	Accuracy: 95.2%, Sensitivity: 97.5%, Specificity: 90.9%. Aids early PD diagnosis and clinical decision-making. Class imbalance. Limited generalizability and dependence on image quality.
[14]	General Disease	General medical datasets including pneumonia, BSI, AKI, and ICU data	Explain ability approaches like LIME, SHAP, Grad-CAM,. Evaluation using AUROC and sensitivity analysis	Resource-intensive, with concerns over legal and ethical uncertainty. Risk of bias in data, impacting model decisions.
[15]	COVID-19	COVID-19 dataset (50,000 + patients, from May to October 2020).	Used MLP and Random Forest for severity prediction (high, medium, low). Integrated LIME for improved model interpretability.	80% accuracy with both MLP and RF. Real-time assessments available via mobile and web apps. Dataset limited to a specific region and time period. Performance variability in medium severity cases, potential overfitting
[18]	Review on Healthcare	150 Articles in Healthcare and interpretability Models.	LIME, Transfer Learning	Systematic review of 53 articles, categorizing XAI methods like SHAP, LIME, and Grad-CAM. Discusses applications to diseases like brain tumors, COVID-19, and chronic kidney disease.

the field of medical disease diagnosis using ML and DL techniques.

Proposed system methodology

The primary goal of this system is to diagnose blood-related diseases using various health indicators while offering clear explanations through XAI techniques. The steps of proposed methodology as illustrated in Fig. 1 includes data collection, preprocessing, EDA, Balancing, splitting in train -test, examining ML models for performance and system integration with XAI for interpretability [21]. These steps are pictorially represented in Fig. 1.

Data collection (Blood test report)

The data analyzed in this research was drawn from Kaggle and comprised of blood samples containing 25 health-related attributes. These attributes included parameters such as hemoglobin, platelets, glucose, cholesterol, red blood cells, white blood cells, and other biochemical indices. Each sample was associated with specific illness categories, providing a strong foundation for forecasting. To make sure it satisfied quality requirements for analysis, the dataset was first examined for consistency, duplication, and completeness [17]. Tabular representation of attributes considered for disease diagnosis is shown in Table 3.

Data pre-processing

Data cleaning is an essential first step in the data preprocessing phase, where duplicates are removed and missing values are addressed to ensure the dataset is accurate and consistent. Once the dataset is clean, standardization is applied to all numerical features using the StandardScaler technique. This process transforms the data to have a mean of zero and a standard deviation of one, ensuring that each feature contributes equally to the model and preventing any potential bias during the training phase.

In the preprocessing phase, the dataset underwent a comprehensive cleaning process, including the removal of duplicate entries and handling of missing values to ensure consistency and reliability [22]. Quantitative health indicators such as glucose, hemoglobin, cholesterol, platelet count, and other blood-related biomarkers were standardized using the standardscaler from sklearn, aligning them to a common scale with zero mean and unit variance. This step was essential to prevent any single feature from disproportionately influencing model performance. Additionally, exploratory data analysis *(later part of the paper)* revealed severe class imbalance among disease categories, which could potentially skew model predictions towards the dominant class.

Exploratory data analysis (EDA)

EDA is conducted to profile the feature matrix, examine feature relationships, analyze distributions, and search for outliers. The insights gained from EDA help in handling class imbalance and feature selection for the modeling process [23]. EDA was conducted to estimate the dataset's structure, identify outliers, and examine variable dependencies.

i. Disease Distribution Analysis: The diagram shown in Fig. 2 above demonstrates the significant imbalance in the dataset, underscoring the importance of using SMOTE to tackle this issue. By applying SMOTE, one can generate synthetic samples for the minority class, helping to create a more balanced distribution. This process is essential for improving model performance and ensuring it doesn't favor the majority class, thereby enhancing the overall accuracy and fairness of the predictions [24]. This step enhanced model performance by reducing bias towards the majority class and improving prediction quality.
ii. Correlation Matrix: The correlation matrix as shown in Fig. 3 highlighted significant relationships between variables:

From correlation matrix one can infer that systolic and diastolic blood pressures showed high inter-correlation, while biomarkers like C-reactive protein and cholesterol exhibited low R-values, indicating independence.

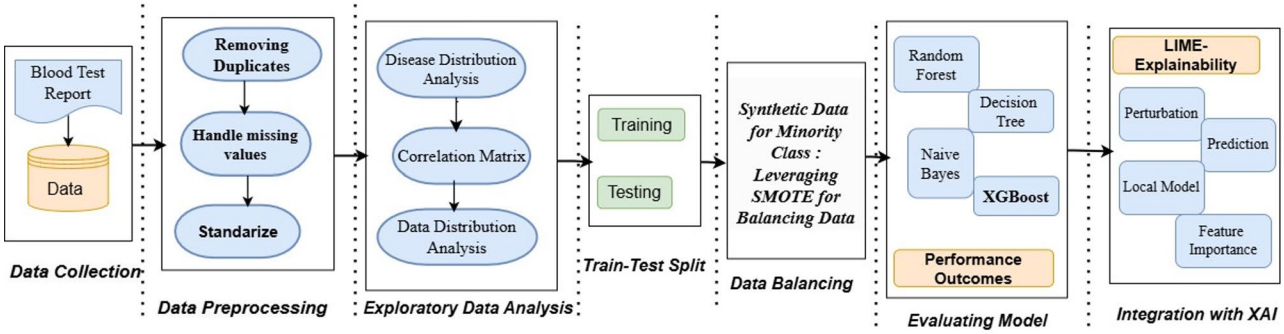

Fig. 1 Flow of methodology for healthcare prediction system

Table 3 Dataset description

S.no	Attribute Name	Data Type, Description and Normal Range
1	Glucose	Numeric This measures Blood Sugar Levels, significant for diagnosing Diabetes. Normal Range 70–140 mg/dL
2	Cholesterol	Numeric: Measures total Cholesterol in Blood. High levels increase risk of cardiovascular disease. Desirable Range 125-200-mg/dL
3	Hemoglobin	Numeric: Represents protein in RBC that carries oxygen. Low value indicates Anemia, whereas high signifies dehydration. Range 13.5–17.5 g/dl
4	Platelets	Numeric: Helps in Clotting of Blood. Low level (thrombocytopenia) causing increase in risk of bleeding, while high level (thrombocytosis) causes clot formation. Range 150,000-450,000 per µL
5	White Blood Cells	Numeric: Vital for immune system Range 4000–11,000 per mm^3
6	Red Blood Cells	Numeric: Main function is to carry oxygen throughout the body. Abnormal levels indicate anemia. Range 4.2–5.4 million/ µL.
7	Hematocrit	Numeric: This indicates percentage of RBC in blood. Range: 38–52%
8	Mean Corpuscular Volume	Numeric: Indicates a measure of average size of RBC. Low value indicates iron deficiency. Range:80–100 fL.
9	Mean Corpuscular Hemoglobin	Numeric: This shows average hemoglobin content per RBC. Range 27–33 pg
10	Mean Corpuscular Hemoglobin Concentration	Numeric: Indicates Hemoglobin concentration in RBC. Range 32–36 g/dL.
11	Insulin	Numeric: Information about the hormone that regulates blood glucose levels. Range 5–25 µU/mL
12	BMI	Numeric: Body Mass Index, analyses weight status of a person. Higher values indicates overweight and vice versa. Range 18.5–24.9 kg/m^2
13	Systolic Blood Pressure	Numeric: Informs about pressure in arteries during heartbeats, high values indicates hypertension Range 90–120 mmHg.
14	Diastolic Blood Pressure	Numeric: Informs about pressure in arteries between heartbeats, low value suggests shock or dehydration Range: 60–80 mmHg.
15	Triglycerides	Numeric: Indicates whether a person has fat in the blood or not. High level indicates risk of cardio disease. Range 50–150 mg/dL
16	HbA1c	Numeric: Measures average blood sugar over 2–3 months, where values > 6.5% indicates diabetes. Range 4–6%
17	LDL Cholesterol	Numeric: Indicates amount of 'Bad' cholesterol that causes artery blockage. Range 70-130 mg/dL
18	HDL Cholesterol	Numeric: Indicates amount of 'Good' cholesterol that removes excess cholesterol from blood. Low level is undesirable as it might cause heart disease. Range 40-60 mg/dL
19	ALT	Numeric: Alanine Aminotransferase, a liver enzyme level, whose high value indicates liver damage. Range 10–40 U/L
20	AHT	Numeric: Aspartate Aminotransferase, another liver enzyme level, whose high value indicates liver disease or muscle damage. Range 10–40 U/L
21	Heart Rate	Numeric: Number of heartbeats per minute. Both extreme low and high are undesirable. Range: 60-100 bpm.
22	Creatinine	Numeric: Indicates level of waste product filtered by kidneys, where high level indicates abnormality in kidney functioning. Range 0.6–1.2 mg/dL
23	Troponin	Numeric: Indicates protein in heart muscle, where slight elevations indicate cardiac damage. Range: 0-0.04 ng/mL
24	C-Reactive Protein	Numeric: Marker of inflammation. High level undesirable. Range: 0–3 mg/L

