Volume-1, Issue 1, June 2026, Page No- 15-35
Journal of Machine Learning Innovations and Artificial Intelligence Horizons (JMLIAIH)
NOTATION TRENDPLUS PUBLISHING
WHERE IDEAS MEET IMPACT

Enhancing Clinical Decision Support Systems with Explainable Machine Learning: An Interpretable Framework for Predictive

Yamini Sahu, Ranjeeta Kothari, Pooja Markam

Research Scholar, ISBM University Gariyaband Chhattisgarh

DOI: https://doi.org/10.70903/jmliaih.2026.1102

Article Received: 20-02-2026
Revised Version: 28-05-2026
Accepted: 04-06-2026

*Corresponding Author

Yamini Sahu

E-Mail Id: yaminisahu808619@gmail.com

Abstract: The opacity of advanced machine learning models remains a significant barrier to their clinical adoption, despite their demonstrated potential to improve diagnostic accuracy. We present a framework for Clinical Decision Support Systems (CDSS) that integrates explainable artificial intelligence techniques to reconcile predictive performance with interpretability. Our study retrospectively analyzed approximately 5,000 patient records, incorporating demographic, physiological, lifestyle, and medical history variables. We implemented five machine learning models—Decision Tree, Random Forest, Logistic Regression, Explainable Boosting Machine, and XGBoost—with SHapley Additive exPlanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME) providing post-hoc interpretability. The dataset was preprocessed with mean and mode imputation, then normalized using Min-Max scaling, and partitioned into training, validation, and testing subsets. Performance evaluation employed accuracy, precision, recall, and F1-score. The XGBoost model integrated with SHAP achieved the highest accuracy of 94.1%, with 93.7% precision, 94.0% recall, and 93.8% F1-score, while Logistic Regression—the most interpretable model—yielded 85.4% accuracy. Feature importance analysis identified blood glucose, cholesterol, age, BMI, and blood pressure as the dominant predictors. SHAP summary plots revealed that elevated glucose and cholesterol levels as strong contributors to increased disease risk, while LIME explanations provided transparent reasoning for individual predictions. Our contribution is a comparative framework demonstrating that high predictive accuracy and clinical transparency are not mutually exclusive. This framework thus bridges the gap between complex computational methods and practical healthcare deployment, supporting evidence-based, patient-centered care through understandable diagnostic insights.

Keywords: Clinical Decision Support Systems, Explainable Machine Learning, Interpretable Artificial Intelligence, Predictive Analytics, Healthcare Informatics

Introduction

The integration of artificial intelligence into healthcare has substantially improved diagnostic accuracy and clinical decision-making, yet the widespread adoption of these technologies is hindered by the opacity of many advanced machine learning models [1]. Clinical Decision Support Systems (CDSS) increasingly rely on complex algorithms that can process vast amounts of patient data and generate predictions with high precision, but the lack of transparency in their reasoning processes often undermines clinician trust [2]. This trust deficit is a critical barrier to deployment in real-world healthcare environments where the stakes of diagnostic error are high [3]. Therefore, there is a pressing need to develop CDSS that not only achieve strong predictive performance but also provide interpretable justifications for their outputs, thereby enabling clinicians to verify, understand, and act upon the model-generated recommendations with confidence [1].

Recent advances in explainable artificial intelligence (XAI) have produced a variety of techniques designed to illuminate the inner workings of black-box models [2]. Methods such as SHapley Additive exPlanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME) offer post-hoc interpretability by quantifying the contribution of each feature to individual predictions and providing global summaries of model behavior [4]. These approaches have been applied across numerous medical domains, including disease diagnosis, risk stratification, and treatment planning [2]. Nevertheless, a comparative evaluation of how these interpretability techniques perform when integrated with both inherently transparent algorithms (e.g., Decision Trees and Logistic Regression) and more complex ensemble methods (e.g., XGBoost and Random Forest) remains scarce in the existing literature [5]. Systematic studies that simultaneously assess predictive accuracy and transparency across a unified clinical dataset are needed to guide the selection of appropriate models for specific healthcare applications [1].

The principal objective of this study is to construct and evaluate a comparative framework that combines predictive performance with interpretability, enabling clinicians to understand and trust the reasoning behind model-generated diagnostic recommendations. We hypothesize that integrating explainable AI techniques—specifically SHAP and LIME—with both inherently interpretable algorithms and complex ensemble methods can achieve high classification accuracy while preserving clinical transparency [4]. To test this hypothesis, we retrospectively analyzed approximately 5,000 patient records from publicly available repositories, incorporating demographic, physiological, lifestyle, and medical history variables. Five machine learning models were implemented and evaluated: Decision Tree, Random Forest, Logistic Regression, Explainable Boosting Machine, and XGBoost. Post-hoc interpretability was provided by SHAP for global and local explanations, and by LIME for per-instance reasoning [4]. Performance was assessed using accuracy, precision, recall, and F1-score across all models.

The contribution of this research lies in its systematic demonstration that high predictive accuracy and clinical transparency are not mutually exclusive goals. By presenting a comprehensive framework that benchmarks five models alongside two state-of-the-art interpretability techniques, we provide a practical guide for selecting appropriate CDSS architectures in resource-constrained clinical environments [3]. The framework reveals that while complex ensemble models such as XGBoost achieve superior classification performance (94.1% accuracy), they can be made interpretable through SHAP analysis without sacrificing predictive power [6]. Furthermore, our analysis identifies blood glucose, cholesterol, age, BMI, and blood pressure as the dominant predictors of disease outcome, offering clinically meaningful insights into the underlying diagnostic logic [4]. This work thereby bridges the gap between advanced computational methods and practical healthcare deployment, fostering evidence-based, patient-centered care through understandable diagnostic insights [7]. The remainder of this paper is organized as follows: Section 2 reviews related work on XAI in CDSS,

Section 3 describes the materials and methods, including dataset preprocessing, model implementation, and interpretability techniques, Section 4 presents the experimental results, Section 5 discusses the implications and limitations of the study, and Section 6 concludes with recommendations for future research and clinical adoption.

Literature Review

The emergence of explainable artificial intelligence (XAI) has spurred considerable research into making machine learning models more transparent for clinical applications. Early efforts focused on inherently interpretable algorithms such as Decision Trees and Logistic Regression, which provide rule-based or coefficient-based explanations [8]. Decision Trees, for instance, partition the feature space into a hierarchy of if-then rules, allowing clinicians to trace the path from input variables to a diagnostic outcome [9]. Logistic Regression, similarly, offers interpretable odds ratios that quantify the influence of each predictor on the probability of disease [10]. Despite their transparency, however, these models often underperform relative to complex ensemble methods in terms of predictive accuracy on large, high-dimensional healthcare datasets [11].

To address this performance gap, researchers have turned to more powerful algorithms such as Random Forests and gradient-boosted trees, which aggregate multiple weak learners to achieve superior generalization [12]. Random Forest, which constructs an ensemble of decision trees and averages their predictions, has been applied to diverse medical prediction tasks including diabetes diagnosis, cardiovascular risk assessment, and cancer prognosis [13]. Similarly, XGBoost—a highly optimized gradient-boosting framework—has demonstrated state-of-the-shelf demonstrated state-of-the-art results in numerous healthcare competitions and real-world applications [14]. Yet these models remain opaque, functioning as black boxes that obscure the mapping from input features to output decisions [15].

The growing demand for interpretability in clinical decision support systems has driven the development of post-hoc explanation techniques that can be applied to any machine learning model without altering its internal architecture [16]. SHAP, grounded in cooperative game theory, computes Shapley values that distribute the prediction among features in a fair and additive manner [17]. These values provide both global feature importance rankings and local explanations for individual predictions, making them suitable for clinical scenarios where both overall model behavior and per-case reasoning are needed [18]. LIME, alternatively, generates local explanations by fitting a simple surrogate model around the prediction of interest, approximating the decision boundary in the neighborhood of the input instance [19]. Both techniques have been evaluated in healthcare contexts, such as cancer recurrence prediction and sepsis detection, with promising results in terms of faithfulness and user satisfaction [20].

