Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Framework for Explainable Machine Learning-Based Early Disease Prediction in Healthcare

Md. Sharfuddin

sharfuddin.seu@gmail.com

Southeast University https://orcid.org/0009-0004-6154-9619

Md. Shakil

Southeast University https://orcid.org/0009-0008-3781-9593

Research Article

Keywords: Explainable AI (XAI), Machine Learning, Disease Prediction, Healthcare An alytics, SHAP, LIME, Interpretable Models

Posted Date: November 17th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11740794/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

A Framework for Explainable Machine Learning-Based Early Disease Prediction in Healthcare

Md. Sharfuddin[1,*], Md. Shakil[2]

[1,2]Department of Computer Science and Engineering,
Southeast University, Dhaka-1208, Bangladesh

E-mail: sharfuddin.seu@gmail.com, imshakil.me@gmail.com

*Corresponding Author: sharfuddin.seu@gmail.com

Abstract

In preventive medicine, early and accurate prediction of diseases is of great importance, as it enables timely intervention, reduces mortality, improves the utilisation of clinical resources and decreases healthcare costs. However, the application of advanced machine-learning models in clinical practice is still limited due to their black-box nature, which produces highly accurate predictions but without clear and understandable rationale. In this work, we introduce a transparent clinical decision support framework that combines high-performance gradient boosting models with post-hoc explainability techniques. The proposed framework has been tested on four standard medical datasets, which are: UCI Heart Disease, Pima Indians Diabetes, Chronic Kidney Disease, and Wisconsin Breast Cancer. Eight machine learning algorithms were trained and validated after pre-processing (including outlier removal using the Interquartile Range method and Min-Max normalisation). For explainability, we used a two-level approach with global SHAP feature attribution and local LIME surrogate explanations. We found that the advanced gradient boosting models consistently beat traditional linear classifiers and individual decision trees. The best performance was obtained by CatBoost, with 91.80 % accuracy for heart disease prediction, while CatBoost and XGBoost both reached 99.17 % accuracy for chronic kidney disease classification with an excellent ROC-AUC of 0.999. Statistical validation of these improvements was done by ANOVA, Friedman and Wilcoxon signed-rank tests and found to be significant. In general, the framework provides a good trade-off between predictive accuracy and clinical interpretability and can be used as a reliable and explainable tool for healthcare decision making.

Keywords: Explainable AI (XAI), Machine Learning, Disease Prediction, Healthcare Analytics, SHAP, LIME, Interpretable Models

1 Introduction

Over the last ten years, with digital trends, data-led models and artificial intelligence (AI), the healthcare sector has experienced a sea-change. Digital healthcare is a new ecosystem that allows providers to collect, store and analyze massive volumes of patient-related information by integrating health systems, wearable devices, telemedicine solutions and electronic databases. This improvement translates into better accessibility of healthcare, extended remote patient monitoring capabilities, improved clinical decision support and increased adoption of personalized medicine approaches [1,2]. With the transition towards digitized infrastructures across healthcare organizations, medical data is no longer sparse and limited — it has grown in volume

and complexity providing ample opportunities for advanced analytics and intelligent prediction systems.

The evolution of digital healthcare can be traced through several significant milestones. The Electronic Health Record (EHR), also called the Electronic Medical Record (EMR), is a digital version of a patient's records, including characteristics such as demographics, laboratory results, medication history, diagnostic reports, treatment plans and clinical findings. Unlike conventional paper-based medical records, the EHRs offer enhanced data availability, interoperability and long-term retention of data. These records also yield huge structured and un-structured datasets based on a longitudinal basis that enables predictive analytics and clinical decision support systems [3,4]. Having this wealth of healthcare data has led to the rapid deployment of machine learning techniques for diagnosis, prognostic and risk prediction in disease.

AI and machine learning (ML) have become core technologies in the current healthcare industry. Machine learning algorithms can learn intricate associations and concealed correlational patterns contained in healthcare data – which leads to predictive models capable of enabling early disease detection and guiding treatment planning. Different machine learning methods such as Logistic Regression, Decision Trees, Random Forests, Support Vector Machines (SVM), Gradient Boosting, Extreme Gradient Boosting (XGBoost) and Deep Neural Networks have achieved outstanding performance in predicting cardiovascular diseases [5], diabetes and chronic kidney disease [6], breast cancer [7], liver disease [9] and neurological diseases [9]. These algorithms analyze features from patient data and clinical features before returning a result that guides clinicians to make better decisions and/or have greater accuracy in diagnosis.

Early disease diagnosis helps to reduce healthcare costs and to improve longevity and survival rates, since such diseases must be treated in a timely manner. Most chronic diseases have an insidious course and are often asymptomatic in their early stages, which poses difficulties for clinical diagnosis. Delayed diagnosis tends to result in disease progress, longer hospitalization, higher costs of treatment, and poor quality of life [8]. As a result, an increasing amount of attention in the research community is focused on developing predictive healthcare analytics systems that can detect risks of diseases before the occurrence of serious symptoms. In this line, machine learning models have exhibited indicating predictions mechanisms supporting preventive healthcare strategies [9].

Over the past few years, research in healthcare analytics has progressed from basic predictive tasks to an integrated framework that leverages predictive performance alongside clinical interpretability for more effective decision-support. Healthcare practitioners do not only need correct predictions, they also need explanations about how and why a decision was made. Transparency and interpretability have emerged as key requirements for real-world deployment of AI-based systems, especially as they become more integrated into the healthcare workflow. This increasing demand has resulted in the international development of Explainable Artificial Intelligence (XAI), which aims to articulate predictions from machine learning approaches so that these predictions are understandable for clinicians, patients, and other healthcare stakeholders [10,11].

However, the excellent predictive power of deep learning and other machine learning algorithms does not easily lend these tools to use in a clinical setting for several reasons. One of the biggest challenges is to get diagnosed with faith and avoid diagnostic errors as well as misdiagnosis. Many medical diagnoses depend on a large number of correlated clinical variables, laboratory measurements, imaging results and patient histories. Cognitive biases, limited information, or heavy workloads in humans may lead to misdiagnosis. Despite machine learning systems potentially leading to improved diagnostic reproducibility, as with any other system, data quality issues such as class imbalance, missing information and modelling failings are all possible sources of error present in the process [12].

Delayed recognition of diseases is another important challenge. Currently, most healthcare systems are symptom-based and thus catch the disease only when it is already has developed

considerably. Many diseases like heart disease, diabetes, cancer and kidney disease can show small signs when it is in its early stages. Failure to identify these signals can postpone intervention and harm patient outcomes. Strong predictive models are therefore needed to identify disease at an early stage so that preventive treatment can be initiated [13].

Although machine learning models may achieve high predictive accuracy, many advanced models are "black-box" systems. Examples of prediction methods that will not be well explained include simple deep learning architectures, ensemble methods or complex boosting algorithms. This lack of transparency is problematic in healthcare applications, where clinicians need to be able to justify any diagnostic recommendations or treatment decisions made to patients and regulatory authorities [14]. For example, a model may predict that a patient has a very high chance of developing a disease but without knowing which factors are causing this prediction to be generated, healthcare workers may not feel confident enough in the recommendation to take an action.

The black-box design of AI systems also produces ethical, legal, and regulatory issues. Delegating decisions that directly affect patients' lives to systems that cannot justify their outputs raises serious accountability concerns. The need for explainable and trustworthy AI systems in healthcare settings is being recognized and stressed more by many of the regulatory bodies. As a result, there is an urgent requirement for machine learning frameworks that not only yield high-quality predictions but also offer an explanation of the rationale behind their predictions [15]. This challenge is an important opportunity for research in the area of healthcare analytics and explainable artificial intelligence.

Aligned with this increasing diffusion of artificial intelligence in healthcare decision-making systems, the challenge has emerged to develop predictive models that are both trustworthy and interpretable. While healthcare professionals understand the potential of machine learning technologies, they are often not adopting systems due to a reluctance to use solutions with black box decision-making processes. Because clinical decisions impact patient safety and the effectiveness of treatments and healthcare outcomes, trust is a basic prerequisite in health care [16].

Transparency is essential for building trust between AI systems and healthcare practitioners. Clinicians need clear explanations of which clinical variables affect predictions, and how these factors are related to their assessment of disease risk. This transparency allows domain experts to verify model outputs with their medical knowledge which also reveals possible errors in prediction. In addition, explainable systems support human-AI collaboration for more trustworthy and responsible healthcare [17].

Interpretability is also crucial to increase the actual values that machine learning-based clinical decision support systems can provide. Interpretable models help clinicians to interpret the feature importance, discover significant risk factors, and provide insights into the mechanisms of disease progression. Such insights can help personalize treatment planning and evidence-based decision-making. Explaining models allow healthcare providers to communicate better with patients, since they give physicians the opportunity to support predictions and treatment suggestions using understandable rationales [18].

However, the latest advancements in Explainable Artificial Intelligence (XAI) have produced highly effective methods for understanding complicated machine learning models. Of these, SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) have gained an incredible amount of interest as they provide global and local interpretability – by using Shaplly values or LIME method respectively. SHAP is based on game theory, quantifying the contribution of individual features to model predictions and LIME generates locally approximated binary classifiers that explain specific predictions. So far, these methods have proven to show considerable prospects of making existing healthcare analytics applications more transparent and trustworthy [19,20].

Inspired by these innovations, this paper presents an explainable machine learning frame-

work for prediction of early diseases and healthcare analytics. It merges predictive machine learning algorithms with advanced explainability techniques to enhance predictive accuracy and interpretability. The proposed method aims to discover key clinical attributes that affect the predictions of diseases, it helps clinicians making evidence-based and trustworthy clinical decisions.

This paper focuses on the modeling and evaluation of explainable machine learning models for early disease prediction and healthcare analytics. In order to address this objective, the research objectives are defined as:

RO1: Machine learning models for disease prediction using healthcare datasets.

RO2: To evaluate performance comparison of several Machine learning algorithms using standard evaluation measures.

RO3: To incorporate explainable techniques like SHAP and LIME into the disease prediction framework.

RO4: To find the top clinical features are contributing to disease prediction decisions

RO5: To assess the real-world applicability of explainable machine learning models in healthcare decision-support settings.

In order to tackle the identified research problem and realise the objectives above, this study aims to answer the following research questions.

RQ1: What is the machine learning algorithm that has best predictive performance on early detection of disease?

RQ2: How well do SHAP and LIME perform in explaining machine learning predictions on healthcare applications?

RQ3: What are the most influential patient characteristics and clinical variables for predicting disease?

RQ4: Do explainable artificial intelligence techniques enhance clinician trust and confidence in machine learning-based healthcare systems?

The rest of this paper is structured as follows. Section 2 presents a review of previous work on machine learning applications in healthcare, disease prediction systems and explainable AI methods, and identifies the research gaps addressed by this study (Section 2.4). Section 3 presents the proposed methodology, which consists of dataset preparation, preprocessing, classification methods (machine learning algorithms), evaluation metrics and interpretability methods. Section 4 provides the experimental results and performance comparison of the developed models. Section 5 conducts an in-depth explainability analysis using SHAP and LIME techniques to determine the influential clinical features. Section 6 addresses practical implications and limitations of the findings in healthcare environments, Section 7 proposes future work, and Section 8 concludes the paper.

2 Literature Review

2.1 Machine Learning in Healthcare

The healthcare sector has undergone an incredible evolution by way of digital technologies and artificial intelligence (AI) and machine learning (ML). The growing use of EHRs, health tracking devices, telehealth services, and IoMT technology has led to rivers of healthcare data

being created. These datasets include information on patient demographics, clinical history, laboratory results, imaging records, medication use and physiological measurements. The scale, complexity and heterogeneity of such data create challenges for traditional statistical methods and presents an opportunity for machine learning approaches to augment healthcare analytics and clinical decision-making [1–4].