Clusters of similar biomarkers (e.g., glucose, HbA1c, and triglycerides) reflected their association with metabolic health.

iii. Disease vs. Health Indicators: Boxplots Fig. 4 illustrate variations in health parameters across diseases. These variations aligned with expected indicators for each disease. This figure shows the distribution of key health metrics across different diseases. Diabetes shows a broader range of glucose levels, with higher median values compared to other diseases, suggesting glucose as a critical factor in diabetes diagnosis [25]. Cholesterol levels vary, with Heart Disease and Diabetes displaying higher distributions, as expected, given the role of cholesterol in cardiovascular health. Anemia has a distinct spread in hemoglobin levels, with slightly elevated medians, which aligns with its clinical markers. Thrombocytopenia shows significantly lower platelet levels, a defining feature of this disease. No extreme deviations are visible across diseases, but Thalassemia and Heart Disease show a wider spread in creatinine levels.

Training and testing

The dataset is split into training and testing sets, with the training set further divided for cross-validation.

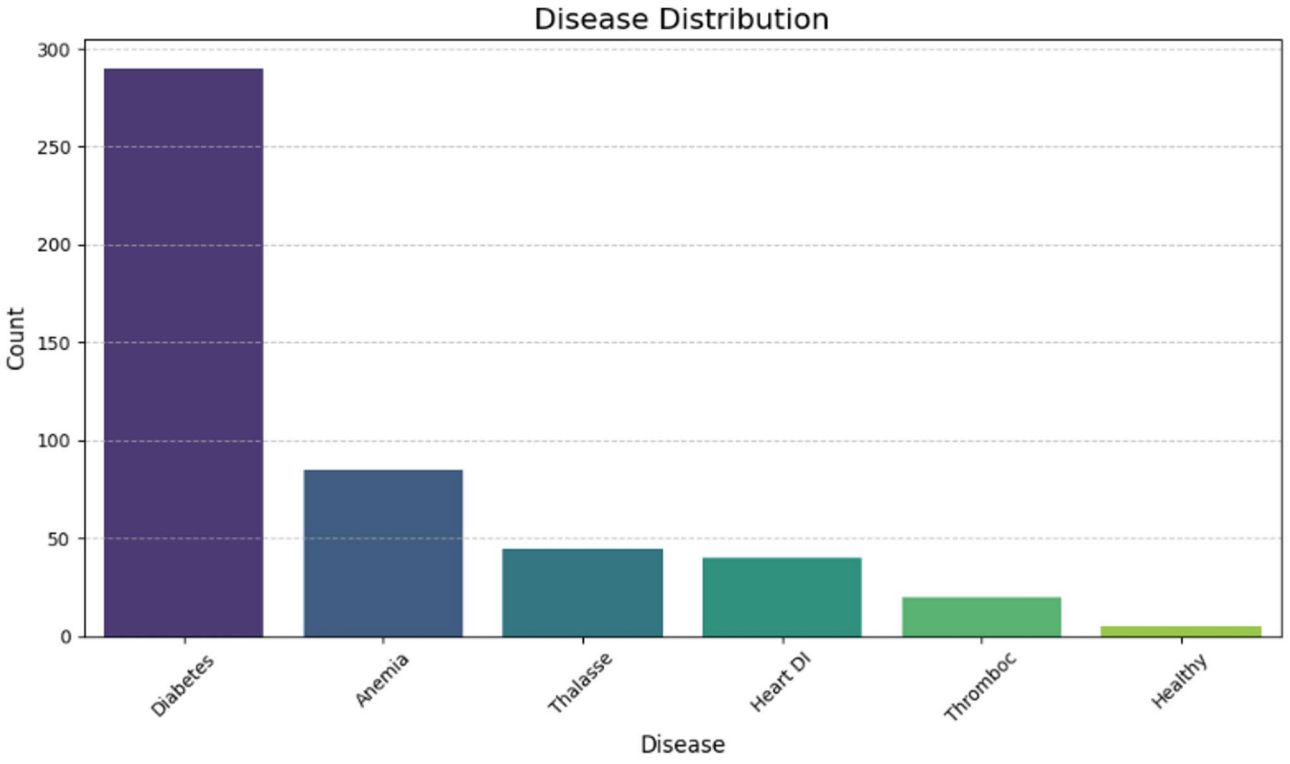

Fig. 2 Disease distribution analysis

- Training: ML models such as Decision Tree, Naive Bayes, Random Forest, and XGBoost are used to train the system on the balanced dataset.
- Testing: The trained models are evaluated on the test set to assess performance on unseen data.

Different permutation of train-test split is done before selecting model and the combination of split which gives best results [26]. In the later stage, when SMOTE is leveraged to balance data, only training samples are balanced and test data remains unchanged.

Handling class imbalance using SMOTE

In the process of analyzing the first part of the dataset, it was observed that there is an appreciative class imbalance for most of the diseases in the current dataset. Some classes, like the Healthy and Thrombocytopenia categories, had far fewer instances than diseases like Diabetes. This difference could lead to model bias, causing a decrease in the total prediction rate for less-represented categories [27]. This issue was resolved with the help of SMOTE. SMOTE works by augmenting the minority classes to generate new samples between existing ones while maintaining the statistical properties of the original dataset [28, 29]. Initially, there was a significant disparity in the number of cases per disease category as shown in Table 4. In the dataset, Diabetes had the highest incidence at 294 cases, whereas Healthy had the least, with only 5 cases. This imbalance caused the model to predict more frequently for diseases such as *Diabetes* and less frequently for rare occurrence class as *Healthy*.

After applying SMOTE, each category was increased to the same number of instances as the largest category, 294 cases, as shown in Table 5. This up-sampling ensured that all categories of illnesses were well-represented in the dataset used to train the model.

This augmentation technique was pivotal in enhancing model generalization and fairness without compromising the original data integrity.