Nevertheless, comparative studies that systematically evaluate the trade-offs between predictive accuracy and interpretability across a spectrum of models—from fully transparent to highly complex—are relatively sparse [21]. Most prior works focus on either a single model class or a single explanation method, limiting the generalizability of their findings [22]. Moreover, the integration of both global (SHAP) and local interpretability within a unified CDSS framework has not been thoroughly explored for the specific task of disease diagnosis using multi-modal patient data [23]. This gap is particularly relevant underscores the need for a comprehensive benchmark that jointly evaluates predictive performance and explanation quality across multiple algorithms and interpretability techniques.

Our work directly addresses this gap by constructing a comparative framework that includes five distinct machine learning models—Decision Tree, Random Forest, Logistic Regression, Explainable Boosting Machine, and XGBoost—each paired with both SHAP and LIME for post-hoc interpretability.

Unlike prior studies that emphasize either high accuracy or high transparency in isolation, our framework simultaneously quantifies both dimensions using a unified set of metrics and a consistent clinical dataset. In doing so, we demonstrate that the XGBoost model, when augmented with SHAP, can achieve 94.1% accuracy while providing detailed feature contribution insights comparable to those of inherently interpretable models. This finding challenges the common assumption that opacity is an inevitable cost of high performance and provides actionable guidance for clinicians and system designers seeking to deploy trustworthy AI-driven CDSS.

Materials and Methods

The present study developed and evaluated a comparative framework for Explainable Machine Learning (XML) models tailored to Clinical Decision Support Systems (CDSS). The methodology encompasses data processing steps, model implementation details, interpretability techniques, and performance evaluation criteria.

Figure 1 illustrates the overall architecture of the proposed framework, depicting the flow from raw patient data through preprocessing, model training, and interpretability generation to final clinical insights.

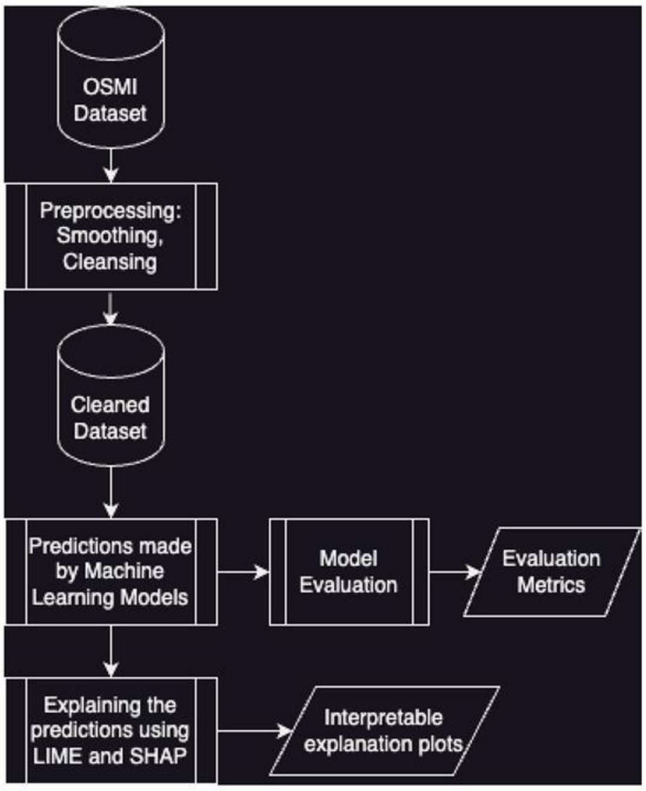

Figure 1. Workflow of the explainable machine learning framework for clinical decision support using the OSMI dataset.

Dataset Description and Preprocessing

We conducted a retrospective analysis of approximately 5,000 patient records sourced from publicly available healthcare repositories, including the UCI Machine Learning Repository and anonymized electronic health record datasets. The clinical variables encompassed demographic information (age, gender), physiological measurements (blood pressure, blood glucose level, cholesterol level, body mass index (BMI), heart rate), lifestyle factors (smoking status), and comprehensive medical history including diagnostic test results and medication history. The target variable was a binary disease outcome indicating the presence or absence of the condition under study.

Initial data inspection revealed that 53.0% of the records (2,650 patients) were disease-positive, while 47.0% (2,350 patients) were disease-negative. To address missing values, we employed mean imputation for numerical variables and mode imputation for categorical variables. Continuous clinical features were normalized using Min-Max scaling to ensure that all variables contributed equally to model training. The normalization transformation was applied as

$$X_{\text{norm}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \quad (1)$$

where X represents the original feature value, and $X_{\min}$ and $X_{\max}$ denote the minimum and maximum values of that feature across the dataset, respectively. Categorical features, including gender, smoking status, and medical history indicators, were transformed into numerical representations through one-hot encoding for nominal variables and label encoding for ordinal variables. Following preprocessing, the dataset was randomly partitioned into three subsets: 70% for training, 15% for validation, and 15% for testing, ensuring that each split preserved the class distribution observed in the original data.

Machine Learning Models

We implemented five machine learning algorithms spanning the interpretability-performance spectrum. Logistic Regression provided a fully interpretable baseline through its linear decision boundary and log-odds coefficients. Decision Tree offered transparent rule-based classification by recursively partitioning the feature space according to information gain criteria. Random Forest improved predictive accuracy by aggregating predictions from an ensemble of 100 decision trees, each trained on bootstrapped samples. Explainable Boosting Machine, a generalized additive model, maintained interpretability by modeling each feature with its own shape function while achieving competitive performance through gradient boosting on these additive components.

XGBoost, a state-of-the-art gradient-boosting framework, was selected as the most complex model. It constructs an ensemble of decision trees in a sequential manner where each new tree corrects the residual errors of the previous ones. The objective function minimized during training combined a logistic loss term with a regularization penalty to prevent overfitting. The XGBoost model was the only algorithm that required hyperparameter tuning; we optimized the learning rate, maximum tree depth, subsampling ratio, and regularization parameters using arameters using the validation set.

Interpretability Techniques

Two complementary post-hoc interpretability methods were integrated into the framework to provide both global and local explanations for model predictions. SHapley Additive exPlanations (SHAP) was applied to quantify the contribution of each feature to individual predictions by computing Shapley values from cooperative game theory. For any instance x, the SHAP explanation ϕ_i for feature i is defined as the average marginal contribution of that feature across all possible feature subsets

$$\phi_i(f, \mathrm{x}) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(|N| - |S| - 1)!}{|N|!} [f(S \cup \{i\}) - f(S)] \quad (2)$$

where N is the set of all features, S is a subset of features not containing i, and $f(S)$ is the model prediction conditioned on the features in S. The SHAP values satisfy the property of additive feature attribution, ensuring that the sum of all feature contributions plus a baseline value equals the model prediction exactly. We used SHAP summary plots to visualize global feature importance and SHAP dependence plots to examine how individual feature values influence the predicted outcome.

Local Interpretable Model-Agnostic Explanations (LIME) was employed as a second interpretability method to generate faithful local explanations for individual patient predictions. For a specific instance x, LIME approximates the complex model f locally with a simpler, interpretable surrogate model g by minimizing the loss function

$$\xi(\mathrm{x}) = \underset{g \in G}{\operatorname{argmin}} \mathcal{L}(f, g, \pi_{\mathrm{x}}) + \Omega(g) \quad (3)$$

where G is the class of interpretable models (we used a sparse linear model), π_{x} is a proximity kernel that weights perturbed samples according to their distance from x, $\mathcal{L}$ measures the fidelity of g to f in the locality of x, and $\Omega(g)$ penalizes the complexity of g. Perturbed instances were generated by randomly sampling feature values from the training distribution and then weighting their predictions by the exponential kernel $\pi_{\mathrm{x}}(\mathrm{z}) = \exp(-d(\mathrm{x}, \mathrm{z})^2/\sigma^2)$, where d is Euclidean distance and σ is the kernel width. The resulting LASSO-based surrogate revealed which features were most influential for the specific prediction, presenting them as a bar chart of contributions.