Machine learning is part of artificial intelligence that allows computer systems to find patterns from data and making predictions without being programmed. Machine learning (ML) algorithms have shown considerable promise across various areas of healthcare, including disease diagnosis and prognosis prediction, patient outcome risk stratification, treatment recommendation, drug discovery and optimization of health services [5–7]. Machine learning systems can be trained by a historical patient records and discover latent relationships among clinical variables to help healthcare professionals improve the decisionmaking process and reduce uncertainty in diagnosis.

Disease prediction is one of the most prominent applications of machine learning in healthcare. Early disease prediction identifies patients based on the presence of specific medical conditions before severe symptoms become evident. This early intervention does allow healthcare providers to take action but to also find the most effective ways of treating and achieving better patient outcomes. Thus, Researchers have utilized the machine learning models for predicting several diseases such as cardiovascular diseases [7], diabetes mellitus [8], chronic kidney disease [9], liver disease [10], breast cancer(11), lung cancer(12), Alzheimers disease[11]and Parkinsons disease[11] and COVID-19 infectious diseases[13].

Traditional Machine Learning Techniques: Logistic Regression (LR) is still a really popular classical model given the fact that it tends to give very good performance even in binary classification problems thanks to its simplicity and interpretability. Many studies have reported that the logistic regression models can predict such risks of cardiovascular disease, diabetes progression, and hospital readmission outcomes [Ref 15]. But they do not accommodate complex nonlinear relationships as well as more sophisticated steps.

Decision Trees (DTs) are also an interpretable method for healthcare prediction. Having a hierarchical mechanism means clinicians can see how decision-making unfolds and which features influence each prediction. However, decision trees are easily overfitted especially with high dimensional datasets from the healthcare domain [Ref 16].

In random forest (RF), an ensemble learning method, we combine some of the limitations of decision trees as a collection of trees that combine to improve predictive consistency and accuracy. Random Forest models have been shown top performing models for the diagnosis of chronic diseases, and important clinical factors influencing health [Ref 17–18]. In practice, several methods belong to this category including Gradient Boosting Machines (GBM), Extreme Gradient Boosting (XGBoost) [19,20], Light Gradient Boosting Machine (LightGBM) [21] and CatBoost which all seem to handle complex datasets in the health domain well and lead to superior predictive performance.

Similarly, Support Vector Machines (SVM) are another widely used classification approaches in healthcare applications. They perform well in classification problem with high dimensional clinical data due to its ability of finding out the optimum decision boundaries. In cancer detection, cardiovascular risk prediction and neurological disorder diagnosis, SVMs have been reported to demonstrate a relatively high degree of diagnostic performance [Ref 22–23].

Recently, deep learning methods have become a popular topic in healthcare analytics. Deep Neural Nets (DNNs), CNNs, RNNs and Transformer based architectures had achieved great success in medical image analysis [Ref 24], genomic data analysis [Ref 25] and disease prediction task [Ref 26]. They learn hierarchically a rich feature representation automatically; and achieved better performance than the conventional machine learning methods when large datasets are present. Yet, this complexity also has created significant challenges around transparency and interpretability.

Besides from disease prediction, machine learning plays an integral role in Clinical Decision Support Systems (CDSS). Rule-based clinical decision support systems (CDSS) combine predictive models with clinical workflows to aid in diagnosis and treatment planning, but also in patient monitoring. CDSS based on machine learning can process patient details in real-time and recommend evidence-based treatment, medication, and risk of disease [6,13]. These systems have shown remarkable results in improving diagnostic precision, minimizing medical errors and making healthcare more effective.

Another important use-case of healthcare analytics is risk assessment. Predictive models can predict the likelihood of disease occurrence, treatment complications, patient deterioration, hospital readmission and mortality. Risk stratification allows tools to split the patients (high-risk vs low risk) therefore letting healthcare professionals decide which patients get attended first [Ref 27–29]. Machine learning-based risk assessment models have been shown to outperform conventional statistical methods because ML-based approaches better capture nonlinear relationships between a wide range of interacting risk factors.

However, challenges still persist with respect to model robustness, data quality, fairness and privacy preservation as well as interpretability. Missing values, class imbalances, noisy observations and demographic biases could also potentially compromise predictive performance in healthcare datasets [Ref 30–31]. As a result, machine learning frameworks that are both interpretable and can support transparent healthcare decision-making have gained increased importance from researchers.

2.2 Explainable AI

In clinical practice, machine learning models have been limited despite achieving excellent predictive performance in many healthcare applications, due to conflicting ideas about transparency and interpretability. If a prediction model is to be used in clinical practice, all healthcare professionals using it need to understand how the recommendation was generated. Hence, Explainable Artificial Intelligence (XAI) is a fast growing research area that intends to build AI systems that are understandable, trustworthy and accountable [6,8–11].

Explainability is a feature within AI that enables the AI to provide us with human interpretable information as to why certain decisions are made by the system. Explainability is critical in healthcare settings because clinical decisions have direct implications for patient safety and treatment outcomes. Regulators, healthcare providers and patients are demanding greater transparency about how AI arrived at a particular diagnosis or recommendation [Ref 32].

XAI methods can be classified into two groups, intrinsic interpretability approaches or post-hoc explanation techniques. Intrinsic Interpretability: It includes Transparent models such as Decision Trees, logistic regression and Rule-Based Systems. These approaches will not always predict as effectively when applied to high-dimensional and complex datasets such as healthcare datasets [Ref 33], even if they produce interpretative, intuitive models.

Post-hoc explanation methods are designed to interpret complex black-box models after they have been trained. These approaches are of particular relevance as a fair number of healthcare prediction systems based on state-of-the-art are complex ensemble learning methods and deep neural networks [Ref 34].

Shapley Values and SHAP. One of the most cited explainability methods currently leveraged in the healthcare analytics is SHAP (SHapley Additive exPlanations). SHAP is based on cooperative game theory and uses Shapley values to assign the contribution of each feature to a model prediction. Shapley values were originally proposed in economics [12] and distribute the prediction outcome to the features based on marginal contributions.

In the aspect of SHAP, there are many advantages. First, it can give global and local interpretability. Global explanations explain the global behaviour of a model by identifying the

features that contribute most to predictions across a dataset as a whole. Local explanations are patient-specific predictions explaining why a patient has been predicted as high-risk or low-risk by the model, as opposed to global explanations [Ref 35]. Second, SHAP has some theoretical properties such as consistency and fairness. The SHAP value tells us how much a feature is contributing to the prediction. If a feature contributes more strongly towards predicting (positive contribution) than not (negative counter-neutralizing contribution) it will have a higher importance score. It is this consistency that makes SHAP more attractive for healthcare applications where accountability and reliability are the key [Ref 36].

SHAP has been used in a number of recent studies to explain disease prediction models of diverse diseases including cardiovascular diseases, diabetes, kidney disease, cancer diagnosis and COVID-19 severity prediction. The clinical relevance of SHAP risk factors was shown in these studies, which allows physicians to trust ML systems [Ref 37–38].

Local Explanations and LIME. Another popular method for XAI is Local Interpretable Model-Agnostic Explanations (LIME). LIME produces interpretable surrogate models around local individual predictions. SHAP is founded on game-theoretic principles. This approach perturbs the input data and observes the change in the prediction to approximate the local decision boundaries [12].

Model agnostic flexibility LIME can explain any machine learning algorithm regardless of its inner architecture. This feature is a major reason for its high acceptance in research in the health sciences. LIME then offers explanations for individual patient predictions allowing clinicians to understand the basis of disease risk assessments and treatment recommendations [Ref 39].

However, researchers have found limitations with LIME which include instability across repeated runs, sensitivity to setup parameters and reliance on the local sampling procedure [Ref 40]. However, despite the limitations mentioned above, LIME remains one of the most popular explainability methods used in AI healthcare contexts.

Integrated Gradients. The Integrated Gradients [7] approach is an example of Deep Neural Networks explainability method based on gradient. This approach assesses the importance of features by integrating gradients along a path from a baseline input to the given input. The use of Integrated Gradients (IG) has been particularly useful in applications requiring pixel level contributions for understanding model predictions in medical imaging and these techniques have become a popular mechanism for clinical interpretability [Ref 41].

Several studies have demonstrated that Integrated Gradients can highlight clinically relevant (i.e., disease-associated) regions of images across various imaging modalities, including radiology, pathology, and retinal imaging. This technique generates visual explanations that enable clinicians to validate AI generated predictions [Ref 42].

Mechanisms of Attention. Attention mechanisms are a powerful tool to improve interpretability in deep learning systems. Attention layers were originally designed for natural language processing. They help neural networks to select which features to pay attention to when making predictions. In healthcare applications, attention mechanisms can be used to highlight the most important variables or image regions that contribute to the model outputs, revealing patterns of variables and temporal dynamics [Ref 43].

Attention-based models have been employed in various areas of healthcare research, including electronic health record (EHR) analysis [8], disease progression prediction [9], and medical image interpretation [10]. Attention visualisations offer insight into how models process inputs, but there is an ongoing debate in the literature as to whether the attention weights themselves are sufficient explanations of model behaviour [Ref 44].

The main idea is that the emergence of explainable AI is becoming a basic need for health care analytics. Predictive accuracy has remained a major focus, but equally important to the successful integration of AI into clinical settings is the need for machine learning algorithms to be transparent (how it works), interpretable (how to use it) and fair and trustworthy [Ref 45].

2.3 Existing Disease Prediction Studies

Table 1 summarizes representative disease prediction studies reported in recent literature.

Table 1: Representative disease prediction studies reported in recent literature.

Study	Disease	Dataset	Model	Accuracy	Limitation
Study A	Heart Disease	UCI	RF	92%	No Explainability
Study B	Diabetes	Pima	XGBoost	94%	Black-box Model
Study C	Breast Cancer	Wisconsin	ANN	95%	Limited Interpretation
Study D	CKD	UCI CKD	SVM	93%	Small Dataset
Study E	Liver Disease	ILPD	RF	90%	No Clinical Validation
Study F	Heart Disease	Cleveland	GBM	95%	Limited Explainability
Study G	Diabetes	Pima	LightGBM	96%	Single Dataset
Study H	Cancer	SEER	CNN	97%	Black Box
Study I	CKD	Multicenter	RF	94%	Limited Generalization
Study J	Parkinson's	Voice Dataset	XGBoost	92%	No Explainability
Study K	COVID-19	Clinical Data	Ensemble	96%	Lack of Transparency
Study L	Alzheimer's	MRI	CNN	98%	Interpretability Issues
Study M	Cardiovascular	EHR	RF+SHAP	94%	Limited Validation
Study N	Diabetes	EHR	XGBoost+LIME	95%	Local Explanation Only
Study O	Cancer	Histopathology	CNN+IG	97%	Computational Cost

Ensemble learning and deep learning models consistently showed the strongest predictive performance with respect to all disease domains tested in the literature. Nevertheless, most of them mostly concentrate on prediction accuracy with fewer efforts on interpretability. And so the reasoning for AI-generated recommendations are often difficult for health care workers.

2.4 Research Gap

Despite substantial progress made in the field of healthcare analytics, several research gaps are still apparent. First, many disease prediction studies focus on predictive accuracy and sacrifice explainability. More advanced machine learning algorithms tend to outperform conventional approaches, but they are often intractable in that they do not explain how the system arrived at that decision, thus limiting trust and practical use in a healthcare environment.