Model evaluation and selection

Upon training multiple machine learning models on the structured healthcare dataset, the selection of the most suitable model was based on a combination of performance metrics—primarily accuracy—and the assessment of potential overfitting. A range of classification algorithms commonly used for structured data was considered, including Decision Trees, Naïve Bayes, Random Forests, and XGBoost. Each model was evaluated using cross-validation and tested on unseen data to ensure generalizability. Among these, XGBoost (Extreme Gradient Boosting) emerged as the best-performing model. Its superior accuracy, robustness against overfitting, and ability to handle complex, nonlinear relationships made it particularly well-suited for the dataset used in this study. XGBoost's regularization techniques, efficient handling

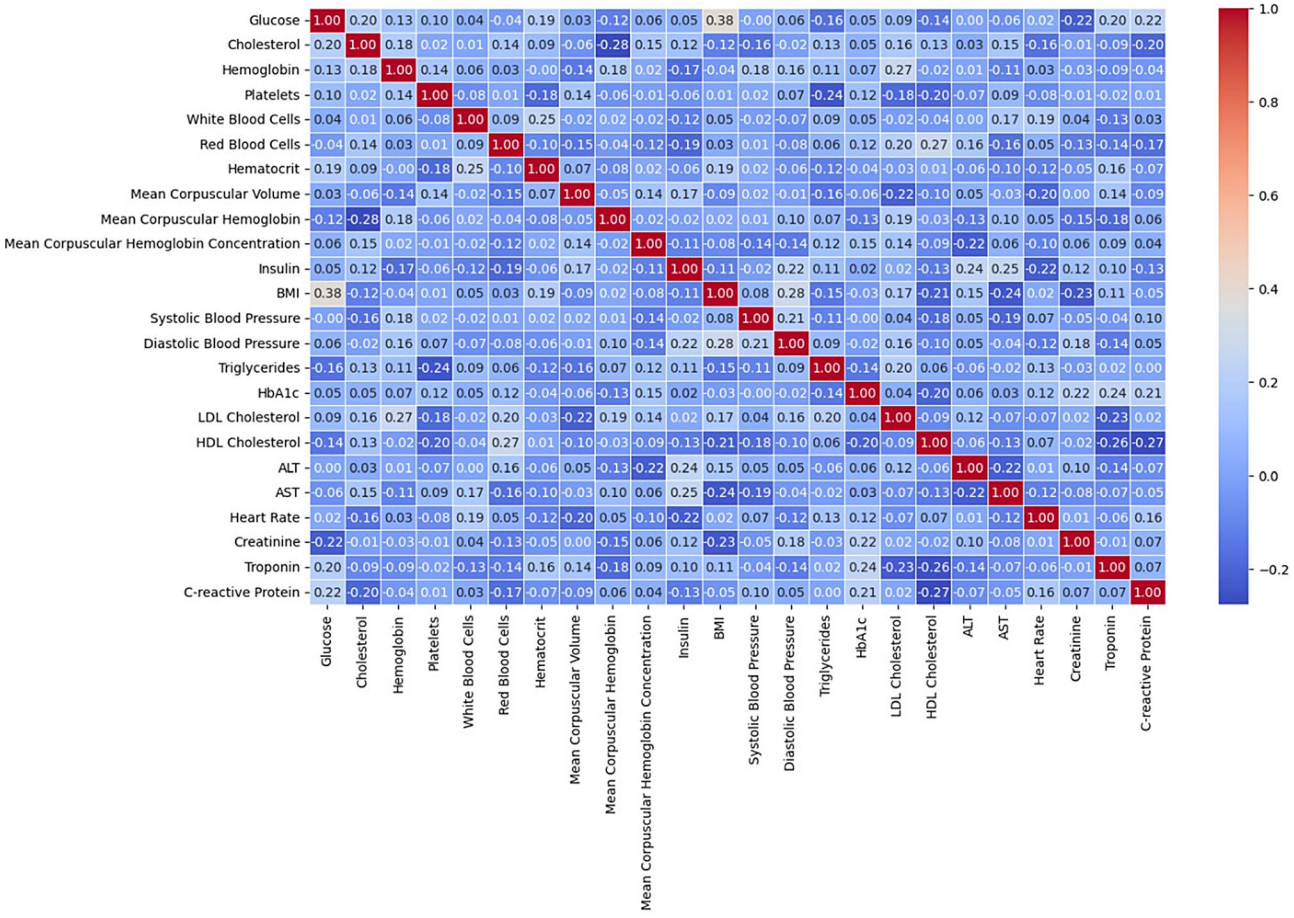

Fig. 3 Correlation matrix

of missing data, and scalability further contributed to its selection as the optimal model for predictive analysis in this healthcare application.

System integration with XAI

LIME is used to make the model's predictions interpretable. - LIME generates perturbations (slightly modified instances of the original data) to analyze the model's behavior. A locally accurate, interpretable model (usually a linear regression) is derived to explain the predictions. The system outputs:

- Feature Importance: The most important features contributing to the prediction.
- Prediction: The final prediction made by the model for a given instance.
- Local Model: A simplified, locally accurate model around the current data point.

A significant challenge addressed by XAI techniques is understanding the manner and rationale behind specific results, which is crucial in critical sectors such as medical, financial, and legal fields, where decisions must be transparent and well-supported. XAI technique, LIME, provides local interpretability for ML models, making it particularly useful in sectors like healthcare for explaining individual predictions and ensuring transparency. Thus, the subsection explains LIME's process, applications, and why it was preferred over SHAP for local interpretability and efficiency in healthcare.

Overview of LIME

LIME is an XAI technique that explains any ML model by approximating the model's behavior locally around an instance using a simpler interpretable model, such as a linear model [30]. It focuses on the local behavior of the model in the vicinity of a specific instance and generates explanations for the features that contributed most to the prediction.

- Process: LIME operates by modifying the input data to create samples similar to the target instance and using the black-box model to predict outcomes for these samples. It then builds an interpretable model around these predictions in the neighborhood of the instance.