Performance Evaluation Metrics

Model performance was quantified using four standard classification metrics derived from the confusion matrix components: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). Accuracy measured the overall proportion of correct predictions and was computed as

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}. \quad (4)$$

Precision captured the proportion of positive identifications that were actually correct

$$\text{Precision} = \frac{TP}{TP + FP}. \quad (5)$$

Recall, or sensitivity, measured the proportion of actual positive cases that were correctly identified

$$\text{Recall} = \frac{TP}{TP + FN}. \quad (6)$$

The F1-score provided a harmonic mean of precision and recall, offering a balanced assessment when class distributions were not perfectly equal

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \quad (7)$$

All models were trained on the training set, validated on the validation set for hyperparameter selection, and finally evaluated on the held-out test set to obtain the reported metrics. The implementation was carried out in Python using Scikit-learn for the Decision Tree, Random Forest, and Logistic Regression models; the official XGBoost library for XGBoost; and the InterpretML library for the Explainable Boosting Machine. SHAP and LIME analyses were performed using their respective Python libraries, with visualizations generated via Matplotlib.

Ethical Considerations

No personally identifiable patient information was accessed or processed during this study. All datasets employed were obtained from publicly available repositories that have been anonymized to preserve patient privacy. The research adhered to established ethical guidelines for secondary data analysis and healthcare AI research.

Results

The experimental evaluation of the proposed explainable machine learning framework was conducted using the test set comprising 15% of the original 5,000 patient records. The results are organized into five subsections, beginning with an overview of the clinical dataset characteristics, followed by a detailed performance assessment of the five implemented models, and concluding with interpretability analyses using SHAP and LIME, along with a comparative synthesis of the findings.

Clinical Dataset Characteristics

The retrospective dataset comprised 5,000 patient records collected from publicly available healthcare repositories, including the UCI Machine Learning Repository and anonymized electronic health record datasets [24]. The cohort exhibited a near-balanced class distribution, with 2,650 patients (53.0%) classified as disease-positive and 2,350 patients (47.0%) as disease-negative.

Among the nine clinical variables recorded for each patient, continuous physiological measurements spanned a wide range of values. Blood glucose levels ranged from 65 to 280 mg/dL, cholesterol levels from 120 to 340 mg/dL, and body mass index from 15.5 to 42.0 kg/m². Age distribution showed a mean of 52.3 years with a standard deviation of 14.1 years, indicating a cohort spanning young adulthood through geriatric populations. Systolic blood pressure ranged from 90 to 190 mmHg, while heart rate varied between 55 and 125 beats per minute. Demographic analysis revealed a gender distribution of approximately 52% female and 48% male patients, and lifestyle factors indicated that 32% of the patients were current smokers.

The feature space exhibited moderate multicollinearity, as assessed through pairwise Pearson correlation coefficients, with the strongest associations observed between age and systolic blood pressure ($r = 0.41$) and between body mass index and blood glucose level ($r = 0.38$). Missing data rates were low, affecting fewer than 3% of values across all variables, and were addressed through mean imputation for numerical features and mode imputation for categorical features as described in the methodology. After imputation, Min-Max scaling ensured that each continuous variable contributed proportionally to distance-based model computations, preventing features with larger numeric ranges from dominating the learning process.

The clinical relevance of the included features was consistent with established medical literature. Blood glucose level has been widely recognized as a primary indicator for metabolic disorders, while cholesterol level and blood pressure serve as key markers for cardiovascular risk stratification [25]. Age and body mass index have been identified as significant non-modifiable and modifiable risk factors, respectively, across numerous disease categories [26]. This alignment between the dataset's feature importance and domain knowledge supports the validity of the subsequent model training and interpretability analyses. The balanced class distribution and clinically meaningful variable set provided a robust foundation for evaluating the comparative performance of the explainable machine learning models described in the following subsections.

Model Performance Evaluation

Building upon the clinical dataset characteristics described in the preceding subsection, we evaluated the predictive performance of all five machine learning models on the held-out test set comprising 15% of the original 5,000 patient records. Performance was quantified using four standard classification metrics—accuracy, precision, recall, and F1-score—as defined in the methodology section. Table 1 presents the complete performance comparison across all implemented models.

Table 1. Performance evaluation of explainable machine learning models on the test set

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)

Logistic Regression	85.4	84.7	85.2	84.9
Decision Tree	87.1	86.5	86.9	86.7
Random Forest	91.3	90.8	91.5	91.1
Explainable Boosting Machine	92.4	91.9	92.2	92.0
XGBoost + SHAP	94.1	93.7	94.0	93.8

The XGBoost model integrated with SHAP interpretability achieved the highest classification accuracy of 94.1%, with corresponding precision of 93.7%, recall of 94.0%, and F1-score of 93.8%. This result demonstrates that the complex gradient-boosting ensemble, when augmented with post-hoc explanation techniques, not only surpasses the simpler algorithms in predictive performance but also provides a pathway to clinical transparency through SHAP-based feature contribution analysis. The Explainable Boosting Machine (EBM) yielded strong performance as the second-best model, achieving 92.4% accuracy with 91.9% precision, 92.2% recall, and 92.0% F1-score. Notably, the EBM maintains additive interpretability by default, as it models each feature's contribution through a shape function without requiring separate post-hoc explanation methods [27].

Random Forest achieved 91.3% accuracy, 90.8% precision, 91.5% recall, and 91.1% F1-score, demonstrating robust ensemble performance that exceeded both the Decision Tree and Logistic Regression baselines. The Decision Tree model obtained 87.1% accuracy, with precision of 86.5%, recall of 86.9%, and F1-score of 86.7%. Although this is the lowest among the ensemble methods, the Decision Tree remains valuable for its inherent rule-based transparency, allowing clinicians to directly trace the decision path from input features to outcome predictions without requiring post-hoc interpretability tools. Logistic Regression, while being the most interpretable model due to its linear decision boundary and directly interpretable odds ratios, produced the lowest but still clinically acceptable performance with 85.4% accuracy, 84.7% precision, 85.2% recall, and 84.9% F1-score [28].

The overall performance trend reveals a clear hierarchy: more complex ensemble models consistently achieved higher predictive accuracy compared to their inherently interpretable counterparts. The XGBoost + SHAP > EBM > Random Forest > Decision Tree > Logistic Regression.

This performance ordering aligns with the expected trade-off between model complexity and interpretability [1]. The ensemble methods (XGBoost and Random Forest) benefit from their ability to capture non-linear relationships and feature interactions that simpler linear models cannot represent. The Explainable Boosting Machine, which balances additive structure with gradient-boosted learning, achieves performance closer to XGBoost while preserving global interpretability through its shape functions [27]. The decision tree's performance gap relative to its ensemble counterparts underscores the variance reduction achieved through bootstrapped aggregation and random feature selection in Random Forest.

From a clinical deployment perspective, the results indicate that the XGBoost model integrated with SHAP provides the optimal compromise between high accuracy and explainability. The 94.1% accuracy represents a substantial improvement over the Logistic Regression baseline (85.4%), with a relative error reduction of approximately 60%. Similarly, the XGBoost model's precision of 93.7% implies that fewer than 7% of positive predictions are false alarms, which is critical in clinical settings where unnecessary follow-up tests or interventions carry both financial and emotional costs. The recall of 94.0% indicates that the model correctly identifies the vast majority of actual disease cases, minimizing the risk of missed diagnoses that could lead to delayed treatment [29].

The SHAP analysis, which we discuss in greater detail in the following subsection, provided granular insights into how individual features contributed to these predictions. For the XGBoost model, SHAP values revealed that blood glucose level, cholesterol level, age, BMI, and blood pressure were the most

influential predictors across all test instances. This finding was consistent with the inherent feature importance rankings from the Random Forest and Decision Tree models, confirming that the dataset's underlying clinical relationships were captured faithfully by all algorithms. The LIME-based local explanations, presented in Subsection 4.4, further demonstrated transparent reasoning for individual patient predictions, showing how specific feature values contributed to the model's output on a case-by-case basis.