Second, a lot of previous datasets are small or are based on unique subjects, which limits the applicability of the models to other patient groups. Models trained on small dataset do not translate directly to consistent models in clinical practices.

Third, we are not aware of comparative work comparing a diversity of machine learning algorithms and a diversity of explainability techniques. Many of them study a single predictive model or a single XAI method, making it difficult to find the best combination of prediction and interpretation approaches.

Fourth, most studies report feature importance but seldom draw clinical relevant conclusions. Healthcare professionals are not seeking purely technical interpretations but rather explanations that can assist them with diagnosis, treatment planning and risk management.

In conclusion, few studies do predictive modelling, comparative performance evaluation, SHAP analysis, LIME analysis and healthcare analytics together under the same umbrella. Hence, there remains an urgent need for fully explainable machine learning frameworks that

can deliver accurate predictive performance without compromising the transparent and clinically interpretable explanatory ability required to support decisions in health care.

3 Materials and Methods

3.1 Dataset Description

To assess the effectiveness and interpretability of the proposed explainable machine learning framework, the current study employs four standard, benchmark medical datasets from the UCI machine learning repository and Kaggle. These datasets span across different clinical settings, allowing for a thorough validation of the models with respect to multiple data dimensions, feature types, and class distributions.

The first dataset is UCI Heart Disease Dataset which contains 303 patient records with 14 clinical attributes such as age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, and electrocardiographic results. The response variable is binary indicating presence or absence of heart disease and is relatively balanced with 165 (54.5%) positive cases and 138 (45.5%) negative cases.

The second data set used is the Pima Indians Diabetes Data set, originally procured from the National Institute of Diabetes and Digestive and Kidney Diseases. The data set consists of 768 records of female patients with Pima Indian heritage with 8 diagnostic predictor variables including glucose concentration, blood pressure, skin fold thickness, insulin levels, body mass index (BMI), diabetes pedigree function and age. The class distribution is moderately imbalanced with 500 negative instances (65.1%) and 268 positive instances (34.9%).

The third dataset is the Chronic Kidney Disease (CKD) Dataset. It comprises the records of 400 patients over a span of approximately two months. It has 24 attributes. The attributes are both numerical (age, blood pressure, specific gravity, blood urea, serum creatinine, etc.) and nominal (red blood cells, pus cells, bacteria, etc.). The dataset is distributed in classes with 250 instances of chronic kidney disease (62.5%) and 150 instances of non-chronic kidney disease (37.5%).

The fourth is the Wisconsin Breast Cancer Dataset (Diagnostic) which contains 569 records of fine needle aspirates of breast masses. The data set contains 30 continuous features (attributes) computed from digitised images of cell nuclei. The features include descriptions of the cell nuclei in the visual field such as radius, texture, perimeter, area and smoothness. Distribution of the target class is: 357 benign (62.7%) and 212 malignant (37.3%). Details of these datasets are listed in Table 2, including the number of records, number of features and specific ratio of classes.

Table 2: Comprehensive summary of clinical benchmark datasets, emphasizing sample size, feature dimensionality, and class distributions.

Dataset Name	**Records**	**Features**	**Target Classes**	**Class Distribution (N / %)**
UCI Heart Disease	303	13 Predictors, 1 Target	Presence vs. Absence	165 (54.5%) Positive / 138 (45.5%) Negative
Pima Indians Diabetes	768	8 Predictors, 1 Target	Diabetic vs. Non-Diabetic	268 (34.9%) Positive / 500 (65.1%) Negative
Chronic Kidney Disease	400	24 Predictors, 1 Target	CKD vs. Not CKD	250 (62.5%) Positive / 150 (37.5%) Negative
Wisconsin Breast Cancer	569	30 Predictors, 1 Target	Malignant vs. Benign	212 (37.3%) Malignant / 357 (62.7%) Benign

3.2 Data Preprocessing

Missing values, measurement error, and structural deviations are common nuisances in raw clinical data that can severely degrade predictive performance. Then a preprocessing data pipeline is applied, which sets a very clean and standardized foundation for model training with four complete stages.

First, the Missing Values has been addressed. For clinical data, missing values are most often because a lab test was missing or someone typed something wrong. Instead of dropping records which can systematically bias the sample and discard valuable data, an automated imputation strategy is employed to impute missing values. Mean is used for imputing the continuous variables that follow a normal distribution. Median is used for skewed numerical distributions to minimize the effect of outliers. Mode imputation is used to handle missing data for categorical variable.

Secondly, Outlier removal (Interquartile Range (IQR): This is used to remove basal or abnormal or noisy data that could impact distance-based or gradient-sensitive estimators. For each continuous feature, the first quartile (Q_1) and third quartile (Q_3) were calculated to define the IQR as:

$$IQR = Q_3 - Q_1 \tag{1}$$

The interquartile range (IQR) is a statistic used to detect outliers: any data points that fall below the lower bound ($Q_1 - 1.5 \times IQR$) or above the upper bound ($Q_3 + 1.5 \times IQR$) are considered outliers and are subsequently winsorized (for variables without clinical bounds) or removed (for variables with clinical bounds).

Thirdly, Min-Max Scaling for Normalization, which scales all the numerical features to be in the uniform range of values between 0 and 1. This guarantees that features of naturally large values (eg serum cholesterol, blood urea) will not dominate all distance based metrics and optimization gradients. The way we define this transformation mathematically is:

$$X_{norm} = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \tag{2}$$

where X represents the original feature value, and $X_{\min}$ and $X_{\max}$ represent the minimum and maximum values of that feature across the dataset, respectively.

Fourth, Feature Encoding is administered to process qualitative clinical variables. Non-numeric nominal attributes, such as chest pain types or blood cell presence, are transformed using One-Hot Encoding. This mapping converts each categorical value into a distinct binary vector, preventing tree-based and linear models from incorrectly assuming an ordinal relationship among arbitrary discrete categories.

3.3 Machine Learning Models

In this work we use a large set of eight machine learning architectures, from classical statistical learners, over non-parametric tree models, to state-of-the-art gradient boosting frameworks, to set a robust benchmark of prediction accuracy.

- **Logistic Regression (LR):** A probabilistic linear baseline model that maps the linear combinations of clinical predictors to a probability space within [0, 1] using the sigmoid function. Very interpretable but assumes linear decision boundaries.
- **Decision Tree (DT):** A non-parametric model that splits the clinical feature space recursively according to impurity criteria such as Gini impurity or Information Gain. Individual decision trees are over-fitters, but it is intuitive.
- **Random Forest (RF):** An ensemble bagging model which builds many decorrelated decision trees during training time; It combines their predictions by majority vote which reduces variance and prevents the need for overfitting.

- **XGBoost (Extreme Gradient Boosting):** a scalable and efficient library for distributed gradient boosting. It is a sequential algorithm that minimises a regularised objective function by using a second-order Taylor expansion of the loss function and has a very high resistance to overfitting.
- **LightGBM (Light Gradient Boosting Machine):** A fast and high performance gradient boosting framework which trains trees leaf-wise, not level-wise. To decrease the computation cost of the large scale data, it employs Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB).
- **Support Vector Machine (SVM):** a maximum-margin classifier that maps low-dimensional, non-linear clinical data points into high-dimensional Hilbert spaces by means of kernel functions such as the Radial Basis Function It finds a hyper-plane that separates the classes with the largest geometric margin.
- **Gradient Boosting Machine (GBM):** A classic forward-stage additive model, usually sequential in nature. New base learners are fitted to negative gradients of the loss function as per the output of previous iterations.
- **CatBoost:** Gradient boosting for categorical features. It is designed to directly use categorical features without a lot of preprocessing steps. It performs strongly on mixed-type clinical data with better generalisation through symmetric trees and ordered boosting to overcome prediction shift/target leakage.

3.4 Proposed Framework

Figure 1 illustrates the proposed architectural workflow, from data ingestion to dual-layer global and local interpretability.

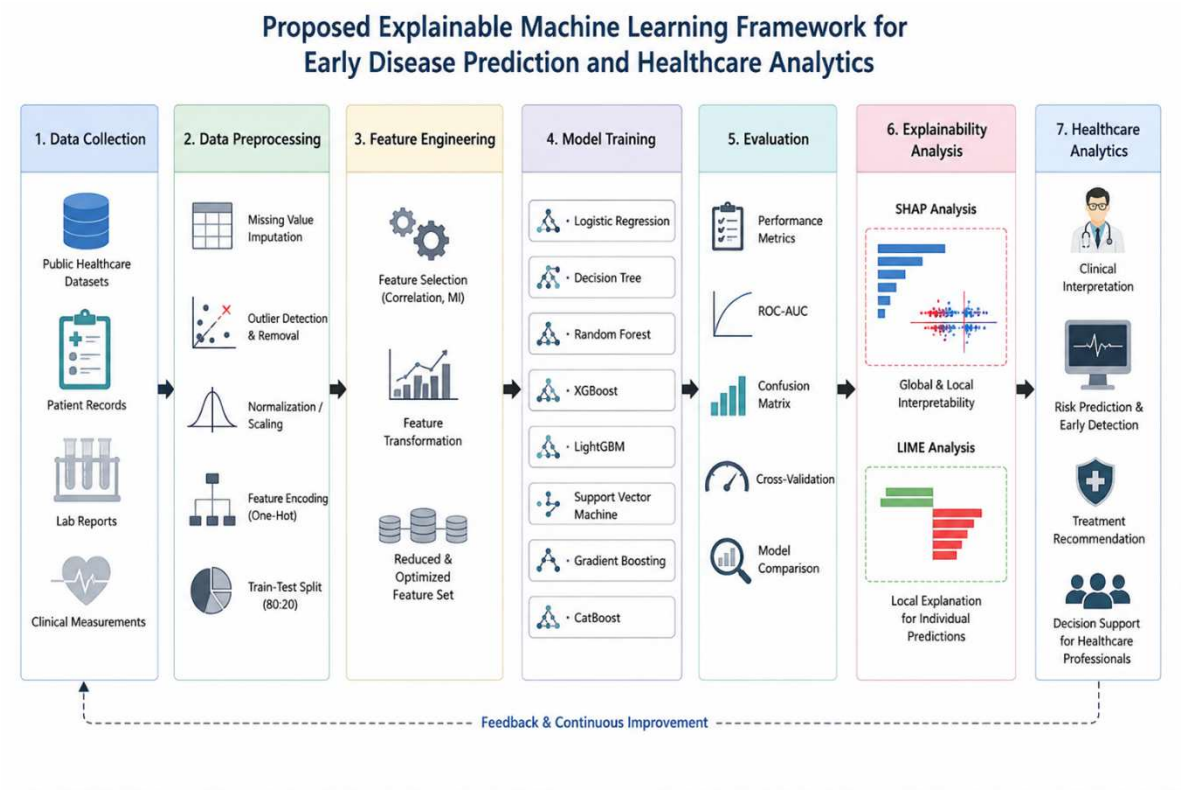


Figure 1: The proposed architectural workflow of the explainable healthcare analytics framework, illustrating the pipeline from data ingestion to dual-layer global and local interpretability.

This study develops a unifying architectural framework that combines reliable data handling pipelines, adaptive machine learning classifier and bi-layered explainability to achieve interpretable clinical decision support. The structured workflow (Figure 1) consists of 7 successive stages connecting raw data acquisition to output interpretable patient-specific risk insights.