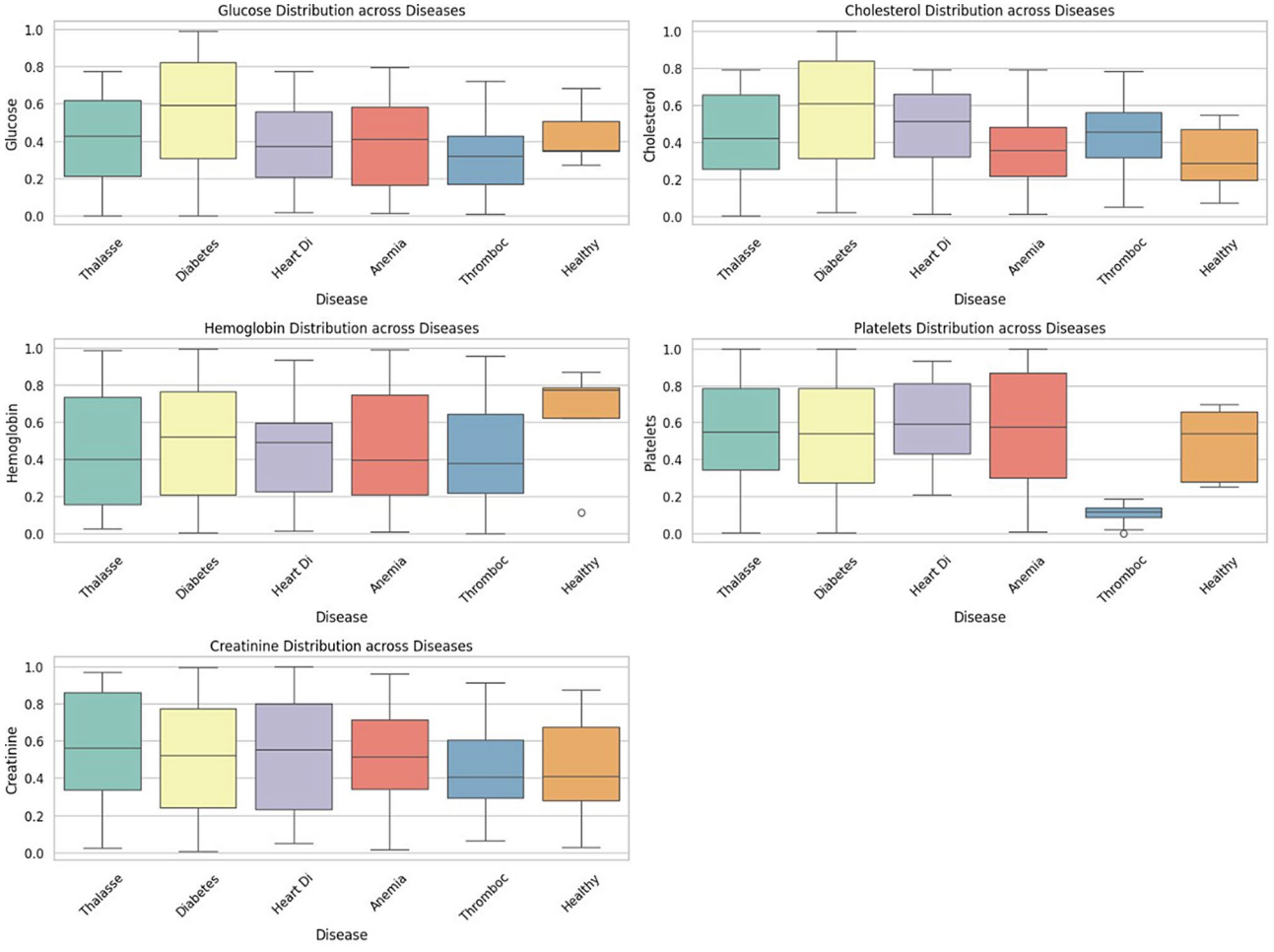

Fig. 4 Health feature boxplots by diseases

Table 4 Initial distribution of classes in dataset

Disease	Instances
Diabetes	294
Anemia	58
Thalassemia	43
Heart Disease	27
Thrombocytopenia	16
Healthy	5

Table 5 Disease distribution after applying SMOTE

Disease	Instances
Diabetes	294
Anemia	294
Thalassemia	294
Heart Disease	294
Thrombocytopenia	294
Healthy	294

- Interpretability: LIME provides explanations for a specific classification or quantity prediction around the instance, highlighting the importance of features proximal to the instance.
- Applications: LIME is widely applied in domains where interpretability of local predictions is essential, such as Loan approval, Medical diagnosis and Recommendation systems.

Model interpretability: LIME over SHAP

XAI techniques, such as LIME and SHAP, are essential in healthcare for ensuring transparency, reliability, and ethical deployment of ML models. While both methods aim to enhance model interpretability, the choice of LIME in this project was motivated by specific considerations relevant to the dataset and healthcare domain.

- Transparency & Ethics: XAI tools like LIME and SHAP ensure trustworthy, ethical use of ML in healthcare.
- Local Interpretability: LIME explains individual predictions using local surrogate models, ideal for patient-specific insights. SHAP offers broader but less personalized interpretations.

Table 6 Accuracy of ML models across splits

Model	80:20 Split	70:30 Split	65:35 Split
XGBoost	**95.92%**	93.84%	87.72%
Random Forest	86.73%	85.62%	81.87%
Decision Tree	65.31%	52.74%	53.80%
Naive Bayes	74.49%	71.23%	66.08%

Table 7 Performance of ML models across various parameters

Model	Accuracy	Precision	Recall	F1 Score
XGBoost	**0.9592**	0.95	0.90	0.91
Random Forest	0.8673	0.94	0.81	0.77
Decision Tree	0.6531	0.72	0.68	0.67
Naive Bayes	0.7449	0.81	0.76	0.78

- Efficiency: LIME is computationally efficient, using sampling and simpler models. SHAP, requiring Shapley values, is resource-intensive.
- Handling Imbalanced Data: LIME adapts well to imbalanced datasets via instance perturbation. SHAP's dataset-wide reliance may introduce bias.
- Human Readability: LIME offers intuitive, clear explanations, helping healthcare professionals understand model decisions.
- Model Flexibility: Being model-agnostic, LIME works across ML algorithms without major adjustments.
- Clinical Alignment: LIME's focus on key features in individual cases aligns with clinical reasoning, supporting actionable insights in diagnoses.
- Domain-Specific Prioritization: In healthcare, explanations must align with clinical reasoning to ensure trust and usability. LIME's ability to isolate and highlight the influence of key features in individual cases makes it ideal for understanding individual patient attributes in diagnosis.

Results

The results section discusses ML model performance across various indices, LIME integration for interpretability, and discussions of key findings.

Performance evaluation

The study evaluated four ML models: Random Forest, Decision Tree, Naive Bayes, and XGBoost, for predicting disease risks based on health metrics. The performance of each model was measured in terms of accuracy across different train-test splits, as shown in Table 6. XGBoost model achieved the best accuracy of prediction of selected disease with the Train: Test split ratio of 80:20. Further Train: test split of 80:20 was selected for comparison of other performance indices across ML models.

Besides Accuracy, other performance indices of the selected ML models have also been considered. These are precision, Recall and F1 Score besides accuracy. Tabular representation of the performance parameters of the models are shown in Table 7. Again, XGBoost comes out be outperform other models selected for predicting disease.

Further 5-fold cross validation and Hyperparameter tuning was also done to check for ML model performance. Error Bar plots of the results were also considered for analysis as shown in Fig. 5. Following interpretations can be made for ML models after 5-fold cross Validation and Hyperparameter tuning the performances obtained.

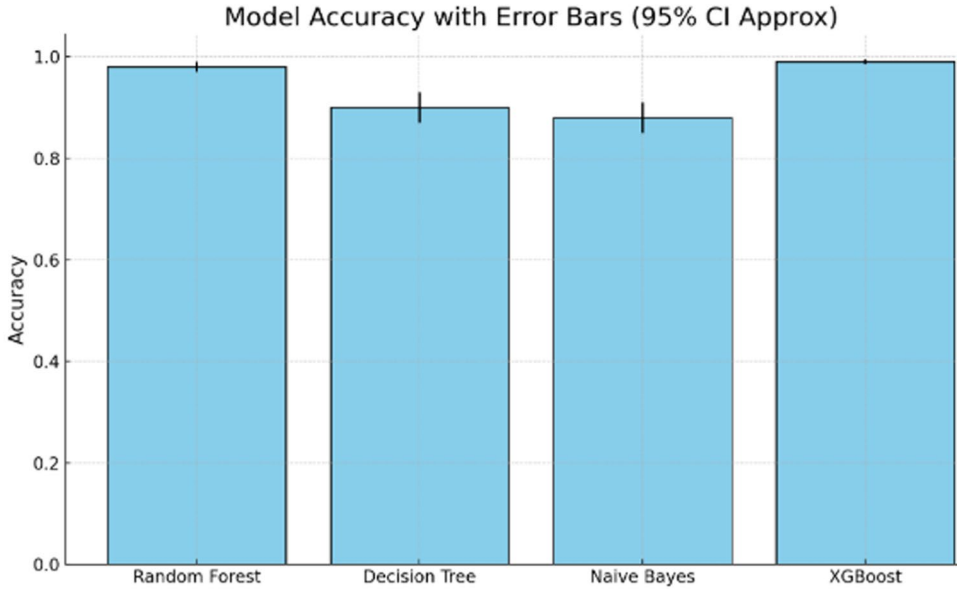

Fig. 5 Error bar plots for ML Models

Model Accuracy with Error Bars (95% CI Approx) is shown in from Fig. 5. Confidence Index (CI) quantifies uncertainty in the model's performance estimate. A narrow CI means high confidence and less variability suggesting the model's accuracy is consistently stable. A wide CI suggests higher variability in accuracy — the model's performance might fluctuate more across different samples.