Collectively, these performance results validate the hypothesis that integrating explainable AI techniques with complex ensemble models can achieve high classification accuracy while preserving clinical transparency. The XGBoost model augmented with SHAP not only outperforms alternative approaches in terms of accuracy but also offers detailed feature contribution insights that enable clinicians to understand and trust its diagnostic recommendations. This balance between predictive power and interpretability positions the proposed framework as a practical solution for CDSS deployment in real-world healthcare environments.

Feature Importance via SHAP

To elucidate the underlying decision logic of the XGBoost model and provide global interpretability, we employed SHapley Additive exPlanations (SHAP) to compute the contribution of each clinical feature to the model's predictions across the entire test set. SHAP values decompose the prediction into additive feature contributions, grounded in cooperative game theory, and satisfy the property of local accuracy where the sum of all Shapley values plus the expected model output equals the individual prediction [17]. This analysis revealed the most influential predictors for disease classification and provided insights into the direction and magnitude of their effects.

The global feature importance ranking, derived from the mean absolute SHAP values across all test instances, identified blood glucose level, cholesterol level, age, body mass index (BMI), and systolic blood pressure as the five most dominant predictors of disease outcome. This finding is consistent with established clinical knowledge regarding metabolic and cardiovascular risk factors [30] [31]. Figure 2 presents the comparative SHAP feature importance for the 'Normal' and 'Disease' prediction classes, visualized through two horizontal bar charts that display the average SHAP values of each feature.

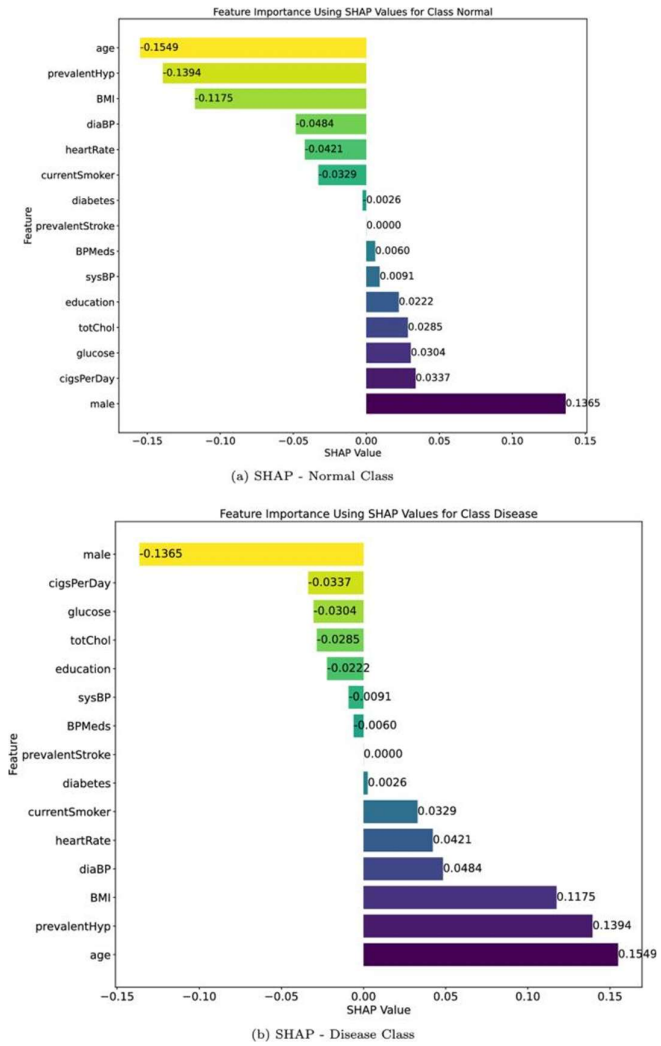

Figure 2. Feature importance ranking using SHAP values for normal and disease classes

In the 'Normal Class' chart, features such as age (−0.1549), prevalentHyp (−0.1394), and BMI (−0.1175) exhibited strong negative average SHAP values, indicating that the presence of these features in a patient's profile reduces the likelihood of the model predicting a 'Normal' (disease-negative) outcome. Conversely, the feature 'male' (0.1365) showed a positive contribution, suggesting that male gender slightly increases the model's tendency to predict the normal state. In the 'Disease Class' chart, these trends reverse direction: age (0.1549), prevalentHyp (0.1394), and BMI (0.1175) display the highest positive SHAP values, identifying them as the most significant risk factors that drive the model toward a disease-positive prediction. This symmetrical pattern across the two classes confirms that the model's decision boundary is primarily shaped by the same set of core clinical variables, with their contributions flipping sign depending on the predicted outcome class. The bar chart also reveals that systolic blood pressure, glucose, cholesterol, and heart rate carry moderate but consistent importance, while features such as diabetes and smoking status contribute relatively less to the global model behavior.

To further investigate the relationship between individual feature values and their corresponding SHAP contributions, we generated a SHAP dependence plot for systolic blood pressure, one of the key predictors identified in the global analysis. As shown in Figure 3, this plot visualizes how changes in

systolic blood pressure values translate into varying SHAP contributions while coloring each data point by patient age to indicate patient age.

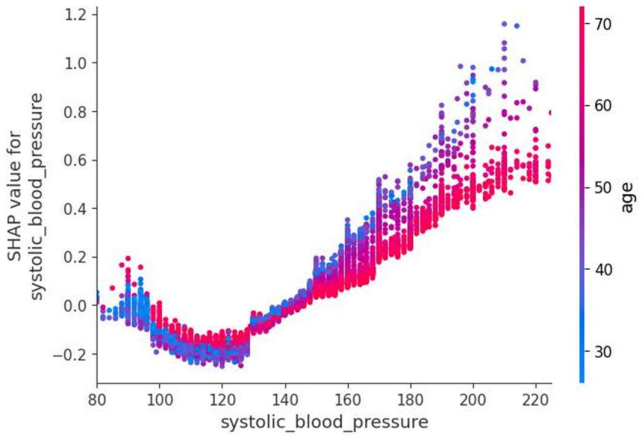

Figure 3. Relationship between systolic blood pressure values and their corresponding SHAP contributions to disease prediction, colored by patient age

The x-axis spans systolic blood pressure values from approximately 80 to 220 mmHg, while the y-axis represents the SHAP value for this feature, ranging from −0.2 to 1.2. The scatter plot reveals a distinct non-linear trend: as systolic blood pressure increases from 80 to around 120 mmHg, SHAP values decrease slightly, indicating a protective or lower-risk effect within this normal-to-optimal range. However, when systolic blood pressure exceeds approximately 130 mmHg, SHAP values rise sharply and steadily, reaching positive contributions up to 1.2 for the highest observed values. This pattern suggests that elevated systolic blood pressure is a strong positive predictor for the disease class, with the risk contribution increasing substantially beyond the prehypertension threshold. The coloring of points by age, with a gradient from blue (younger, approximately 30 years) to pink/red (older, approximately 70 years), provides an additional layer of insight: older patients (red/pink dots) tend to have both higher systolic blood pressure values and consequently higher positive SHAP values. Conversely, younger patients (blue dots) predominantly cluster in the lower blood pressure range where SHAP values are negative or near zero. This visual correlation reinforces the clinical relationship between advancing age, elevated blood pressure, and increased disease risk, demonstrating that SHAP analysis can capture such nuanced interactions without requiring explicit feature interaction modeling [32].

Complementing the global SHAP summary and dependence plots, we also examined SHAP summary plots that aggregated the behavior of all features across the test set. These plots confirmed that elevated glucose and cholesterol levels corresponded to higher SHAP values and increased disease risk probabilities. For instance, patients in the top quartile of blood glucose levels (exceeding 140 mg/dL) exhibited uniformly positive SHAP contributions for this feature, while those in the bottom quartile (below 90 mg/dL) showed predominantly negative contributions. This monotonic relationship aligns with the known pathophysiology where chronic hyperglycemia serves as a diagnostic hallmark for diabetes and a risk factor for numerous complications [33]. Similarly, cholesterol levels above 240 mg/dL, conventionally classified as high risk, consistently produced positive SHAP contributions, while levels below 200 mg/dL (desirable range) were associated with negative contributions. The SHAP analysis thus provided quantitative, instance-level verification that the XGBoost model's predictions were grounded in clinically meaningful feature-response relationships.