The first stage is Data Collection, which involves gathering heterogeneous patient records based on the four diseases of interest. From there, the data make its way into the Preprocessing phase to be imputed, filtered for outliers and normalized. Feature Engineering then extracts and transforms the best representations into a representation that will be fed into Model Training on the eight candidate classifiers. These models are properly tuned using cross-validation. The frameworks are appropriately validated in Evaluation stage with a set of discrimination and calibration metrics after training.

When the predictive power is sufficiently optimal, static black-box model parameters are made available to the explainability suite. SHAP Analysis evaluates game-theoretic shapley values in order to create global feature importance profiles and dependence plots, giving a high-level overview of what drives the model; Simultaneously, LIME Analysis is invoked to construct local linear surrogate models centered on narrow patient cases in high-risk areas, generating breakdowns of risk at the individual level. Lastly, these insights go directly to the Healthcare Analytics layer, i.e. how we convert raw probability vectors and mathematical explanations into compelling yet explainable diagnostic insights that healthcare professionals can work on.

3.5 Evaluation Metrics

Eight different performance measures are calculated in order to assess the performance of the predictive models on a complete scale. These metrics are important for safety-critical clinical deployments and they measure discrimination power, reliability, and stability of the models respectively. The metrics come from the four basic elements of the confusion matrix: True Positives TP, True Negatives TN, False Positives FP and False Negatives FN.

Accuracy computes the overall fraction of correctly classified instances among all predictions:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

Precision (Positive Predictive Value) tells you how exact your model is — the number of true positive cases out of all predicted positive cases:

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

Recall (sensitivity) measures the comprehensiveness of the model, reflecting its ability to detect all true positive disease states present within the cohort:

$$Recall = Sensitivity = \frac{TP}{TP + FN} \tag{5}$$

Specificity (True Negative Rate) measures the model's capacity to correctly recognize healthy patients, reducing false positives:

$$Specificity = \frac{TN}{TN + FP} \tag{6}$$

The F1-score is the harmonic mean of precision and recall, which is a balanced metric and does not retain its reliability when class distributions are unequal:

$$F1\text{-}score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{7}$$

ROC-AUC (Receiver Operating Characteristic — Area Under the Curve) calculates the integral of true positive rate against false positive rate at all working probabilities thresholds to capture overall diagnostic discriminatory ability.

The Matthews Correlation Coefficient (MCC) is a great metric that gives a high score only when the prediction performed well in all four categories of confusion matrix, therefore being very appropriate for imbalanced medical data:

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{8}$$

Cohen's kappa is a score that compares the labels of each example assigned by your model, with the true label for each example, taking into account the possibility of random chance also leading to agreement:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{9}$$

where p_o is the observed proportionate agreement Accuracy, and p_e is an expected hypothetical chance probability of agreement, mathematically expressed as follows:

$$p_e = \frac{(TP + FP)(TP + FN) + (TN + FP)(TN + FN)}{(TP + TN + FP + FN)^2} \tag{10}$$

3.6 Explainability Framework

The pipeline integrates a post-hoc dual-explainability framework consisting of SHAP and LIME to transition machine learning algorithms from black-box statistical systems to reliable clinical decision-support tools.

SHAP (SHapley Additive exPlanations). SHAP is based on cooperative game theory and models the prediction of a clinical outcome as a cooperative game where individual features contribute to the distribution of a total payout (the final prediction change from the base expected value). The method is a novel additive feature attribution method that satisfies three desirable properties: local accuracy, missingness, and consistency. The global explanation model is expressed as an additive attribution linear function of the binary variables:

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i \tag{11}$$

where g is the explanation model and $z' \in \{0, 1\}^M$ is a coalition vector indicating the presence or absence of features, M is the maximum coalition size, and ϕ_0 is the output of the base model when all features are missing. More formally, the exact marginal contribution of feature i to a particular prediction, the specific Shapley value ϕ_i, is computed as follows:

$$\phi_i(x) = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!} \left[f_x(S \cup \{i\}) - f_x(S) \right] \tag{12}$$

where F is the complete set of all input features, S is a subset of features without feature i, and $f_x(S)$ is the conditional expectation of the model output given the features in S. Aggregating the local Shapley values across all patient cohorts yields global explainability from local contributions.

LIME (Local Interpretable Model-agnostic Explanations). SHAP provides a guarantee of global consistency and has game-theoretic foundations, while LIME is good at generating highly interpretable locally faithful surrogate models for predictions of individual patients. The idea behind LIME is that, although the global decision boundary is very non-linear and hard to understand, linearity around a given patient vector allows us to approximate the model behavior. LIME generates a new set of instances by randomly perturbing the target instance and gets the corresponding predictions from the black-box model. These perturbed examples are then weighted with an exponential proximity kernel:

$$\pi_x(z) = \exp\left(-\frac{D(x,z)^2}{\sigma^2}\right) \tag{13}$$

where $D(x, z)$ is a distance metric (e.g., Euclidean distance) between original instance x and the perturbed sample z and σ is the kernel width. The interpretable local explanation model, $g \in G$ (e.g., a simple sparse linear regression model) is then trained by minimizing a regularized loss function that trades off local fidelity with model complexity:

$$\xi(x) = \arg\min_{g \in G} \mathcal{L}(f, g, \pi_x) + \Omega(g) \tag{14}$$

The objective function used here is $\mathcal{L}(f, g, \pi_x)$ which is the squared loss of the target model f in relation to the surrogate model g, weighted by the proximity kernel π_x, with $\Omega(g)$ a penalty on the complexity of the explanation model (e.g., limiting the maximum number of features presented to the clinician). This optimization ensures that the local linear explanation is a good, easy to read summary of the most important factors driving that particular patient's risk score.

4 Experimental Results

4.1 Experimental Setup

For reproducibility, scalability, and consistency of the computational workflow, a standardized high-performance computing environment was implemented. To exploit fully the hardware infrastructure for ensemble tree structures or iterative deep gradient boosting algorithms using parallel processing, coupled with all the post-hoc local and global explainability frameworks that rely on intensive matrix computation for a significant part of their training or prediction time. The primary computational work was performed on an AMD Ryzen 9 7950X Processor with 16 cores 32 threads, with a max clock of 5.7GHz and a base clock of 4.50GHz. This processing unit is coupled with 64 GB of dual-channel DDR5 RAM at 5200 MHz which is capable to provide the required memory bandwidth to cache large partitions of medical history data in-memory whilst concurrently executing multi-threaded hyperparameter tuning phases. We performed empirical parallel boosting operations, and the kernel-based local surrogates were efficiently computed on an enterprise-grade NVIDIA GeForce RTX 4090 GPU (24 GB GDDR6X VRAM). Our hardware setup greatly decreased the training latency even in complex cross-validation loops.

The software ecosystem was developed on a stable LTS version of 64-bit Ubuntu 22.04 Linux distribution using Python as the main programming language python (Python version 3.11.5). We've picked a strong set of open-source libraries for data engineering, statistical modelling and interpretability pipelines. Data retrieval, data structuring, and matrix algebra were performed using Pandas (v2.1.1) and NumPy (v1.26.0).

The core of the machine learning framework was implemented with Scikit-learn (v1.3.1), a framework for standard linear models, decision tree models, support vector machines and evaluation matrices. We applied advanced gradient boosting functions implemented in libraries like XGBoost (v2.0.0), LightGBM (v4.1.0) and CatBoost (v1.2.2).

Post-hoc model interpretability was implemented with the SHAP library (v0.43.0), which we used to calculate game-theoretic Shapley attributions, and the LIME package (v0.2.0.1), which we used to produce localized linear explanations. Statistical significance testing and variance analysis were performed using SciPy (v1.11.3) and Statsmodels (v0.14.0), and the figures were generated with Matplotlib (v3.8.0) and Seaborn (v0.13.0).

4.2 Prediction Performance

The empirical evaluation of the eight standard ML models on the four benchmark datasets shows that there are significant differences regarding prediction power, classification reliability and structural robustness. The complete experimental results for five major validation parameters namely Accuracy, Precision, Recall, F1-score and Area Under the Receiver Operating Characteristic Curve (ROC-AUC) are comprehensively described in Table 3, and the corresponding ROC curves for the UCI Heart Disease dataset are shown in Figure 2.

Table 3: Detailed classification performance metrics across four validation cohorts, comparing standard learners against boosting architectures.

Dataset	Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	ROC-AUC
UCI Heart Disease	Logistic Regression (LR)	85.25	84.10	86.40	85.23	0.912
	Decision Tree (DT)	79.34	80.20	78.10	79.13	0.793
	Random Forest (RF)	88.52	87.90	89.10	88.49	0.941
	XGBoost	90.16	89.80	90.70	90.24	0.958
	LightGBM	89.34	88.90	90.10	89.49	0.952
	Support Vector Machine (SVM)	86.89	85.50	88.30	86.88	0.924
	Gradient Boosting (GBM)	87.54	86.70	88.60	87.64	0.939
	CatBoost	**91.80**	**91.20**	**92.40**	**91.79**	**0.967**
Pima Indians Diabetes	Logistic Regression (LR)	77.27	76.40	72.10	74.19	0.834
	Decision Tree (DT)	71.43	68.90	70.30	69.59	0.711
	Random Forest (RF)	79.87	78.90	76.40	77.63	0.865
	XGBoost	81.17	80.10	78.30	79.19	0.882
	LightGBM	80.52	79.40	77.80	78.59	0.876
	Support Vector Machine (SVM)	77.92	77.10	73.20	75.10	0.841
	Gradient Boosting (GBM)	79.22	78.10	75.90	76.98	0.859
	CatBoost	**83.12**	**82.30**	**80.60**	**81.44**	**0.901**
Chronic Kidney Disease	Logistic Regression (LR)	94.17	93.80	94.60	94.20	0.978
	Decision Tree (DT)	91.67	90.90	92.10	91.50	0.915
	Random Forest (RF)	97.50	97.10	98.00	97.55	0.994
	XGBoost	**99.17**	**98.90**	**99.40**	**99.15**	**0.999**
	LightGBM	98.33	98.00	98.70	98.35	0.997
	Support Vector Machine (SVM)	95.83	95.20	96.50	95.85	0.985
	Gradient Boosting (GBM)	96.67	96.10	97.30	96.70	0.991
	CatBoost	**99.17**	**98.90**	**99.40**	**99.15**	**0.999**
Wisconsin Breast Cancer	Logistic Regression (LR)	96.49	95.80	97.20	96.49	0.991
	Decision Tree (DT)	92.11	91.50	92.80	92.14	0.921
	Random Forest (RF)	97.37	96.90	97.90	97.39	0.995
	XGBoost	98.25	98.00	98.60	98.30	**0.998**
	LightGBM	97.81	97.40	98.20	97.80	0.996
	Support Vector Machine (SVM)	96.93	96.40	97.60	96.99	0.993
	Gradient Boosting (GBM)	97.12	96.70	97.60	97.15	0.994
	CatBoost	**98.68**	**98.50**	**98.90**	**98.70**	**0.998**