The information that can be inferred from the plot of accuracy for different models, with error bars indicating the approximate 95% confidence interval (CI). To calculate the 95% confidence interval for the accuracy of a classification model, the normal approximation to the binomial distribution can be used:

$$CI = \hat{p} \pm Z_{\alpha}/2 \cdot \sqrt{((\hat{p}(1-\hat{p}))/n)} \quad (1)$$

Where:

$\hat{p}$ = observed accuracy (e.g., 0.99).

$Z_{\alpha/2}$ = critical value for 95% CI (≈ 1.96).

n = number of samples.

Example (for 99% accuracy for XGBoost model and say 1000 test samples):

$CI = 0.99 \pm 1.96 \cdot \sqrt{((0.99 \times (1–0.99))/1000)}$

$= 0.99 \pm 1.96 \cdot \sqrt{(0.0099/1000)}$

$= 0.99 \pm 1.96 \cdot 0.00315$

$= 0.99 \pm 0.0062$

Final CI ≈ [0.9838, 0.9962]

So with 99% accuracy and 95% CI it can be inferred that With 95% confidence, the true model accuracy lies between 98.38% and 99.62%.

- For Decision Tree: Accuracy increased from 65.3 to 89% post 5-fold cross validation and tuning of hyperparameters.
- For RF the accuracy improved to almost 99% and small error bar indicates low variability across folds with highly consistent performance.
- For NB classifier accuracy reached to 87% with larger error bar indicating more variation in accuracy across folds and less consistent.
- *Finally, for XGBoost model accuracy is highest and very tight error bar signifying stable and reliable results.*

XAI integration

To ensure model transparency, LIME was used to explain the predictions. LIME provided insights into the importance of features for each disease prediction. For instance:

- **Diabetes**: High cholesterol and specific insulin levels were critical contributors.
- **Thrombocytopenia**: Low platelet count was identified as the most significant factor.
- **Anemia and Thalassemia**: Hemoglobin levels and RBC counts were key predictors.

For each disease, LIME visualizations highlighted the relative importance of features, as shown in Figs. 6, 7, 8, 9 and 10.

The figure shown in Fig. 6 represents LIME interpretation for an individual who is anemic. As is clear the probability of the sample being of an anemic case is 99.26%, whereas those belonging to Diabetes or other disease is almost negligible. The model is 99% confident that the input sample belongs to a sample of a person having anemia. LIME also identifies which features pushed the ML model's decision towards a particular outcome, here disease being Anemia. As can be interpreted from Fig. 5, if the value of WBC count is less than 0.27, the person may

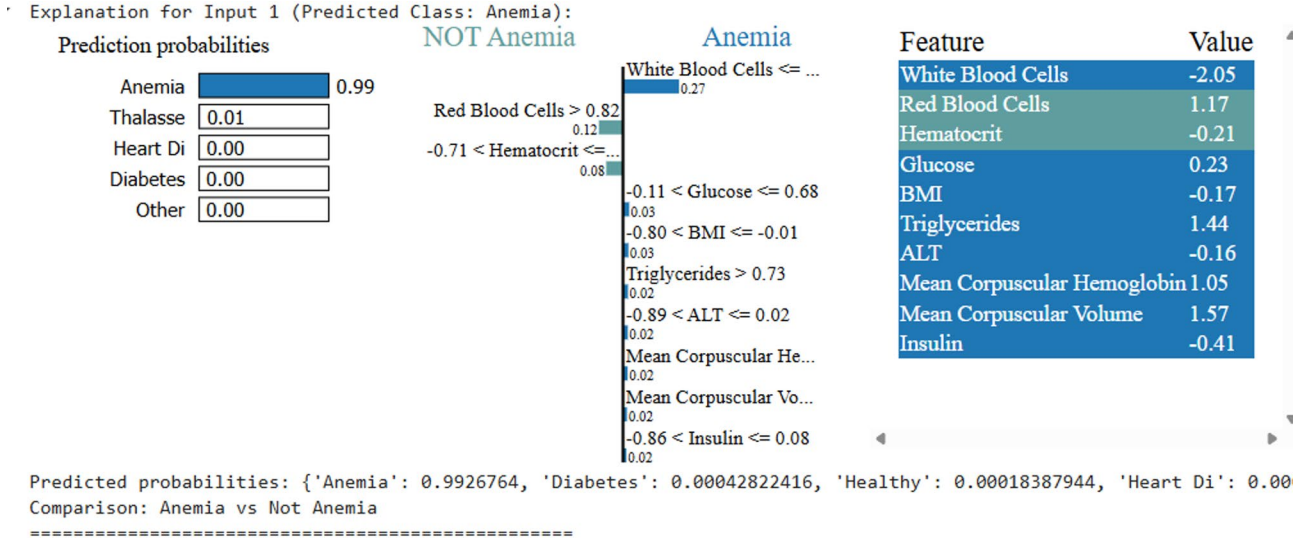

Fig. 6 LIME explanation for anemia prediction. low hemoglobin and RBC counts contributed most significantly

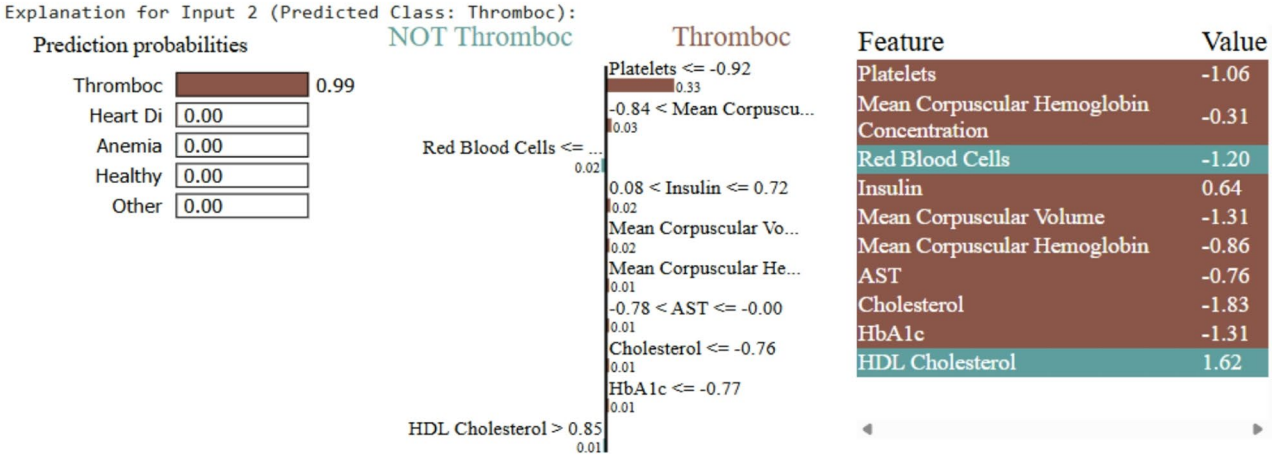

Fig. 7 LIME explanation for thrombocytopenia prediction. low platelet count was the key factor