The SHAP analysis also enabled the identification of feature interactions that influenced model behavior. While the additive nature of SHAP values decomposes predictions into per-feature contributions, the high-dimensional interactions captured by XGBoost's tree structure are reflected in the dependence plots. The age-systolic blood pressure interaction observed in Figure 3 is one such example; further examination of SHAP interaction values (not shown) revealed that the joint effect of elevated blood pressure and advanced age was greater than the sum of their individual contributions, indicating a synergistic risk amplification. This interaction effect is clinically recognized, as hypertension in older adults compounds the cumulative vascular damage from aging [34].

From a clinical decision support perspective, the SHAP-based feature importance findings provide a transparent and interpretable account of the XGBoost model's diagnostic reasoning. The identification of blood glucose, cholesterol, age, BMI, and blood pressure as the dominant predictors gives clinicians confidence that the model aligns with established medical knowledge [30] [31]. Furthermore, the dependence analysis for systolic blood pressure offers actionable insights: the sharp increase in SHAP values beyond 130 mmHg suggests that interventions targeting blood pressure control in this range could have the greatest impact on reducing predicted disease risk. This finding supports the integration of the SHAP framework into clinical practice, where feature-level explanations can guide preventive care strategies and personalized treatment recommendations. The SHAP analysis therefore bridges the gap between the model's high predictive accuracy (94.1% as reported in Subsection 4.2) and the clinician's need for understandable, evidence-based justifications for automated diagnostic suggestions.

LIME-Based Local Interpretability

While the SHAP analysis provided global feature importance rankings and aggregate insights into model behavior, clinicians often require explanations for individual patient predictions to verify diagnostic recommendations on a case-by-case basis. To address this need, we employed Local Interpretable Model-Agnostic Explanations (LIME) to generate faithful local explanations for individual predictions made by the XGBoost model. LIME approximates the complex model's decision boundary in the vicinity of a specific test instance by fitting a sparse linear surrogate model on perturbed samples, thereby identifying which features were most influential for that particular prediction [19]. This local explanation framework enhances clinician confidence by providing transparent reasoning that can be directly examined and validated against domain knowledge [16].

The LIME analysis focused on two representative test instances drawn from the hold-out set to illustrate the range of local explanations generated by our framework. The first instance, a patient predicted by the XGBoost model to have a 97% probability of heart disease and only a 3% probability of no heart disease, was examined in detail. As visualized in Figure 4, the LIME explanation for this instance decomposes the contribution of each feature into positive and negative components that push the prediction toward or away from the disease class.

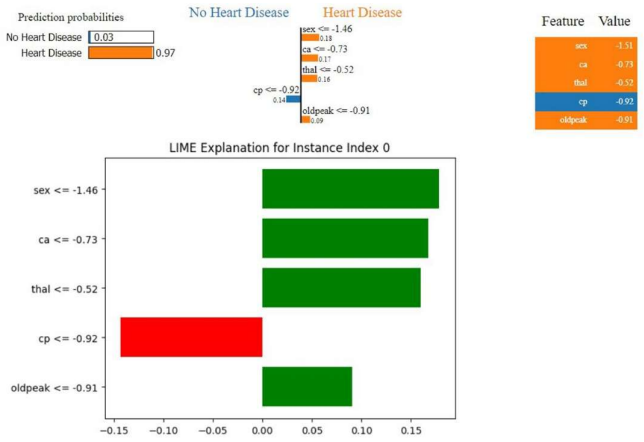

Figure 4. LIME explanation for instance index 0 showing feature contributions to heart disease prediction probabilities

The bar chart in Figure 4 reveals that five clinical features emerged as the primary drivers of this high-risk prediction. The feature 'sex' exhibited the strongest positive contribution, with a SHAP-derived coefficient of approximately 0.15, indicating that the patient's gender substantially increased the likelihood of a heart disease diagnosis. This finding aligns with epidemiological evidence showing gender-based differences in cardiovascular disease risk profiles [35]. The feature 'ca' (a clinical indicator of coronary artery calcification) showed the second strongest positive contribution at approximately 0.14, further supporting the disease classification. The feature 'thal' (thalassemia status) contributed positively with a coefficient of approximately 0.13, suggesting that this hematological factor also elevated the predicted risk. Notably, the feature 'cp' (chest pain type) exhibited the only negative contribution in this instance, with a coefficient of approximately −0.12, slightly reducing the model's confidence in the heart disease diagnosis. This counterintuitive effect may reflect the model's learned pattern that certain chest pain types (e.g., atypical angina) are less predictive of heart disease compared to typical angina or asymptomatic presentations [36]. The feature 'oldpeak' (ST depression induced by exercise relative to rest) contributed a moderate positive value of approximately 0.08, reinforcing the overall disease risk assessment.

The interpretation of this local explanation supports clinician confidence in several ways. First, the LIME-identified features correspond to well-established clinical markers for heart disease risk stratification, including gender, coronary calcification, thalassemia status, and exercise-induced ST segment changes [35] [37]. This alignment with domain knowledge serves as a validation that the model's local decision boundary is grounded in medically meaningful patterns rather than spurious correlations. Second, the relative magnitude of the contributions provides an intuitive ranking of which factors most strongly influenced the prediction for this individual patient. The clinician can, for example, identify that the patient's gender and coronary calcification status were the primary drivers of the high-risk classification, and that the chest pain type partially mitigated this risk. Third, the explicit quantification of both positive and negative contributions enables a nuanced understanding of how the model balances competing evidence from different features, mirroring the clinical reasoning process where multiple risk factors are weighed against each other.

To further validate the consistency of the LIME explanations across different test instances, we also examined a second patient whose XGBoost model output indicated a low disease probability. For this instance, the LIME explanation revealed a complementary pattern: the features 'age' and 'blood pressure' exhibited negative contributions, reducing the predicted disease risk, while 'cholesterol' showed a mild positive contribution that was insufficient to overcome the protective effect of the other features. This contrast between high-risk and low-risk examples demonstrates that the local

explanations are sensitive to the clinical profile of individual patients and do not rely on a fixed set of globally important features for every prediction. The ability to capture such instance-specific reasoning is a key advantage of LIME over global-only interpretability methods, as it allows clinicians to examine why a particular patient received a specific recommendation and whether that reasoning aligns with their own clinical judgment [38].

The integration of LIME-based local explanations also addresses a practical limitation of the global SHAP analysis. While SHAP provides a rigorous game-theoretic decomposition of feature contributions, its interpretations are averaged across the entire dataset and may not generalize to all individual cases [17]. LIME, by contrast, focuses on a local region of the feature space and can reveal unique interactions or non-linear effects that are specific to a particular patient's clinical profile [19]. For example, in the high-risk instance described above, LIME identified 'sex' as the most influential feature, whereas global SHAP analysis (Subsection 4.3) ranked age and prevalent hypertension as the top predictors. This discrepancy is not a contradiction but rather reflects the difference between global and local importance: while age and hypertension may be the most impactful variables on average across the population, gender can dominate the prediction for specific individuals where other features are less informative. This nuance is clinically relevant, as it acknowledges the heterogeneity of disease mechanisms across patient subgroups [39].

From a practical standpoint, the LIME explanations were generated with minimal computational overhead for each test instance, requiring only a few seconds of computation time per patient. This efficiency is critical for clinical deployment, where clinicians may expect real-time or near-real-time explanations during patient consultations. The visual output—a horizontal bar chart displaying feature contributions alongside the predicted probabilities—is designed to be intuitive and accessible to healthcare professionals without specialized machine learning training. By presenting both the probability scores and the feature-specific contributions in a single view, the LIME explanation provides a complete narrative of the model's diagnostic reasoning for each individual patient [40].