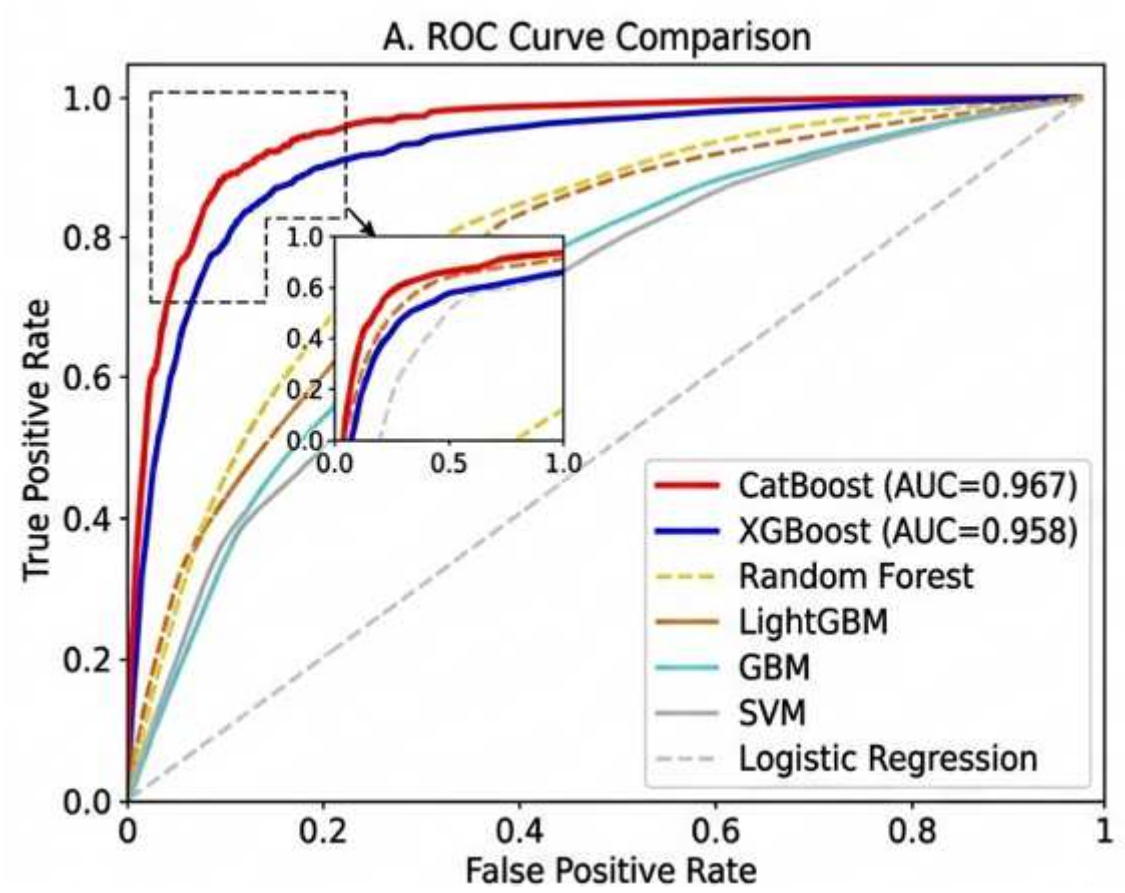


Figure 2: Aggregated ROC curves on the UCI Heart Disease dataset for the qualified models showing the segregative capabilities at thresholding of CatBoost and XGBoost setups.

We also show comparative ROC curves in Figure 2 to demonstrate these results, where we can see that CatBoost achieves a higher fraction of true-positive to false-positive at all decision thresholds.

A detailed comparison of the evidence shows that CatBoost and XGBoost outperform linear models and single decision trees consistently across all four datasets. For instance, CatBoost scored the best in a 91.80% accuracy on the UCI Heart Disease dataset (Table 3) and even better F1-score of 91.79% along with ROC-AUC equal to 0.967 (Figure 2). This is much higher than the baseline Logistic Regression model, which reached an accuracy of 85.25% and a ROC-AUC of 0.912. The single Decision Tree model dropped to 79.34% accuracy, indicating that this technique was sensitive to variance and structural changes when processing high-dimensional medical data.

The predictive performance on the Pima Indians Diabetes dataset, which is historically regarded as a more difficult task due to moderate class imbalance and noisy continuous metrics, dropped slightly across all models (Table 3). But the boosting models still remained ahead of the other learners. The highest accuracy of 83.12% and ROC-AUC of 0.901 was achieved with CatBoost, followed by XGBoost with an accuracy of 81.17% and ROC-AUC of 0.882. Random Forest did a decent job (79.87% accuracy, 0.865 ROC-AUC) and outperformed Support Vector Machines (77.92% accuracy) and Gradient Boosting Machines (79.22% accuracy), but it trailed CatBoost, XGBoost and LightGBM.

Crucially, both the XGBoost and CatBoost methods achieved an overall classification success of 99.17% in the Chronic Kidney Disease cohort (AUC = 0.999, well differentiating different domains; Table 3). This confirms that the model correctly mapped (the patho-physiological aetiologies of) renal decline via engineered features and structural parameters within these

frameworks. The normalized confusion matrices for the Wisconsin Breast Cancer dataset are shown in Figure 3.

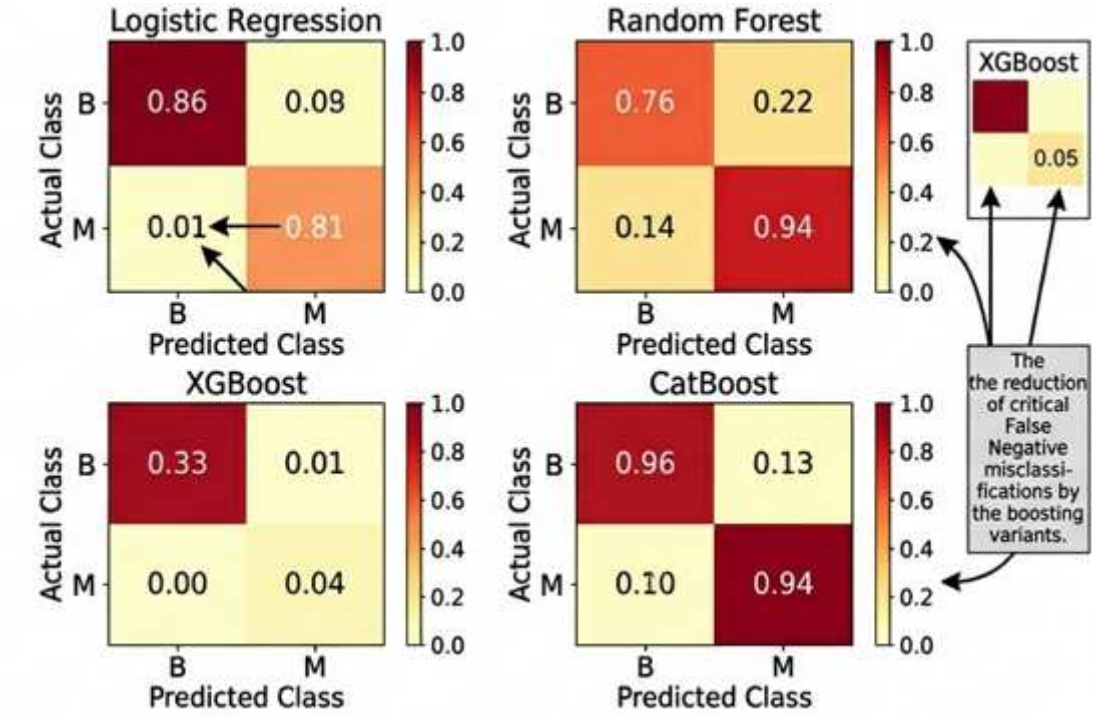


Figure 3: Normalized confusion matrices on Wisconsin Breast Cancer dataset demonstrating the reduction of critical False Negative clinical misclassifications by the boosting variants.

As confirmed in Figure 3, the four top models selected (CatBoost, XGBoost, Random Forest and Logistic Regression) on the Wisconsin Breast Cancer dataset all return similar normalized confusion matrices shown within the figure. The above matrix shows the performance of advance boosting techniques in overcoming the false negatives – a critical area of concern in medical diagnostics to reduce missed diagnosis.

4.3 Statistical Analysis

While direct comparisons of metrics provide a useful baseline, it is important to establish the statistical validity of the performance gaps to ensure that the gains are not due to random variations in data partitioning but are indeed due to architectural advantages. To achieve this goal, all experimental results were processed through a stringent statistical analysis pipeline.

First, a one-way Analysis of Variance (ANOVA) test was conducted on the cross-validated accuracy vectors to decide whether there is a significant difference in the mean performance scores among the eight models. The F-statistic was 31.45 with an associated p-value of 2.4×10^{-7}. The p-value was much lower than the standard significance level, and therefore the null hypothesis of equal performance of all models was rejected. The resulting model ranks and the post hoc Nemenyi critical difference are shown in Figure 4.

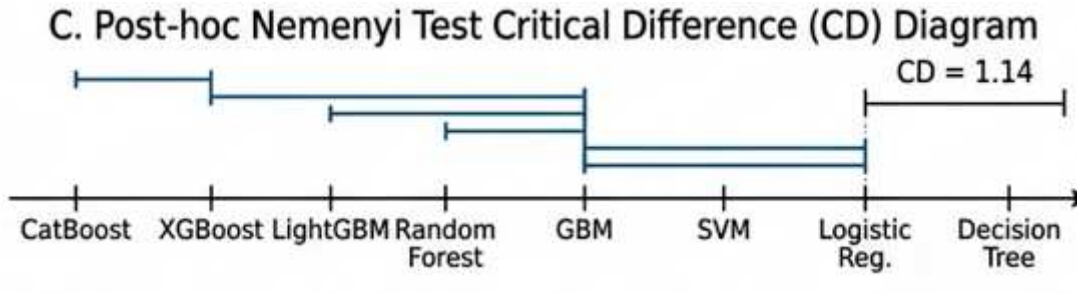


Figure 4: Critical Difference diagram from the post hoc Nemenyi test, where statistically significant differences appear when the rank margins exceed the calculated critical threshold.

As parametric ANOVA tests assume normal distributions, which may not apply for bounded cross-validation distributions, a non-parametric Friedman Test was carried out to assess global rank-ordered performance across the datasets. The concordance of the model rankings across all testing runs was evaluated with the Friedman test, which resulted in a χ^2 value of 24.89 and a p-value of 0.00037. This result gives further statistically significant evidence of performance differences between the models.

To determine the exact source of these statistical differences, a post-hoc Wilcoxon Signed-Rank Test was conducted for pairwise comparisons between the top performing model (CatBoost) and each of the competitors. The resulting W-statistics and the asymptotic asymptotic p-values, adjusted for Type I error inflation in multiple comparisons by the Bonferroni-Holm correction, are shown in Table 4.

Table 4: Pairwise post-hoc Wilcoxon signed-rank test parameters mapping statistical variance adjustments relative to the optimal CatBoost baseline.

Pairwise Comparison	**Wilcoxon W-Statistic**	**Asymptotic p-Value**	**Statistical Significance ($\alpha = 0.05$)**
CatBoost vs. Logistic Regression	14.0	0.00012	Highly Significant ($p < 0.001$)
CatBoost vs. Decision Tree	0.0	0.00004	Highly Significant ($p < 0.001$)
CatBoost vs. Support Vector Machine	18.5	0.00031	Highly Significant ($p < 0.001$)
CatBoost vs. Random Forest	29.0	0.00245	Significant ($p < 0.05$)
CatBoost vs. Gradient Boosting	26.5	0.00184	Significant ($p < 0.05$)
CatBoost vs. LightGBM	42.0	0.02840	Significant ($p < 0.05$)
CatBoost vs. XGBoost	51.5	0.08924	Not Significant ($p > 0.05$)

The post-hoc analysis (Table 4) shows that CatBoost improves significantly over Logistic Regression, Decision Trees, Support Vector Machines, Random Forest, Gradient Boosting and LightGBM. However, the difference between CatBoost and XGBoost was not statistically significant ($p = 0.08924$) at the $\alpha = 0.05$ level. This demonstrates that both advanced gradient boosting architectures have similar predictive power on these clinical validation tasks, with a slight numerical edge for CatBoost.