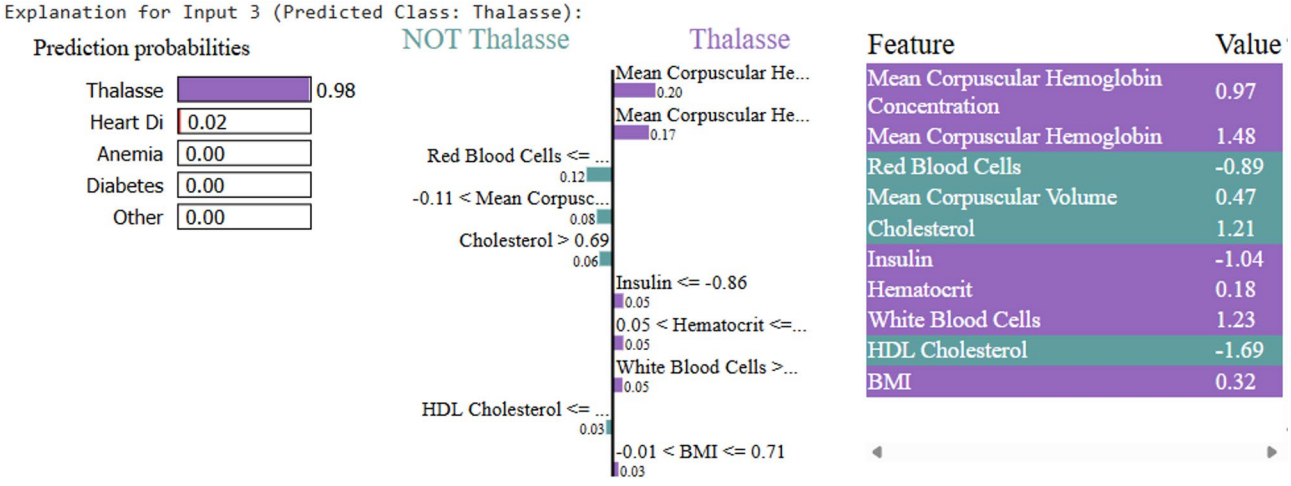

Fig. 8 LIME explanation for thalassemia prediction

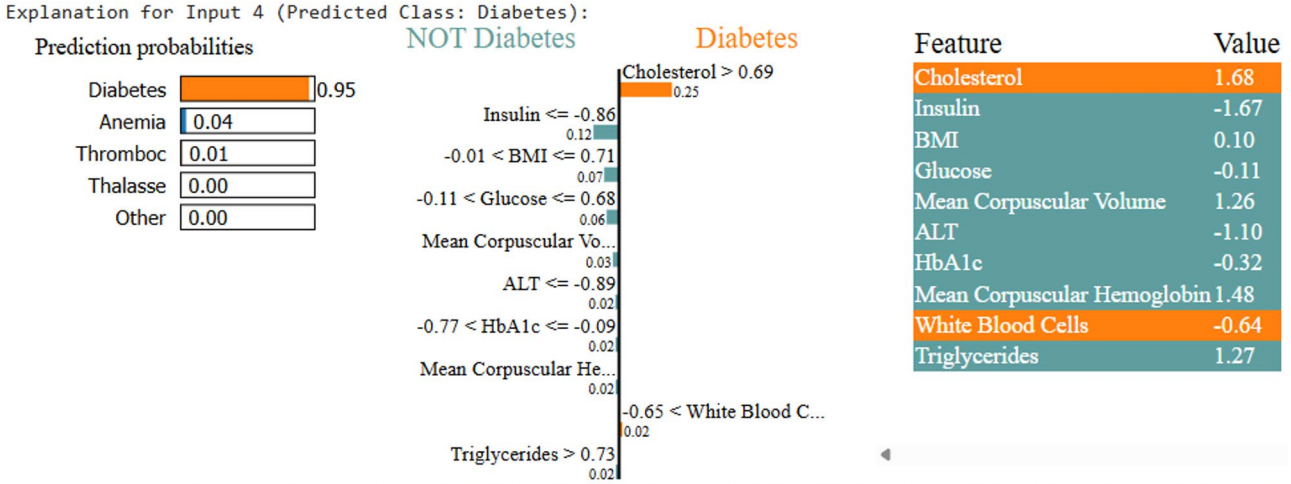

Fig. 9 LIME explanation for diabetes prediction

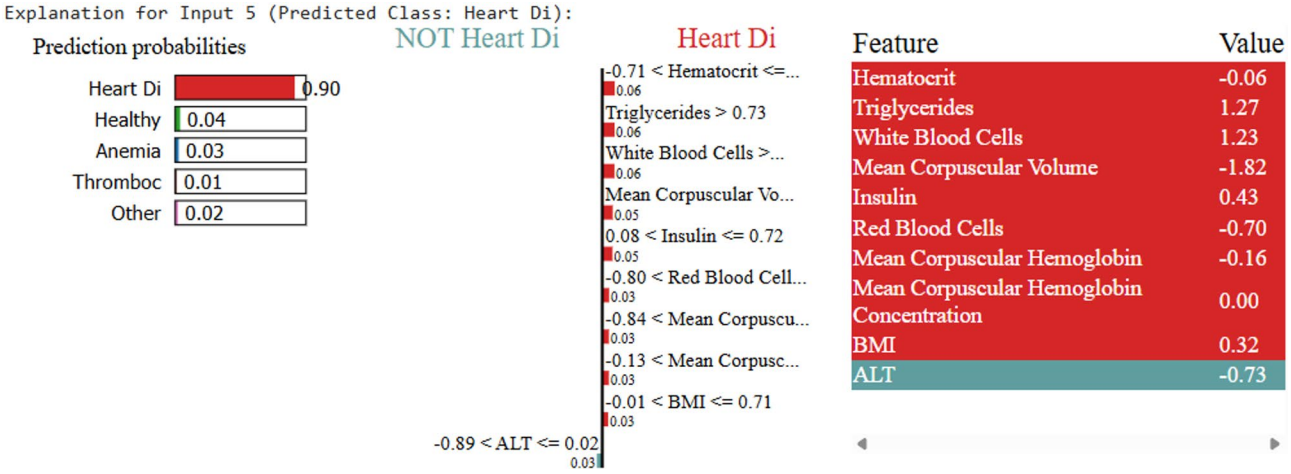

Fig. 10 LIME explanation for heart disease prediction

be suffering from Anemia. The WBC count of sample provided for validation is -2.05 which clearly is less than desired. Low WBC count suggests bone marrow suppression and immunity being low, this is a feature which clinically supports anemia. Clinically, WBC may might not be primary indicator, but low count signal underlying conditions that may co-exist with Anemia. For the particular sample this count has become the primary feature responsible for disease. Still another feature responsible for detection of Anemia by LIME model is RBC which in healthy case should be greater than 0.82. The RBC count of sample is 1.17, just slightly higher than needed, is another feature which lead ML model to draw the conclusion of the sample being that of an anemic person. Low RBC count indicates fewer number of cells to carry oxygen in blood thus making a person anemic.