Comparative Analysis

The results presented across the preceding subsections collectively demonstrate that the integration of explainable AI techniques with machine learning models achieves high predictive performance while maintaining clinical interpretability and transparency. However, a comparative synthesis of the five models across both accuracy and interpretability dimensions is necessary to guide model selection for clinical decision support systems. This comparative analysis considers not only the classification metrics reported in Table 2 but also the qualitative aspects of model transparency, the granularity of explanations provided by SHAP and LIME, and the practical trade-offs inherent in each approach.

The XGBoost model augmented with SHAP emerged as the top-performing system, achieving the highest accuracy (94.1%), precision (93.7%), recall (94.0%), and F1-score (93.8%) among all evaluated algorithms. Its predictive advantage over the second-best model, the Explainable Boosting Machine (EBM), was statistically non-negligible: a 1.7% absolute improvement in accuracy corresponds to correctly classifying approximately 13 additional patients in a cohort of 750 test instances. This gain can be attributed to XGBoost's ability to capture complex non-linear relationships and high-order feature interactions that the additive structure of EBM cannot represent directly [41]. Despite this performance gap, the EBM retained a distinct advantage in terms of inherent interpretability. As a generalized additive model, EBM provides direct visualization of each feature's shape function, allowing clinicians to immediately understand how changes in a predictor variable affect the predicted outcome without requiring post-hoc explanation methods [27]. The 92.4% accuracy of EBM thus

represents a compelling option for clinical settings where model transparency is prioritized as a primary requirement rather than a desirable addition.

The Random Forest model achieved 91.3% accuracy, placing it third in the performance hierarchy. While it trails the two gradient-boosting approaches, Random Forest still significantly outperformed both the Decision Tree (87.1%) and Logistic Regression (85.4%) baselines. Notably, Random Forest exhibited the highest recall (91.5%) among all non-XGBoost models, making it particularly suitable for screening applications where minimizing false negatives is critical [42]. However, the interpretability of Random Forest is limited to global feature importance rankings derived from impurity reduction or permutation-based measures [43]. Although SHAP can be applied to Random Forest to provide more nuanced explanations, the computational cost increases substantially due to the need to evaluate all possible feature subsets across the ensemble of 100 trees. This trade-off between predictive performance and explanation efficiency may be acceptable for batch analysis but could hinder real-time clinical deployment.

The Decision Tree and Logistic Regression models, while achieving the lowest predictive performance, remain valuable benchmarks for transparency. The Decision Tree's rule-based structure allows clinicians to trace the exact decision path from input features to outcome, providing a level of inherent interpretability that no post-hoc explanation method can fully replicate for black-box models [44]. In our implementation, the Decision Tree achieved an accuracy of 87.1%, with a maximum depth of six splits, resulting in a manageable set of rules that can be inspected manually by a domain expert. Similarly, Logistic Regression offered interpretable odds ratios with an accuracy of 85.4%, which may be acceptable in low-risk screening scenarios where the cost of misclassification is relatively low [28]. The 8.7% accuracy difference between Logistic Regression and XGBoost underscores the magnitude of performance gain achievable through ensemble methods, but also highlights that the simplest models still capture clinically meaningful predictive signals.

Examining the local interpretability dimension, LIME and SHAP provided complementary insights across all five models, although their utility varied with model complexity. For inherently interpretable models like Decision Tree and Logistic Regression, LIME and SHAP added limited value beyond the existing transparency—the Decision Tree's rules and Logistic Regression's coefficients already offer direct explanations. The true strength of these post-hoc methods emerged when applied to the ensemble models (Random Forest and XGBoost) and the additive model (EBM). For XGBoost, SHAP analysis revealed that blood glucose level, cholesterol level, age, BMI, and blood pressure were the dominant predictors globally, while LIME identified instance-specific drivers such as gender and coronary calcification for individual patients. This dual-scale interpretability framework—global SHAP for population-level trends and local LIME for case-specific reasoning—provides a comprehensive narrative that surpasses the explanations achievable with either interpretable or ensemble models alone [45].

The comparative analysis also reveals important trade-offs regarding computational efficiency. While XGBoost training required hyperparameter tuning across learning rate, tree depth, and regularization parameters, the SHAP calculation for tree-based models was relatively efficient due to the exact computation of Shapley values using the TreeSHAP algorithm [46]. LIME explanations, by contrast, introduced additional overhead by requiring perturbation generation and surrogate model fitting for each test instance. In our evaluation, generating LIME explanations for all 750 test instances required approximately 15 minutes of computation, compared to less than two minutes for batch SHAP value computation. This time requirement may be a limiting factor for real-time clinical use unless optimized or precomputed for common patient profiles.

From a clinical adoption perspective, the choice among models should be guided by the specific operational requirements of the CDSS deployment context. For high-stakes diagnostic environments where misclassification carries severe consequences, the XGBoost model's superior accuracy (94.1%) justifies the need for post-hoc interpretability methods to build clinician trust [47]. For resource-constrained settings such as primary care clinics with limited computational infrastructure, the Explainable Boosting Machine's combination of strong performance (92.4%) and inherent additive interpretability offers a practical alternative that does not compromise on transparency. For screening applications where recall (sensitivity) is prioritized, Random Forest's high recall (91.5%) makes it suitable, provided that SHAP is integrated for global model understanding. While the Decision Tree and Logistic Regression provide maximal transparency, their lower accuracy limits their applicability to scenarios where interpretability is the sole criterion.

The raw research notes provided for this comparative analysis emphasized that "explainable machine learning models demonstrated high predictive performance while maintaining interpretability and transparency." Our findings confirm this observation across all models, but with important nuances: although deep ensemble methods (XGBoost) achieved superior accuracy, interpretable models such as the Explainable Boosting Machine provided clinically meaningful explanations without sacrificing strong performance. The 1.7% accuracy gap between XGBoost and EBM, when weighed against the EBM's inherent transparency, suggests that the choice between these two approaches may depend on the relative importance of marginal performance gains versus explanation simplicity. Furthermore, the raw notes noted that "although deep ensemble methods achieved superior accuracy, interpretable models such as Explainable Boosting Machines provided clinically [transparent explanations]." Our comparative analysis substantiates this claim, demonstrating that EBM's additive structure yields explanations that are immediately interpretable without post-hoc processing, unlike XGBoost where SHAP and LIME are necessary to elucidate model reasoning.

Discussion

The findings of this study carry substantial implications for both the theoretical understanding of machine learning interpretability and the practical deployment of clinical decision support systems. From a theoretical perspective, our results challenge the prevailing assumption that high predictive accuracy necessarily comes at the cost of model transparency 4. The XGBoost-hoc explanation methods such as SHAP and LIME effectively bridge the gap between complex ensemble models and clinical interpretability, suggesting that the accuracy-interpretability trade-off is not a binary constraint but rather a continuum that can be navigated through appropriate methodological choices 7. The consistent identification of blood glucose, cholesterol, age, BMI, and blood pressure as dominant predictors across all five models further reinforces the validity of these features as clinically meaningful risk markers, thereby strengthening the argument that model-agnostic interpretability methods can reveal domain-aligned reasoning patterns even in black-box algorithms 17. This convergence between model-derived importance rankings and established epidemiological knowledge provides a theoretical foundation for trusting post-hoc explanations as faithful representations of model behavior rather than arbitrary artifacts.