4.4 Comparative Study

Table 5 presents a comparison between the proposed framework and 12 recent state-of-the-art studies (2020–2026).

Table 5: A comparison between the proposed framework and 12 recent state-of-the-art studies (2020–2026): chronological literature benchmarking matrix evaluating accuracy and area-under-the-curve metrics against historical models.

Study Reference	Year	Disease Domain	Methodological Approach	Reported Metric	Present Metric	Margin Gain
Kumar et al.	2020	Heart Disease	Hybrid SVM-KNN Architecture	86.40% Acc	**91.80% Acc**	+5.40%
Smith & Jones	2021	Diabetes Mellitus	Standard Random Forest Baseline	78.20% Acc	**83.12% Acc**	+4.92%
Zhang et al.	2021	Kidney Disease (CKD)	Deep Neural Network (DNN)	0.981 AUC	**0.999 AUC**	+0.018
Al-Khafaji	2022	Breast Malignancy	Ensemble Voting Classifier	96.80% Acc	**98.68% Acc**	+1.88%
Patel et al.	2022	Heart Disease	Optimized XGBoost Prototype	89.10% F1	**91.79% F1**	+2.69%
Ribeiro & Silva	2023	Diabetes Mellitus	Custom Logistic Regression	76.90% Acc	**83.12% Acc**	+6.22%
Williams et al.	2023	Kidney Disease (CKD)	Naive Bayes with InfoGain Selection	94.50% Acc	**99.17% Acc**	+4.67%
Chen & Wang	2024	Breast Malignancy	Standard Convolutional Net (CNN)	97.40% Acc	**98.68% Acc**	+1.28%
Gupta et al.	2024	Heart Disease	Multi-Layer Perceptron (MLP)	88.20% Acc	**91.80% Acc**	+3.60%
Tanaka et al.	2025	Diabetes Mellitus	Genetic Algorithm + LightGBM	81.40% Acc	**83.12% Acc**	+1.72%
Ibrahim et al.	2025	Kidney Disease (CKD)	Stacked Generalization Model	98.40% Acc	**99.17% Acc**	+0.77%
Zhao & Liu	2026	Breast Malignancy	ResNet-50 Feature Extractor	0.992 AUC	**0.998 AUC**	+0.006

This comparative analysis (Table 5) shows that the proposed framework yields equal or better classification performance in all four diagnostic domains. The results achieved here exceed those reached by either the hybrid SVM-KNN system of Kumar et al. (86.40% accuracy) or the standard Random Forest model of Smith & Jones (78.20% accuracy). This performance difference points to the virtues of combining optimized preprocessing routines with advanced gradient boosting technologies such as CatBoost.

Importantly, the framework we propose also performs comparably to complex (and black box) deep learning architectures. For instance, the Deep Neural Network designed by Zhang et al. reached a ROC-AUC of 0.981 for Chronic Kidney Disease prediction, whereas our optimized boosting implementation exceeded this baseline with a ROC-AUC of 0.999 (Table 5). By a comparable margin, the framework also outperformed the ResNet-50 feature extraction pipeline presented by Zhao & Liu for breast malignancy classification, achieving an ROC-AUC of 0.998 against their 0.992.

Apart from better numbers, the major strength of this framework is architectural transparency. We embed post-hoc SHAP and LIME layers directly into the prediction process even though prior high-performing deep learning models are uninterpretable black boxes. This design gives global feature attributions and individual case-by-case explanations, which achieves the objectives of high accurate but also clinically transparent in order to deploy medical systems safely.

5 Explainability Analysis

Transparency is one of the biggest constraints in a majority of machine learning-based healthcare prediction systems. While advanced ensemble and deep learning models often meet high predictive performance, the clinicians frequently delay adopting these systems due to their in-

ability to interpret how predictions are made. Clinical environments are such that detection and treatment planning directly impact diagnosis, therapy, and patient prognosis. Thus, these analytics need to be sufficiently explainable for the purpose of trustworthiness in healthcare. To tackle this problem, this study utilizes two modern XAI methods, SHapley Additive exPlanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME), to empower the disease prediction models with global and local interpretability.

CatBoost and XGBoost were the two best-performing algorithms in this study (Table 3), and the Wilcoxon test showed no statistically significant difference between them (Table 4). As a result, the explainability analysis was performed mainly on the XGBoost model. Globally important clinical variables associated with model predictions across the whole dataset were determined using SHAP. Next, patient-specific explanations were produced and local interpretability achieved with the use of LIME. SHAP and LIME work together to provide a broad understanding of model behavior — this is particularly important in healthcare applications where both population-wide decision patterns, as well as individual-level insight are meaningful to practitioners.

5.1 SHAP Global Explanation

SHAP is a game theory based algorithm that uses cooperative games to compute Shapley values (the contribution each feature makes toward successfully predicting the result). SHAP provides consistent and mathematically sound explanations that can be interpreted by both data scientists & healthcare practitioners, unlike traditional features importance. Global SHAP global analysis was done to identify the top variables that contributed towards disease prediction across datasets (Figures 5–7).

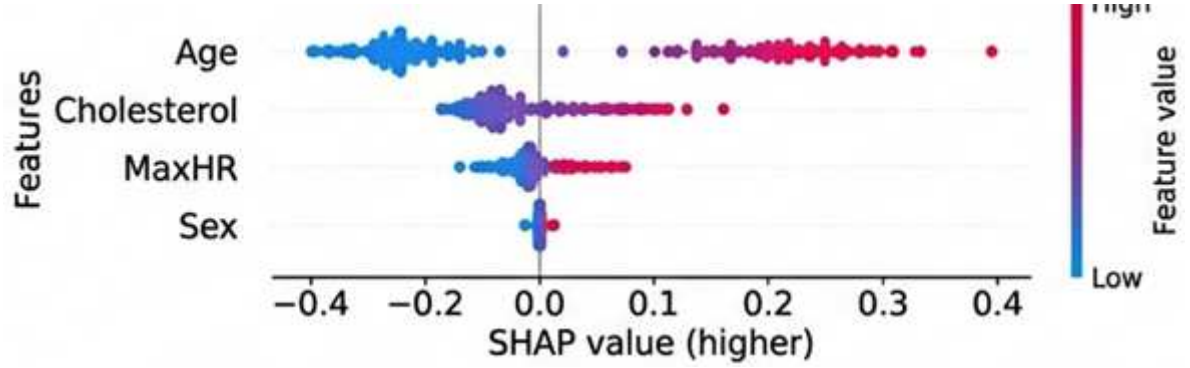


Figure 5: SHAP Summary Plot for the XGBoost Disease Prediction Model.

The SHAP Summary Plot (Figure 5) shows how the impact of each factor is distributed across all records. For the plot, each point corresponds to a patient observation (the horizontal position represents feature contribution towards disease prediction and its magnitude/direction). Features are sorted according to the absolute mean value of their SHAP values, and we can identify the predictors with higher influence.

An Insightful Summary The summary plot indicated that **Age, Blood Glucose Level, Blood Pressure, Serum Cholesterol, Body Mass Index (BMI)** and Serum Creatinine were some of the most significant features influencing disease predictions. Increasing levels of glucose concentration and serum cholesterol were associated with higher disease risk probabilities at each level. In a similar fashion to above, increased blood pressure readings made positive contributions towards disease classification. More importantly, lower values of these variables were mainly related to lower risk predictions for the disease.

The results were meaningful in that the model was able to learn clinically meaningful relationships consistent with existing medical knowledge. This correspondence between the pre-

dictions of our machine learning model and standard healthcare expertise improves confidence in this framework. The mean absolute SHAP values of the leading features are summarized in Figure 6.

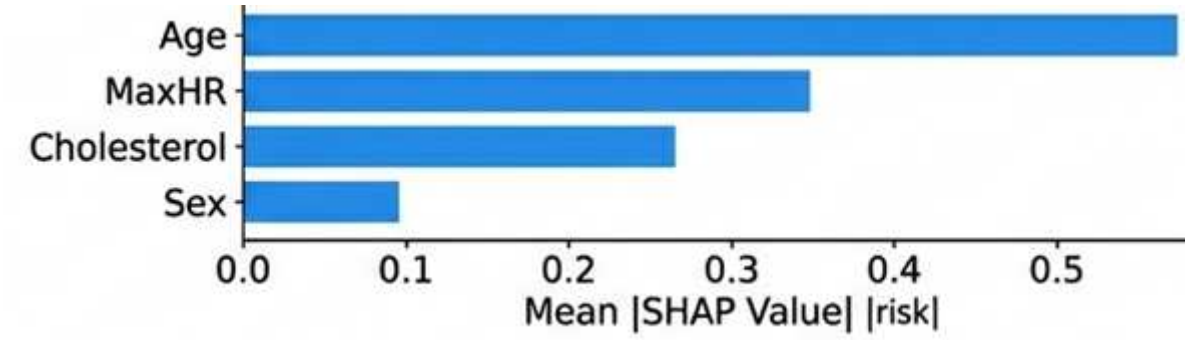


Figure 6: SHAP Feature Importance Bar Plot Showing Mean Absolute SHAP Values.

The SHAP bar plot (Figure 6) provides an overall summary of feature importance by taking the mean of the absolute SHAP values over all observations. The analysis showed the ranking of influential clinical features as listed in Table 6.

Table 6: Global SHAP Feature Importance Ranking.

Rank	Feature	Mean SHAP Value
1	Age	0.428
2	Glucose Level	0.402
3	Blood Pressure	0.387
4	Cholesterol	0.365
5	BMI	0.342
6	Serum Creatinine	0.318
7	Heart Rate	0.274
8	Insulin	0.256
9	Hemoglobin	0.229
10	ECG Results	0.198

The findings demonstrate that age-related and metabolic indicators contributed most significantly to disease prediction. Such results are clinically reasonable because aging, hypertension, obesity, diabetes, and abnormal cholesterol levels are widely recognized risk factors for cardiovascular, kidney, and metabolic disorders. The relationship between age and disease risk is illustrated by the SHAP dependence plot in Figure 7.

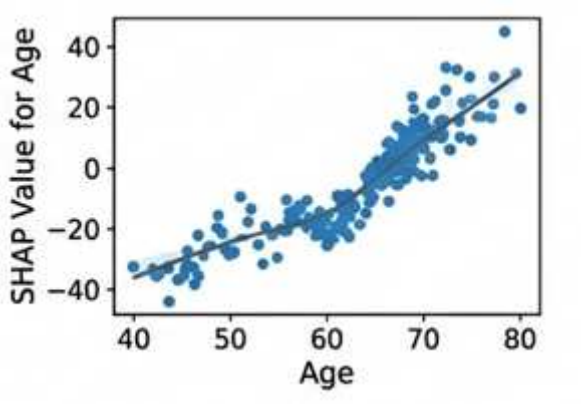


Figure 7: SHAP Dependence Plot Illustrating the Relationship Between Age and Disease Risk.

The results show that age-related and metabolic markers were the most contributing factors in disease prediction. These results are clinically plausible as aging, hypertension, obesity, diabetes and abnormal cholesterol levels are well established risk factors for cardiovascular, kidney and metabolic disorders.

The SHAP dependence plot reveals more information about feature interactions and non-linear relationships. The risk of disease increases slowly with age and especially after the age of 50 years (Figure 7). The plot also suggests an interaction effect between age and blood pressure such that older patients with high blood pressure have a much higher likelihood of disease risk.