Samples fed to LIME for other person can be interpreted similarly. Upon testing, the ML model predicts Thrombosis, with LIME providing the interpretability. As shown in Fig. 7, the low platelet count emerges as the key contributing factor. For the disease Thalassemia, a validated sample with an elevated Mean Corpuscular Hemoglobin Concentration leads to a Thalassemia prediction, as illustrated in Fig. 8. In the case of Diabetes, Fig. 9 highlights abnormal cholesterol levels as the primary indicator influencing the model's decision. Finally, for predicting cardiovascular or heart-related ailments, Fig. 10 shows that Hemocrit levels play a significant role in the model's interpretation for the specific sample, as revealed by LIME. Thus, post prediction from Ml model, LIME will interpret the reason for the disease, so that appropriate precautions can be taken by individuals to maintain the normal range.

Discussions and limitations

Practical considerations include infrastructural limitations such as the availability of electronic health record (EHR) systems, reliable internet connectivity, and adequate computational resources, which are often lacking in under-resourced healthcare environments. Furthermore, challenges such as data interoperability, model retraining with localized data, and the need for clinician training and acceptance are critical for successful deployment.

Regarding real-time feasibility, while the current model demonstrates strong predictive performance in a controlled experimental setting, its integration into Clinical Decision Support Systems (CDSS) requires further optimization. This includes streamlining data input pipelines, ensuring model inference efficiency, and embedding explainability features (through XAI) that align with clinicians' cognitive workflows to foster transparency and trust.

However, the study's reliance on a single Kaggle dataset and potential limitations in real-time clinical settings should be considered. Further research is needed to evaluate the scalability of this approach for broader healthcare applications and to test its feasibility across diverse clinical environments.

Conclusion

In this research, a novel approach that integrates ML and XAI techniques to enhance transparency and trust in healthcare applications is proposed. By leveraging the predictive power of XGBoost models alongside interpretability tools as LIME, the system achieves not only high accuracy but also clear and actionable insights into its decision-making process. This interpretability is crucial in healthcare, where trust and accountability are paramount. Table 8 presents a comparative analysis of the proposed work against existing state-of-the-art methods. While some studies show good performance, they often rely on smaller datasets or focus on a single disease. In contrast, our model achieves high accuracy while

Table 8 Comparative analysis of proposed work with current state of Art

Ref.	Multiple ML models were tested	Ensemble model selected	Data Balancing	Predicted Single/ Multiple Diseases	Included XAI	Model Accuracy	Limitation
[31]	CNN	Yes	No	No, chest related radiographs for detecting Pneumonia	No	88.1	With explanations behind outcome was not included. fully assess what factors might be contributing to the hospital system–specific biasing of the models.
[32]	Yes: SVM, NB and DT	NO-	No	Multiple: Diabetes Related Heart Disease	No	90%	Diabetes Related Heart Disease a single disease is predicted with 90% accuracy
[33]	Yes: 7 ML models	NA	NA	Yes: Multiple Disease, a comparative analysis	NO	NA	Basically, a research review article, that includes review of 48 articles done for disease prediction
[34]	Yes DT, XGB, RF, SVM, KNN, NB, GB, SG, LGBM, ET, ANN, HML	Yes-XG Boost	NO	No, single disease for Chronic Kidney Disease (CKD) prediction	No	99.2%	Small dataset (400 instances), generalizability concerns lacking expandability behind predictions
[35]	Yes: DT, KNN, NB, LR, RF, AB, SVM	Yes	Yes (Used SMOTE) 0	Single : Cardiac disease	No	96.6%	Small dataset (303 instances), overfitting risk and interpretability missing. Only predictions of Heart disease
Proposed work	Yes: DT, RF, NB and XGBoost	Yes: XGBoost	Yes, using SMOTE	5 Disease Predicted: Heart Disease, Diabetes, Thalassemia, Thrombocytopenia and Anemia	Yes	99%	Predicted Anemia, Diabetes, Heart Disease, Thalassemia, Thrombocytopenia with 99.1% accuracy with interpretability of diagnosis including features responsible for the cause.

predicting multiple diseases and enhances trust through integrated explainability.

A look in the table reflects that not only model predictions across multiple disease has been improved but also interpretations for occurrence of disease can be made to understand with the integration of XAI. This is required specifically in healthcare sector where patient's trust is much needed for faster recovery and also to gain acceptance. The study highlights that the XGBoost model out- performed others in terms of accuracy and robustness, while LIME effectively identified the contributions of individual features, ensuring transparency and fostering trust. Additionally, the application of SMOTE successfully addressed data imbalance, improving fairness and generalization across different disease categories. These findings demonstrate that XAI can bridge the gap between complex ML models and real-world clinical applications, empowering early disease risk assessment and supporting informed decision-making. The healthcare dataset utilized in the study was sourced from a publicly available repository, which, while widely used and benchmarked in research, may still carry inherent biases due to factors such as regional demographic representation, imbalance in class distribution, and variability in data acquisition protocols across collection sites. To mitigate these concerns, several preprocessing techniques including class balancing using SMOTE, normalization of feature values, and rigorous cross-validation to minimize overfitting to specific data segments are included. Additionally, XAI techniques were leveraged to further inspect and interpret model predictions, helping to uncover any potential model reliance on spurious or biased features. In future work, validation of models across multiple datasets from diverse populations is required.

This study shows that using XAI with machine learning can help make disease predictions more accurate and easier to understand. However, there are still many ways this work can be improved in the future. First, the model should be tested on more real-world medical data from different hospitals and regions to make sure it works well everywhere. Second, the system can be expanded to detect more types of diseases, including rare ones. It's also important to make the system easier to use in hospitals, so that doctors can get fast and helpful predictions during their work. Working closely with doctors and medical experts can help improve how the system explains its results, so it makes more sense to them. In the future, the model should also be able to update itself when new data is available and use privacy-friendly methods to keep patient information safe. Lastly, more research is needed to develop standards for how explainable AI should work in the field of pathology.

Author contributions
All authors contribute equally to manuscript.

Funding
Open access funding provided by Symbiosis International (Deemed University).

Data availability
No datasets were generated or analysed during the current study.

Declarations

Competing interests
The authors declare no competing interests.

Author details
[1]Department of Computer Science, Symbiosis Institute of Technology, Symbiosis International (Deemed University), Pune, Maharashtra 412115, India
[2]Department of Computer Science, Nottingham Trent University, College Drive, Clifton Lane, Nottingham NG11 8NS, UK

Received: 13 March 2025 / Accepted: 9 July 2025
Published online: 25 September 2025