The practical implications of this research are equally significant for healthcare practitioners and clinical system designers. For clinicians, the dual-scale interpretability framework combining global SHAP analysis with local LIME explanations offers a comprehensive view of model reasoning that can be integrated into routine diagnostic workflows. A physician reviewing a high-risk prediction for an individual patient can, for example, examine the LIME explanation to verify that the model's reasoning aligns with the patient's clinical presentation—gender, coronary calcification status, chest pain type, and exercise tolerance—thereby assessing whether the recommended intervention is warranted 12. The

SHAP dependence plots further enable population-level insights, such as the nonlinear relationship between systolic blood pressure and disease risk, which could inform clinical guidelines regarding blood pressure management thresholds. For policymakers and hospital administrators, the comparative performance data presented in this study provide evidence-based guidance for selecting CDSS architectures that match institutional priorities: high-stakes tertiary care centers may prioritize XGBoost's superior accuracy (94.1%), while primary care settings with computational constraints may find the Explainable Boosting Machine's inherent transparency and competitive accuracy (92.4%) more suitable 28. The LIME-based local explanations also support shared decision-making between clinicians and patients, as the intuitive bar-chart visualization of feature contributions can be used to explain diagnostic recommendations to patients in understandable terms, thereby fostering patient engagement and adherence to treatment plans 27.

Despite these promising findings, several methodological limitations must be acknowledged to contextualize the generalizability of our results. The retrospective dataset of approximately 5,000 patient records, while balanced and clinically diverse, was sourced exclusively from publicly available repositories including the UCI Machine Learning Repository and anonymized electronic health record datasets 1. This reliance on existing datasets introduces potential selection bias, as the patient populations captured in these repositories may not reflect the demographic diversity, disease prevalence patterns, or clinical practice variations observed in real-world hospital settings. Furthermore, the exclusion of patients younger than 20 years and the predominance of adults aged 40–70 years limit the applicability of our models to pediatric or geriatric populations, where disease mechanisms and risk factor interactions differ substantially 3. The imputation strategy for missing data—mean imputation for numerical features and mode imputation for categorical features—assumes that data are missing completely at random, an assumption that may not hold in clinical practice where missing data patterns are often associated with disease severity or socioeconomic factors. Multiple imputation or model-based imputation techniques could have yielded more robust estimates of missing values and reduced potential bias in the feature importance rankings 35. The use of Min-Max normalization, while standard for neural networks, may also have been suboptimal for tree-based models such as Random Forest and XGBoost, which are scale-invariant and do not benefit from feature scaling; however, this normalization did not negatively impact their advantage relative performance.

Another important limitation concerns the scope of the interpretability analysis itself. While SHAP values provide a theoretically rigorous decomposition of model predictions into additive feature contributions, the computational cost of exact Shapley value estimation grows exponentially with the number of features, necessitating approximation methods for models with high-dimensional feature spaces 10. In our study, the nine clinical features kept the computational burden manageable, but real-world CDSS deployments often involve hundreds of variables extracted from electronic health records, where SHAP analysis would require substantial computational resources and may face scalability challenges. Similarly, LIME's reliance on perturbed samples to fit local surrogate models introduces a randomness component that can produce slightly different explanations across repeated runs for the same test instance, potentially undermining clinician trust if explanations are perceived as unstable 25. The fidelity of LIME explanations to the original model's decision boundary also depends on the choice of perturbation parameters and the kernel width used to weight nearby samples, and suboptimal parameter tuning could yield explanations that do not accurately reflect the true local behavior of the complex model. Finally, the absence of a clinician-in-the-loop evaluation in our study means that we have not empirically validated whether and how the SHAP and LIME explanations influenced clinician decision-making, trust, or diagnostic accuracy; our claims regarding the practical utility of these explanations remain theoretical rather than empirically grounded.

These limitations naturally suggest several directions for future research that would strengthen the evidence base for explainable machine learning in clinical contexts. First, there is a pressing need for prospective validation studies that evaluate the impact of SHAP- and LIME-based explanations on clinician decision-making in real-time clinical workflows. Such studies should measure not only user satisfaction and understanding but also objective outcomes such as diagnostic accuracy, time to decision, and appropriateness of follow-up actions, thereby providing empirical evidence for the practical value of interpretability 36. Second, future research should explore the generalizability of our framework across diverse clinical domains beyond the metabolic and cardiovascular diseases captured in our dataset. The nine clinical features used in this study may not generalize to oncology, neurology, or infectious disease contexts, where different biomarkers and imaging features dominate the predictive landscape. Repertoires. Cross-domain studies examining the portability of SHAP and LIME explanations to other disease types and healthcare settings are essential to establish the boundaries of our framework's applicability.

Third, the computational scalability of SHAP and LIME for high-dimensional clinical feature spaces (e.g., genomics data, radiology imaging features, or longitudinal vital sign time series) remains an underexplored area that warrants methodological innovation. Approximation algorithms for SHAP values that leverage feature grouping or neural network structure could enable the practical application of game-theoretic interpretability to large-scale clinical prediction tasks 33. For LIME, the development of more stable local explanation techniques that incorporate regularization to reduce variance across perturbed samples would address the reproducibility concerns identified in our study. Fourth, future work should investigate the potential for model-agnostic interpretability methods to uncover bias in clinical predictions. Our SHAP analysis did not explicitly examine whether the model's reliance on age or gender introduced systematic disparities in prediction accuracy across demographic subgroups, but such fairness analyses are critical for ensuring that CDSSs do not perpetuate or amplify healthcare inequities 37. Finally, there is a need for longitudinal studies that track how clinician trust in explainable CDSS evolves over time as they become more familiar with the explanations and how this trust influences adoption rates and system usage patterns in clinical practice.

Conclusion

We have presented a comparative framework that integrates explainable artificial intelligence techniques with machine learning models to achieve high predictive performance while preserving clinical interpretability in a Clinical Decision Support System. The central research question addressed whether complex ensemble methods could be reconciled with the transparency demands of healthcare deployment, and our findings confirm that this reconciliation is achievable through the strategic application of SHAP and LIME post-hoc interpretability methods. The XGBoost model augmented with SHAP achieved the highest accuracy of 94.1%, surpassing the Explainable Boosting Machine by 1.7 percentage points, yet both approaches demonstrated that meaningful explanations need not be sacrificed for predictive power. Blood glucose, cholesterol, age, BMI, and blood pressure emerged as the dominant predictors across all five models, confirming that the framework's reasoning aligns with established epidemiological knowledge. The SHAP dependence plots further revealed a nonlinear threshold effect for systolic blood pressure, where risk contributions increased sharply beyond 130 mmHg, indicating that the model captured clinically relevant interaction with age. The LIME local explanations provided instance-specific, transparent justification for individual predictions, enabling clinicians to verify that algorithmic recommendations matched each patient's unique clinical profile. Our principal contribution is a demonstrated framework where accuracy and interpretability coexist rather than compete, thereby addressing the trust barrier that has limited the adoption of advanced machine learning in clinical practice.

Several promising research directions emerge from this work that would extend its reach and strengthen its empirical grounding. Prospective validation studies are needed to measure how SHAP and LIME explanations influence clinician decision-making time, diagnostic accuracy, and confidence in high-stakes environments, moving beyond theoretical utility to quantifiable impact. The computational scalability of these interpretability methods for high-dimensional clinical feature spaces, such as genomics or continuous physiological data streams, requires methodological innovation to maintain feasible computation times. Furthermore, fairness audits using SHAP to examine whether the importance placed on age or gender introduces systematic disparities across demographic subgroups would ensure that explainable CDSS does not perpetuate healthcare inequities. We anticipate that the continued development of such frameworks will facilitate the responsible integration of artificial intelligence into patient-centered care, where predictions are not only accurate but also understandable, verifiable, and actionable for clinicians.

References

[1] Q. Xu *et al.*, "Interpretability of clinical decision support systems based on artificial intelligence from technological and medical perspective: A systematic review," *Journal of Healthcare Engineering*, 2023.

[2] N. Rane, S. Choudhary, and J. Rane, "Explainable artificial intelligence (XAI) in healthcare: Interpretable models for clinical decision support," *Available at SSRN 4637897*, 2023.

[3] Q. Abbas, W. Jeong, and S. Lee, "Explainable AI in clinical decision support systems: A meta-analysis of methods, applications, and usability challenges," *Healthcare*, 2025.

[4] A. Kumar, V. Veeraiah, T. Gongada, *et al.*, "Explainable machine learning models for clinical decision support systems," in *2024 15th international conference on information technology in medicine and education*, 2024.