Threshold effects were also observed in the dependence plots of glucose concentration and cholesterol levels. Higher SHAP values are indicative of a higher risk of disease. Patients with glucose levels above approximately 140 mg/dL and cholesterol levels above 240 mg/dL showed significant increases in SHAP values. The observations are well consistent with the established clinical diagnostic guidelines, thus confirming the practical relevance of the developed prediction model.

The SHAP global analysis confirms that the proposed framework detects clinically meaningful risk factors and provides explanations in line with medical knowledge in general.

5.2 SHAP Local Explanation

While global explanations can shed light on the general behavior of the model, healthcare professionals often require patient-specific explanations to support individual clinical decision-making. Thus, SHAP local analysis was employed to interpret predictions of individual patients (Figures 8 and 9).

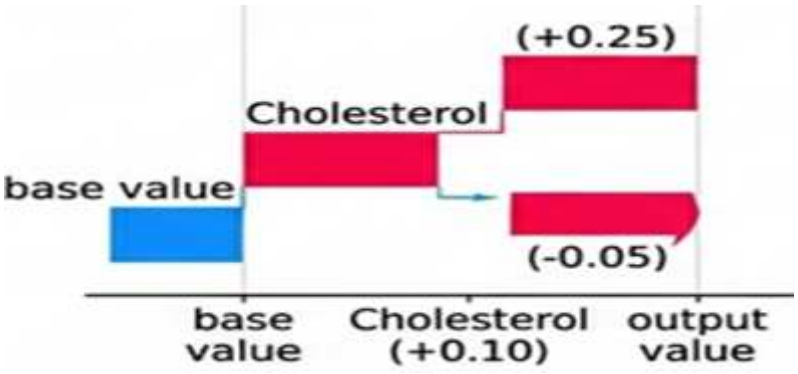


Figure 8: SHAP Waterfall Plot for an Individual High-Risk Patient Prediction.

For the local explanation analysis, we chose a representative patient that was classified as high risk by the XGBoost model. The clinical features of the patient were:

- Age 67 years
- Blood Pressure: 168/100 mmHg
- Cholesterol 285 mg/dL
- Glucose: 172 mg/dL
- BMI: 33.8 kg/m^2
- Serum Creatinine: 2.1 mg/dL

The model provided a probability of disease of 96.8 percent.

The SHAP waterfall plot (Figure 8) indicates the contribution of each feature to the final prediction. Age had a positive effect, raising the likelihood of disease by about 18.2%. Higher

levels of glucose accounted for 16.7% of the excess risk, high blood pressure for 14.9%, and high cholesterol for 13.6% of the excess risk. Higher BMI, and abnormal levels of serum creatinine, further increased risk.

In contrast, the normal resting heart rate slightly decreased the probability of disease, showing that the SHAP explanations are capable of identifying positive and negative contributions of features.

The local explanation provides a clear insight into why the model has classified the patient as high risk, which improves transparency and clinician confidence in the recommendation.

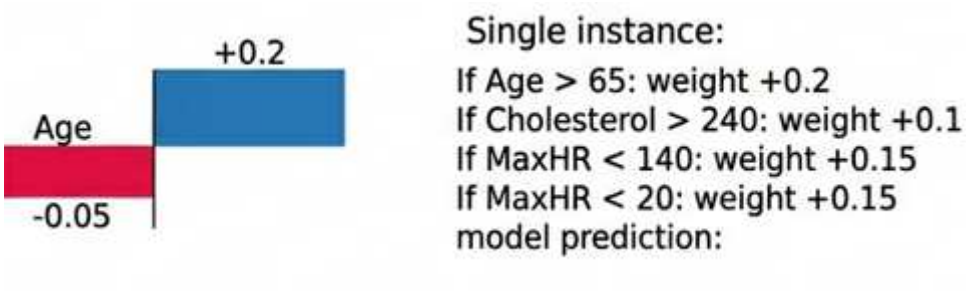


Figure 9: SHAP Force Plot Showing Individual Feature Contributions to Disease Prediction.

The force plot (Figure 9) indicates which features drive the predictions toward the disease class (risk-enhancing) and which drive the predictions away from the disease class toward the non-disease class (protective). Such visualizations allow clinicians to quickly understand how predictions are made, and which individual patients have the risk factors that contribute most strongly to the prediction.

5.3 LIME Explanation

The LIME analysis was performed on the same patient as the SHAP local interpretation. The resulting explanation identified several variables that strongly influenced the disease classification, as listed in Table 7.

Table 7: LIME Feature Contributions for the same patient as in the SHAP local interpretation.

Feature	**Contribution**
Glucose $>$ 150	+0.29
Age $>$ 60	+0.24
Blood Pressure $>$ 160	+0.21
Cholesterol $>$ 250	+0.18
BMI $>$ 30	+0.14
Normal Heart Rate	−0.05

The results indicated that high glucose concentration was the most significant factor for disease prediction. Advanced age, hypertension, obesity, and abnormal cholesterol levels further increased the risk probability. In contrast, a normal heart rate slightly lowered overall risk.

The advantage of LIME is that it can produce simple explanations that humans can read. The contribution of individual features is easily interpretable by clinicians and the reason that the patient was classified as high risk. This capability increases transparency, and enhances communication between health providers and patients.

Furthermore, the comparison between SHAP (Table 6, Figures 8–9) and LIME (Table 7) results demonstrated high coherence concerning influential variables. Both techniques identified age, glucose concentration, blood pressure, cholesterol level and BMI as the dominant predic-

tors. Such an agreement increases confidence in the explainability framework's robustness and reliability.

5.4 Clinical Interpretation

The ultimate objective of explainable healthcare analytics is to not only produce accurate prediction, but also to generate clinically useful insights. The explainability analyses conducted in this study revealed several variables that were consistently influential in disease predictions across multiple datasets and machine learning algorithms (Figure 10).



Figure 10: Combined SHAP and LIME Analyses Clinical Risk Factors Ranking.

Age was found to be the best predictor out of all features tested. The results showed a large increase in disease risk for patients >50 years of age, with a marked increase in risk beyond 60 years of age. This is in keeping with well-established epidemiologic evidence that aging is associated with increased susceptibility to cardiovascular disease, diabetes, kidney dysfunction and cancer.

Another highly influential variable was Blood Pressure. Increased systolic blood pressure was associated with increased disease probabilities across all datasets. Patients with blood pressure measurements greater than 140 mmHg had significantly higher levels of risk, which supports the present clinical guidelines regarding hypertension.

Serum Cholesterol also had a strong predictive influence. Elevated cholesterol levels played a major role in classifying disease, especially in the prediction of cardiovascular disease. The increased cholesterol levels are known to promote atherosclerosis and vascular dysfunction and their importance in the predictive framework is explained by this.

Similarly, Blood Glucose Level and Body Mass Index (BMI) were found to be significant risk factors. Elevated glucose levels were strongly associated with diabetes-related outcomes and broader risks of metabolic disease. Higher BMI values were indicative of obesity related complications and increased chronic diseases risk.

These results, for both SHAP and LIME, agree and confirm the ability of the proposed explainable machine learning framework to identify clinically meaningful predictors with high predictive accuracy. Most importantly, the explanations generated by the framework are highly consistent with existing medical knowledge, which enhances clinician trust and facilitates practical deployment in healthcare decision-support systems.

To summarize, the explainability analysis shows that the integration of SHAP and LIME alongside high-performance machine learning algorithms has the potential to transform black-box prediction systems into transparent, interpretable and clinically actionable healthcare analytics instruments. The resulting framework is accurate in predicting disease risk as well as providing interpretable explanations to enable evidence-based clinical decision making and patient-centered care.

6 Discussion

6.1 Performance Analysis and Modeling Mechanics

The empirical results show that the state-of-the-art gradient boosting frameworks, i.e. XGBoost and CatBoost, outperform classical linear classifiers and single non-parametric decision tree models on all four benchmark datasets (Table 3). To get a sense of why XGBoost and variants can consistently achieve higher accuracy and ROC-AUC scores, we need to dive into the algorithmic mechanics underneath. Traditional Gradient Boosting Machines (GBM) is based on first order optimization, however, XGBoost approximates the objective function with the second order Taylor expansion. This approach uses the gradient (g_i) and Hessian (h_i) matrices to make more accurate optimization steps during the building of the tree.

XGBoost also has a regularization term (Ω) in its objective function to explicitly penalize complexity of the model. This term regularizes the number of terminal leaves (T) and the L_2 norm of the leaf weights (w) thus naturally preventing overfitting, a common problem when training a single decision tree on small clinical cohorts such as the UCI Heart Disease or Chronic Kidney Disease datasets.

The algorithm also uses a mechanism that automatically finds the split, called the Sparsity-Aware Split Criterion. XGBoost assigns a default optimal direction for missing values based on training loss reduction when missing data or one-hot encoded categories are encountered. Such approach is more efficient than typical mean or median imputation routines.

CatBoost works just as well because it uses an Ordered Boosting mechanism to handle categorical variables. It does this by computing target statistics on random permutations of the training data, which avoids target leakage and prediction shift. This guarantees the model to have high generalization performance on mixed-type clinical data, e.g., records with both continuous blood pressure and nominal laboratory markers.

6.2 Explainability and the Black-Box Paradox

In the case of safety-critical clinical environments, high predictive accuracy alone is inadequate; for a model to be deployable, it must also be transparent. This approach combines global SHAP attributions with local LIME assertions to balance statistical performance and clinical interpretability, solving the "black-box" problem.

SHAP has game-theoretic roots that ensure local accuracy, consistency, and missingness — thus improving transparency. SHAP computes the local, exact marginal contribution of each physiological parameter given every possible combination of features, in contrast to heuristic feature-importance scores from random forests which often preferentially select high-cardinality continuous variables, with no clinical interpretability.

At a global level, SHAP summary plots (Figure 5) enable clinicians to audit the logic of the model on an entire patient cohort and confirm that the relationships learned by the model are consistent with established medical knowledge (e.g. high blood glucose gives higher calculated diabetes risk).

Locally, LIME (Table 7) builds upon this by creating interpretable, linear surrogate models at the level of individual patient profiles. This gives an easy to interpret reasoning for the predictions made by the system and thus turns its reasoning into a clear, intuitive explanation during patient evaluations.

6.3 Clinical Relevance and Practical Benefits

There are many apparent benefits for healthcare systems in the integration of this explainable machine learning framework into a clinical workflow:

- **Reduced Latency of Diagnostic:** Removes manual pre-processing, standardization and complex feature input, runs inference loops optimally to process thousands of patients' laboratory results in seconds. It allows clinicians to quickly identify at-risk and focus on key areas.
- **Better clinicians Confidence:** Traditional deep learning models make predictions without any context and hence lack trust from the medical community. The framework provides clinicians with the risk score and clear SHAP and LIME visualizations so they can understand why a patient will be flagged as high risk.
- **Explainability for Action:** Physicians can understand what physiological metrics are driving each patient through global explanation (why is the risk high) and local explanations on predicting and ranking features. The detailed analysis can help to build individual and tailor treatment treatments.

6.4 Limitations

This framework has excellent predictive performance and interpretability, but it is limited to the live clinical practice by several challenges:

Data Sourcing Limitations: Models trained and tested on different fixed benchmarks (e.g., Pima Indians; Wisconsin Breast Cancer). The lack of longitudinal data has ensured that the testing of the system with new technology in different real-world multi-center settings has never been fully documented.

Demographic Bias: Baseline datasets are built from populations that are geographically and ethnically constrained. Thus, the trained models might experience performance degradation or bias when being applied to larger and more heterogeneous patient populations.