References
1. Mirbabaie M, Stieglitz S, Frick NRJ. Artificial intelligence in disease diagnostics: A critical review and classification on the current state of research guiding future direction. Health Technol. 2021;11:693–731. https://doi.org/10.1007/s12553-021-00555-5.
2. Biswas AA. A comprehensive review of explainable AI for disease diagnosis. Array. 2024;22:100345. https://doi.org/10.1016/j.array.2024.100345.
3. Alowais SA, Alghamdi SS, Alsuhebany N, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Med Educ. 2023;23:689. https://doi.org/10.1186/s12909-023-04698-z.
4. Khanom F, Biswas S, Uddin MS, et al. XEMLPD: an explainable ensemble machine learning approach for Parkinson disease diagnosis with optimized features. Int J Speech Technol. 2024;27:1055–83. https://doi.org/10.1007/s10772-024-10152-2.
5. Kamal Alsheref F, Hassan W. Blood diseases detection using classical machine learning algorithms. Int J Adv Comput Sci Appl. 2019;10. https://doi.org/10.14569/IJACSA.2019.0100712.
6. Dehghani M, Yazdanparast Z. Discovering the symptom patterns of COVID-19 from recovered and deceased patients using apriori association rule mining. Inf Med Unlocked. 2023;42:101351. https://doi.org/10.1016/j.imu.2023.101351.
7. Mohit, Indukuri K, Santhosh Kumar UAK, Reddy, Badhagouni Suresh Kumar. An Approach to detect multiple diseases using machine learning algorithm. Journal of Physics:Conference Series. 2021;2089(1):012009.
8. Khanom F, Uddin MS, Mostafiz R. PD_EBM: an integrated boosting approach based on selective features for unveiling parkinson's disease diagnosis with global and local explanations. Eng Rep. 2025;7(1):e13091.
9. Elkenawy ESM, Alhussan AA, Khafaga DS, et al. Greylag Goose optimization and multilayer perceptron for enhancing lung cancer classification. Sci Rep. 2024;14:23784. https://doi.org/10.1038/s41598-024-72013-x.
10. Ramesh, Banoth G, Srinivas P, Ram Praneeth Reddy MD, Huraib Rasool D, Rawat, Sundaray M. Feasible Prediction of Multiple Diseases using Machine Learning. In E3S Web of Conferences. 2023;430:01051.
11. Alkhammash EH, Assiri SA, Nemenqani DM, Raad MM, Althaqafi M, Hadjouni F, Saeed, Elshewey AM. Application of machine learning to predict COVID-19 spread via an optimized BPSO model biomimetics. 2023;8(6):457. https://doi.org/10.3390/biomimetics8060457
12. Band SS, Yarahmadi A, Hsu CC, Biyari M, Sookhak M, Ameri R, Dehzangi I, Chronopoulos AT, Liang HW. Application of explainable artificial intelligence in medical health: A systematic review of interpretability methods. Inf Med Unlocked. 2023;40:101286. https://doi.org/10.1016/j.imu.2023.101286.
13. Guleria P, Srinivasu PN, Ahmed S, Almusallam N, and Fawaz Khaled Alarfaj. XAI frame work for cardiovascular disease prediction using classification techniques. Electronics. 2022;11(24):4086. https://doi.org/10.3390/electronics11244086
14. Amann J, Blasimme A, Vayena E, Frey D, Madai VI, Precise, 4Q Consortium. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Med Inf Decis Mak. 2020;20:1–9.
15. Magesh PR, Myloth RD, Tom RJ. An explainable machine learning model for early detection of parkinson's disease using LIME on DaTSCAN imagery. Comput Biol Med. 2020;126:104041. https://doi.org/10.1016/j.compbiomed.2020.104041.
16. Sheu RK, Pardeshi MS. A survey on medical explainable AI (XAI): recent progress, explainability approach, human interaction and scoring system. Sensors. 2022;22(20):8068. https://doi.org/10.3390/s22208068.
17. Gabbay F, Bar-Lev S, Montano O, Hadad N. A lime-based explainable machine learning model for predicting the severity level of covid-19 diagnosed patients. Applied Sciences. 11(21):10417. https://doi.org/10.3390/app112110417
18. Gaurav K, Kumar A, Singh P, Kumari A, Kasar M, Suryawanshi T. Human disease prediction using machine learning techniques and Real-life parameters. Int J Eng. 2023;36(6):1092–8. https://doi.org/10.5829/ije.2023.36.06c.07.
19. Khanom F, Mostafiz R, Uddin KMM. Exploring multimodal framework of optimized Feature-Based machine learning to revolutionize the diagnosis of parkinson's disease:AI-Driven insights. Biomedical Materials & Devices. 2025:1–20.
20. Bedi P, Thukral A, Dhiman S. Explainable AI in disease diagnosis. In: Aluvalu R, Mehta M, Siarry P, editors. Explainable AI in health informatics. Computational intelligence methods and applications. Singapore: Springer; 2024. https://doi.org/10.1007/978-981-97-3705-5_5.
21. Elshewey AM, Alhussan AA, Khafaga DS, et al. EEG-based optimization of eye state classification using modified-BER metaheuristic algorithm. Sci Rep. 2024;14:24489. https://doi.org/10.1038/s41598-024-74475-5.
22. Patel P, Chauhan S, Gupta S, Gupta T, Agrawal R. Mitigating class imbalance with ensemble SMOTEfied-GAN: advancing detection strategies for automobile insurance fraud. Int J Fuzzy Log Intell Syst. 2024;24(4):333–42.
23. Zahraa Tarek AA, Alhussan DS, Khafaga, El-Sayed M, El-Kenawy AM, Elshewey. A snake optimization algorithm-based feature selection framework for rapid detection of cardiovascular disease in its early stages. Biomed Signal Process Control. 2025;102:1746–8094. https://doi.org/10.1016/j.bspc.2024.107417.
24. El-Kenawy ESM, Khodadadi N, Mirjalili S, Abdelhamid AA, Eid MM, Ibrahim A. Greylag Goose optimization: nature-inspired optimization algorithm. Expert Syst Appl. 2024;238:122147.
25. Atteia G, El-kenawy ESM, Samee NA, Jamjoom MM, Ibrahim A, Abdelhamid AA, Azar AT, Khodadadi N, Ghanem RA, Shams MY. Adaptive dynamic dipper throated optimization for feature selection in medical data. Computers Mater Continua. 2023;75(1):1883–900.
26. El-Kenawy M, Khodadadi ES, Eid N. Improved cancer detection through feature selection using the binary al Biruni Earth radius algorithm. Sci Rep. 2025;15:9483. https://doi.org/10.1038/s41598-025-92187-2.
27. El-Kenawy ESM, Rizk FH, Zaki AM, Mohamed ME, Ibrahim A, Abdelhamid AA, Khodadadi N, Almetwally EM, Eid MM. Football optimization algorithm (fboa): A novel metaheuristic inspired by team strategy dynamics. J Artif Intell Metaheuristics. 2024;8(1):21–38.
28. Chunduru A, Kishore AR, Sasapu BK, et al. Multi chronic disease prediction system using CNN and random forest. SN COMPUT SCI. 2024;5:157. https://doi.org/10.1007/s42979-023-02521-6.
29. Kumar G, Agrawal R, Sharma K, Gundalwar PR, Agrawal P, Tomar M, Salagrama S. Combining BERT and CNN for sentiment analysis A case study on COVID-19. Int J Adv Comput Sci Appl. 2024;15(10).
30. Muhammad D, Keles A, Bendechache M. Towards Explainable Deep Learning in Oncology: Integrating EfficientNet-B7 with XAI techniques for Acute Lymphoblastic Leukaemia. 2024
31. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Vari able generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. PLOS Med. 2018;15:e1002683.
32. Arumugam K, Naved M, Shinde PP, Leiva-Chauca O, Huaman-Osorio A, Gonzales-Yanac T. Multiple disease prediction using machine learning algorithms. Mater Today: Proc. 2023;80:3682–5.
33. Uddin S, Khan A, Hossain ME, Moni MA. Comparing different supervised machine learning algorithms for disease prediction. BMC Med Inf Decis Mak. 2019;19(1):1–16.
34. Islam MA, Majumder MZH, Hussein MA. Chronic kidney disease prediction based on machine learning algorithms. J Pathol Inf. 2023;14:p100189. https://doi.org/10.1016/j.jpi.2023.100189. PMID: 36714452; PMCID: PMC9874070.
35. Narayanan J, Jayashree N. Implementation of efficient machine learning techniques for prediction of cardiac disease using Smote. Procedia Comput Sci. 2024;233:558–69.

Publisher's note