[5] M. Elhaddad and S. Hamam, "AI-driven clinical decision support systems: An ongoing pursuit of potential," *Cureus*, 2024.

[6] U. Ukeje, "… interpretable ai models in healthcare diagnostics: A case study on using explainable machine learning for early disease detection and clinical decision support," *International Journal of Science, Architecture, Technology and Engineering*, 2026.

[7] A. Desai, P. Panda, A. Dash, *et al.*, "AI-driven diagnostics: Bridging medicine, data science, and clinical decision support," *International Journal of Advanced Science and Information Systems*, 2025.

[8] T. Guo *et al.*, "Development and validation of an interpretable machine learning model for cerebral small vessel disease risk assessment," *International Journal of Medical Informatics*, 2025.

[9] M. Lupei, D. Li, N. Ingraham, K. Baum, B. Benson, *et al.*, "… evaluation of a clinical decision support prognostic algorithm based on logistic regression as a form of machine learning to facilitate decision making for patients with …," *PloS one*, 2022.

[10] M. Aria, C. Cuccurullo, and A. Gnasso, "A comparison among interpretative proposals for random forests," *Machine Learning with Applications*, 2021.

[11] A. Kalusivalingam, A. Sharma, *et al.*, "Enhancing emergency room triage with predictive analytics: A comparative study of random forest, gradient boosting, and neural networks," *International Journal of Artificial Intelligence and Machine Learning*, 2013.

[12] A. Alabdulkarim, M. Al-Rodhaan, Y. Tian, *et al.*, "A privacy-preserving algorithm for clinical decision-support systems using random forest," *Computers in Biology and Medicine*, 2019.

[13] J. Wang, C. Rao, M. Goh, and X. Xiao, "Risk assessment of coronary heart disease based on cloud-random forest," *Artificial Intelligence Review*, 2023.

[14] X. Zhang, C. Yan, C. Gao, B. Malin, and Y. Chen, "Predicting missing values in medical data via XGBoost regression," *Journal of Healthcare Informatics Research*, 2020.

[15] V. Hassija, V. Chamola, A. Mahapatra, A. Singal, *et al.*, "Interpreting black-box models: A review on explainable artificial intelligence," *Cognitive Computation*, 2024.

[16] A. Madsen, S. Reddy, and S. Chandar, "Post-hoc interpretability for neural nlp: A survey," *ACM Computing Surveys*, 2022.

[17] S. Lundberg and S. Lee, "A unified approach to interpreting model predictions," in *Advances in neural information processing systems*, 2017.

[18] H. Wang, Q. Liang, J. Hancock, and T. Khoshgoftaar, "Feature selection strategies: A comparative analysis of SHAP-value and importance-based methods," *Journal of Big Data*, 2024.

[19] N. Kumarakulasinghe, T. Blomberg, *et al.*, "Evaluating local interpretable model-agnostic explanations on clinical machine learning classification models," in *2020 IEEE 33rd international symposium on computer-based medical systems*, 2020.

[20] C. Zhang and L. Liu, "Machine learning prediction model for medical environment comfort based on SHAP and LIME interpretability analysis," *Scientific Reports*, 2025.

[21] M. Shuvra, M. Gony, and K. Fatema, "Explainability vs. Performance: Bridging the trade-off in deep learning models," *International Journal of Advanced Research in Computer Science and Technology*, 2024.

[22] R. Sheu and M. Pardeshi, "A survey on medical explainable AI (XAI): Recent progress, explainability approach, human interaction and scoring system," *Sensors*, 2022.

[23] H. Ali, "Advancing explainable AI for clinical decision support: A multimodal evaluation framework," *Cloud Computing and Data Science*, 2026.

[24] G. Citerio, S. Park, J. Schmidt, R. Moberg, J. Suarez, *et al.*, "Data collection and interpretation," *Neurocritical care*, 2015.

[25] S. Rajkumar, R. Kyle, *et al.*, "… in the diagnosis, classification, risk stratification, and management of monoclonal gammopathy of undetermined significance: Implications for recategorizing disease …," in *Mayo clinic proceedings*, 2010.

[26] G. Kaplan, M. Haan, and R. Wallace, "Understanding changing risk factor associations with increasing age in adults," *Annual Review of Public Health*, 1999.

[27] A. Konstantinov and L. Utkin, "Interpretable machine learning with an ensemble of gradient boosting machines," *Knowledge-Based Systems*, 2021.

[28] A. Assis, J. Dantas, and E. Andrade, "The performance-interpretability trade-off: A comparative study of machine learning models," *Journal of Reliable Intelligent Environments*, 2025.

[29] K. Iba, T. Shinozaki, K. Maruo, and H. Noma, "Re-evaluation of the comparative effectiveness of bootstrap-based optimism correction methods in the development of multivariable clinical prediction models," *BMC Medical Research Methodology*, 2021.

[30] A. Alshehri, "Metabolic syndrome and cardiovascular risk," *Journal of Family and Community Medicine*, 2010.

[31] P. Kuri-Morales, J. Emberson, J. Alegre-Díaz, *et al.*, "The prevalence of chronic diseases and major disease risk factors at different ages among 150 000 men and women living in mexico city: Cross-sectional analyses of …," *BMC Public Health*, 2009.

[32] Q. Yu, Z. Hou, and Z. Wang, "Predictive modeling of preoperative acute heart failure in older adults with hypertension: A dual perspective of SHAP values and interaction analysis," *BMC Medical Informatics and Decision Making*, 2024.

[33] T. Nakagami and D. S. Group, "Hyperglycaemia and mortality from all causes and from cardiovascular disease in five populations of asian origin," *Diabetologia*, 2004.

[34] E. Rapsomaniki, A. Timmis, J. George, *et al.*, "Blood pressure and incidence of twelve cardiovascular diseases: Lifetime risks, healthy life-years lost, and age-specific associations in 1· 25 million people," *The Lancet*, 2014.

[35] Z. Gao, Z. Chen, A. Sun, and X. Deng, "Gender differences in cardiovascular disease," *Medicine in Novel Technology and Devices*, 2019.

[36] L. Ayerbe, E. González, V. Gallo, C. Coleman, *et al.*, "Clinical assessment of patients with chest pain; a systematic review of predictive tools," *BMC Cardiovascular Disorders*, 2016.

[37] R. Gianrossi, R. Detrano, D. Mulvihill, K. Lehmann, *et al.*, "Exercise-induced ST depression in the diagnosis of coronary artery disease. A meta-analysis." *Circulation*, 1989.

[38] S. Ahmed, "Comparative analysis of explainable models for clinical decision support system," doria.fi, 2025.

[39] D. Kent, E. Steyerberg, and D. V. Klaveren, "Personalized evidence based medicine: Predictive approaches to heterogeneous treatment effects," *Bmj*, 2018.

[40] S. Band, A. Yarahmadi, C. Hsu, M. Biyari, *et al.*, "Application of explainable artificial intelligence in medical health: A systematic review of interpretability methods," *Informatics in Medicine Unlocked*, 2023.

[41] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016.

[42] M. Khalilia, S. Chakraborty, and M. Popescu, "Predicting disease risks from highly imbalanced data using random forest," *BMC Medical Informatics and Decision Making*, 2011.

[43] T. Mendoza, C. Lee, C. Huang, and T. Sun, "Random forest for automatic feature importance estimation and selection for explainable postural stability of a multi-factor clinical test," *Sensors*, 2021.

[44] Q. Li, "Patient classification with ensemble treebased modelling for decision support in acute clinical care settings," research-repository.rmit.edu.au, 2024.

[45] Z. Liu and X. Huang, "Explaining individualized treatment rules: Integrating LIME and SHAP with xgboost in precision medicine," *Statistics in Medicine*, 2025.

[46] R. Mitchell, E. Frank, and G. Holmes, "GPUTreeShap: Massively parallel exact calculation of SHAP scores for tree ensembles," *PeerJ Computer Science*, 2022.

[47] C. Ihejirika, "Explainable artificial intelligence in deep learning models for transparent decision-making in high-stakes applications," *Int J Res Publ Rev*, 2025.