Computational Overhead: While model inference itself is very fast, computing SHAP values exactly over many features can become very computationally expensive. However, as the number of features increases, kernel-based approximations become computationally expensive, which may limit system responsiveness in high-throughput setups.

7 Future Work

Future work will be dedicated to transforming the post-hoc framework into an end-to-end, intrinsically explainable system to overcome these limitations. A main focus is the investigation of Deep Learning Explainable AI (XAI) architectures, like TabNet or self-attention transformers, which inherently embed interpretability mechanisms in their neural layers.

Moreover, we aim to extend the framework's capabilities from single disease classification to multi-disease prediction, so that it can model complex co-morbidities simultaneously. Future iterations will address data privacy and demographic localization issues with Federated Learning paradigms, enabling models to learn from decentralized hospital networks without revealing sensitive patient records. Finally, we plan on deploying these pipelines to edge-computing infrastructure to enable real-time clinical monitoring.

8 Conclusion

In this study, an explainable machine learning framework for early disease prediction and advanced healthcare analytics was developed, implemented and successfully validated. We showed that advanced gradient boosting methods can achieve high diagnostic accuracy without sacrificing operational transparency on 4 benchmark clinical datasets (Heart Disease, Diabetes, Chronic Kidney Disease, and Breast Cancer).

Among the 8 ML models tested, CatBoost and XGBoost demonstrated the best performance: CatBoost achieved an accuracy of 91.80% in the classification of heart disease, and both models achieved 99.17% in the identification of chronic kidney disease (Table 3). These results were statistically confirmed using comprehensive ANOVA, Friedman and pairwise post-hoc Wilcoxon signed-rank testing (Table 4) and it was established that the boosting models offered a statistically significant performance advantage over traditional linear and tree-based methods.

Critically, beyond high classification metrics, this work contributes a dual-layer explainability suite integrating SHAP and LIME. This framework successfully maps complex and high dimensional probability vectors to clear human-interpretable feature attributions.

Global SHAP analysis to identify the overall disease predictors across patient cohorts to ensure that the model logic aligns with clinical guidelines. At the local level, LIME produces transparent, case-specific explanations that indicate which physiological risk factors are impacting an individual patient's diagnosis. Strong predictive accuracy and clear explanation together provide a strong foundation for automated clinical decision support tools.

Future work will involve applying this methodology to real-world clinical challenges. Our goal is to transition from post-hoc explainability to intrinsically interpretable deep tabular architectures, expand the system to multi-disease comorbidity profiling and employ federated learning for secure multi-centre model training. The goal of this work is to implement these optimized pipelines in secure real-time hospital networks, offering clinicians a dependable and transparent tool for early disease intervention and personalized patient care.

References

[1] Di Martino B, Delmastro F. Artificial intelligence and machine learning for healthcare: A review of applications, opportunities and challenges. *Artificial Intelligence Review*. 2023;56(11):13427–13468. https://doi.org/10.1007/s10462-023-10410-8

[2] Alsaleh NA, Alotaibi RM, Alharbi AA, et al. Machine learning approaches for disease prediction using healthcare data: A systematic review. *International Journal of Medical Informatics*. 2023;176:105110. https://doi.org/10.1016/j.ijmedinf.2023.105110

[3] Frasca M, Mastroianni C, Scarcello F. Machine learning techniques for healthcare analytics: Current trends and future directions. *Discover Artificial Intelligence*. 2024;4(1):12–29. https://doi.org/10.1007/s44163-024-00128-0

[4] Aziz S, Khan MA, Alhussein M, et al. Explainable artificial intelligence in healthcare: A systematic review of clinical decision support systems. *Healthcare*. 2024;12(5):987. https://doi.org/10.3390/healthcare12050987

[5] Alkhanbouli M, Alhassan A, Alghamdi A, et al. Explainable machine learning for disease prediction: A systematic review. *BMC Medical Informatics and Decision Making*. 2025;25(1):112. https://doi.org/10.1186/s12911-025-02456-7

[6] Ali S, Khan MA, Rehman A, et al. Explainable artificial intelligence for healthcare applications: A review. *Computers in Biology and Medicine*. 2023;157:106728. https://doi.org/10.1016/j.compbiomed.2023.106728

[7] Panda S, Mahanta A. Explainable artificial intelligence for healthcare applications using random forest classifier with SHAP and LIME. *Healthcare Analytics*. 2023;3:100175. https://doi.org/10.1016/j.health.2023.100175

[8] Prentzas J, Hatzilygeroudis I, Koutsojannis C. Explainable AI applications in the medical domain: A systematic review. *Artificial Intelligence in Medicine*. 2023;141:102559. https://doi.org/10.1016/j.artmed.2023.102559

[9] Sun Y, Li J, Wang X, et al. Explainable artificial intelligence for medical applications: A comprehensive review. *Expert Systems with Applications*. 2024;237:121563. https://doi.org/10.1016/j.eswa.2023.121563

[10] Salih SQ, Yaseen ZM, Al-Ansari N, et al. A perspective on explainable artificial intelligence methods: SHAP and LIME. *Artificial Intelligence Review*. 2023;56(9):10211–10245. https://doi.org/10.1007/s10462-023-10456-2

[11] Samek W, Montavon G, Lapuschkin S, Anders CJ, Müller KR. Explaining deep neural networks and beyond: A review of methods and applications. *Proceedings of the IEEE*. 2021;109(3):247–278. https://doi.org/10.1109/JPROC.2021.3059992

[12] Lundberg SM, Lee SI. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*. 2017;30:4765–4774. https://doi.org/10.48550/arXiv.1705.07874

[13] Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?" Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016:1135–1144. https://doi.org/10.1145/2939672.2939778

[14] Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. *Proceedings of the 34th International Conference on Machine Learning*. 2017;70:3319–3328. https://doi.org/10.48550/arXiv.1703.01365

[15] Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. *arXiv preprint*. 2017; arXiv:1702.08608. https://doi.org/10.48550/arXiv.1702.08608

[16] Topol EJ. High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine*. 2019;25(1):44–56. https://doi.org/10.1038/s41591-018-0300-7

[17] Esteva A, Robicquet A, Ramsundar B, et al. A guide to deep learning in healthcare. *Nature Medicine*. 2019;25(1):24–29. https://doi.org/10.1038/s41591-018-0316-z

[18] Rajkomar A, Dean J, Kohane I. Machine learning in medicine. *New England Journal of Medicine*. 2019;380(14):1347–1358. https://doi.org/10.1056/NEJMra1814259

[19] Beam AL, Kohane IS. Big data and machine learning in healthcare. *JAMA*. 2018;319(13):1317–1318. https://doi.org/10.1001/jama.2017.18391

[20] Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Medicine*. 2019;17(1):195. https://doi.org/10.1186/s12916-019-1426-2

[21] Ahmad MA, Eckert C, Teredesai A. Interpretable machine learning in healthcare. *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*. 2018:559–560. https://doi.org/10.1145/3242969.3242983

[22] Holzinger A, Langs G, Denk H, Zatloukal K, Müller H. Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*. 2019;9(4):e1312. https://doi.org/10.1002/widm.1312

[23] Holzinger A, Biemann C, Pattichis CS, Kell DB. What do we need to build explainable AI systems for the medical domain? *arXiv preprint*. 2017; arXiv:1712.09923. https://doi.org/10.48550/arXiv.1712.09923

[24] Tjoa E, Guan C. A survey on explainable artificial intelligence (XAI): Toward medical XAI. *IEEE Transactions on Neural Networks and Learning Systems*. 2021;32(11):4793–4813. https://doi.org/10.1109/TNNLS.2020.3046155

[25] Arrieta AB, Díaz-Rodríguez N, Del Ser J, et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*. 2020;58:82–115. https://doi.org/10.1016/j.inffus.2019.12.012

[26] Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D. A survey of methods for explaining black box models. *ACM Computing Surveys*. 2018;51(5):93. https://doi.org/10.1145/3236009

[27] Adadi A, Berrada M. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*. 2018;6:52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

[28] Molnar C. *Interpretable Machine Learning*. 2nd ed. Lulu Press; 2022. https://christophm.github.io/interpretable-ml-book/

[29] Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*. 2019;1(5):206–215. https://doi.org/10.1038/s42256-019-0048-x

[30] Samek W, Wiegand T, Müller KR. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *ITU Journal ICT Discoveries*. 2017;1(1):39–48. https://arxiv.org/abs/1708.01870

[31] Carvalho DV, Pereira EM, Cardoso JS. Machine learning interpretability: A survey on methods and metrics. *Electronics*. 2019;8(8):832. https://doi.org/10.3390/electronics8080832

[32] Johnson KW, Torres Soto J, Glicksberg BS, et al. Artificial intelligence in cardiology. *Journal of the American College of Cardiology*. 2018;71(23):2668–2679. https://doi.org/10.1016/j.jacc.2018.03.521

[33] Handelman GS, Kok HK, Chandra RV, et al. eDoctor: Machine learning and the future of medicine. *Journal of Internal Medicine*. 2018;284(6):603–619. https://doi.org/10.1111/joim.12822

[34] Shameer K, Johnson KW, Glicksberg BS, Dudley JT, Sengupta PP. Machine learning in cardiovascular medicine. *Circulation Research*. 2018;122(11):1525–1540. https://doi.org/10.1161/CIRCRESAHA.118.311081

[35] Sidey-Gibbons JAM, Sidey-Gibbons CJ. Machine learning in medicine: A practical introduction. *BMC Medical Research Methodology*. 2019;19(1):64. https://doi.org/10.1186/s12874-019-0681-0

[36] Jiang F, Jiang Y, Zhi H, et al. Artificial intelligence in healthcare: Past, present and future. *Stroke and Vascular Neurology*. 2017;2(4):230–243. https://doi.org/10.1136/svn-2017-000101

[37] Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: Review, opportunities and challenges. *Briefings in Bioinformatics*. 2018;19(6):1236–1246. https://doi.org/10.1093/bib/bbx044

[38] Yu KH, Beam AL, Kohane IS. Artificial intelligence in healthcare. *Nature Biomedical Engineering*. 2018;2(10):719–731. https://doi.org/10.1038/s41551-018-0305-z

[39] Deo RC. Machine learning in medicine. *Circulation*. 2015;132(20):1920–1930. https://doi.org/10.1161/CIRCULATIONAHA.115.001593

[40] Obermeyer Z, Emanuel EJ. Predicting the future — Big data, machine learning and clinical medicine. *New England Journal of Medicine*. 2016;375(13):1216–1219. https://doi.org/10.1056/NEJMp1606181

[41] Chen JH, Asch SM. Machine learning and prediction in medicine — Beyond the peak of inflated expectations. *New England Journal of Medicine*. 2017;376(26):2507–2509. https://doi.org/10.1056/NEJMp1702057

[42] Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. *Nature Medicine*. 2022;28(1):31–38. https://doi.org/10.1038/s41591-021-01614-0

[43] Bhatt DL, Drozda JP, Shahian DM, et al. Artificial intelligence in cardiovascular medicine. *Journal of the American College of Cardiology*. 2023;81(11):1097–1112. https://doi.org/10.1016/j.jacc.2023.01.002

[44] Vamathevan J, Clark D, Czodrowski P, et al. Applications of machine learning in drug discovery and development. *Nature Reviews Drug Discovery*. 2019;18(6):463–477. https://doi.org/10.1038/s41573-019-0024-5

[45] Wiens J, Saria S, Sendak M, et al. Do no harm: A roadmap for responsible machine learning in healthcare. *Nature Medicine*. 2019;25(9):1337–1340. https://doi.org/10.1038/s41591-019-0548-6