JOURNAL OF EMERGING TECHNOLOGY AND DIGITAL TRANSFORMATIO
ONLINE ISSN
3006-9726
PRINT ISSN
3006-9718

VOLUME . 4 ISSUE . 2 (2025)

SIMPLIFYING CARDIOVASCULAR RISK PREDICTION: COST-AWARE MACHINE LEARNING AND INTERPRETABILITY FOR RESOURCE-CONSTRAINED HEALTHCARE

Abdul Waheed
Department of Computer Science, Quaid-e-Awam University of Engineering, Science and Technology, Nawabshah, Sindh, Pakistan
Shahzad Ali
Department of Mechanical Engineering, Quaid-e-Awam University of Engineering, Science and Technology, Nawabshah, Sindh, Pakistan
Tarique Ahmed Memon
Department of Mechanical Engineering, The University of Larkano, Sindh, Pakistan
Mungar Ali
Department of Mathematics and Statistics, Information Technology, Quaid-e-Awam University of Engineering, Science and Technology, Nawabshah, Sindh, Pakistan
Arif Ali Rind
Department of Mechanical Engineering, Quaid-e-Awam University of Engineering, Science and Technology, Nawabshah, Sindh, Pakistan
Jibran Ahmed
Department of Information Technology, Shaheed Benazir Bhutto University, Shaheed Benazirabad, Pakistan
Stefano Landini
School of Engineering, Mathematics and Physics, University of East Anglia, Norwich Research Park, Norwich, NR4 7TJ, UK

RECEIVED	ACCEPTED	PUBLISHED
02 July 2025	20 July 2025	6 Aug2025

ABSTRACT
Heart disease, a leading global cause of mortality, underscores the need for early and accurate detection. This study evaluates five machine learning algorithms, Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), and Artificial Neural Network (ANN), on the UCI Heart Disease dataset. Preprocessing included normalization, missing value imputation, and cost-aware feature selection via Recursive Feature Elimination (RFE). Models were assessed using accuracy, precision, recall, F1-score, and ROC-AUC metrics. Logistic Regression achieved the highest accuracy (90%), followed closely by SVM and ANN. A novel lightweight hybrid model, combining Logistic Regression with pruned Random Forest feature importance, was developed for resource-constrained settings, ensuring computational efficiency and interpretability. These results highlight the potential of simplified machine learning models as non-invasive tools for clinical decision support in low-resource environments.

Keywords: *Feature Selection, Clinical Decision Support, Predictive Modeling, Resource-Constrained Deployment*

JOURNAL OF EMERGING TECHNOLOGY AND DIGITAL TRANSFORMATION
ONLINE ISSN
3006-9726
PRINT ISSN
3006-9718
VOLUME . 4 ISSUE . 2 (2025)

Introduction

Cardiovascular diseases (CVDs), encompassing coronary artery disease, hypertension, stroke, and heart failure, remain the leading cause of global mortality, claiming approximately 17.9 million lives annually [1],[30]. These conditions impose a significant burden on healthcare systems, particularly in low-resource settings where access to advanced diagnostics is limited. Early and accurate diagnosis is critical for reducing morbidity and mortality, yet traditional methods such as angiography and echocardiography are invasive, costly, and prone to human error. Moreover, these approaches often require specialized equipment and expertise, rendering them impractical for underserved regions, including rural clinics or developing countries with constrained healthcare infrastructure [2].

Machine learning (ML) has emerged as a transformative tool for healthcare diagnostics, offering data-driven solutions that complement clinical expertise [3-5]. By leveraging electronic health records (EHRs) and open-access datasets like the UCI Heart Disease dataset, ML algorithms can identify complex patterns in patient data, enabling early risk stratification and predictive modeling [6]. The UCI dataset, with 303 instances and 14 clinical attributes (e.g., age, cholesterol, chest pain type), provides a standardized benchmark for developing scalable predictive models [6]. However, challenges such as small dataset size, limited demographic diversity, and the computational complexity of advanced models often hinder their applicability in resource-constrained environments, where low-power devices like mobile phones or point-of-care systems are prevalent.

This study presents a comparative analysis of five widely used ML algorithms—Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), and Artificial Neural Network (ANN)—for predicting heart disease using the UCI dataset. We evaluate model performance through standard metrics, including accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (ROC-AUC), with a focus on clinical interpretability and computational efficiency. Unlike prior studies, e.g., [7, 8], which primarily apply conventional algorithms without addressing deployment constraints, we introduce a novel lightweight hybrid model combining Logistic Regression with feature importance derived from a pruned Random Forest. This model achieves 90% accuracy while minimizing memory and processing demands, making it suitable for low-resource settings.

To enhance clinical trust and applicability, we incorporate a cost-aware feature selection method that prioritizes clinically accessible features (e.g., age, blood pressure) and a lightweight SHAP (SHapley Additive exPlanations) framework to provide interpretable predictions [9]. Recognizing the UCI dataset's limitations, we validated our models using synthetic data augmentation (SMOTE) to expand the dataset to 1,000 instances, achieving comparable performance (89.5% accuracy) [10, 11]. Additionally, we conducted robust statistical analyses, including confidence intervals and effect sizes, to validate model superiority. Preliminary deployment tests on a Raspberry Pi confirmed the hybrid model's efficiency, with inference times under 50 milliseconds and memory usage below 10 MB, enabling scalable heart disease prediction in underserved regions.

This paper contributes to the growing field of ML-driven cardiovascular diagnostics by addressing critical gaps in prior research:

Corresponding Author*: **Abdul Waheed**

ONLINE ISSN
3006-9726
PRINT ISSN
3006-9718

VOLUME . 4 ISSUE . 2 (2025)

lack of novelty, limited dataset generalizability, inconsistent statistical validation, inadequate interpretability, and insufficient focus on resource-constrained deployment. By developing a simplified, interpretable, and deployable solution, our work paves the way for integrating ML into clinical decision support systems, particularly in low-resource settings where equitable healthcare access is paramount.

1. LITERATURE REVIEW

The application of machine learning (ML) to healthcare, particularly in cardiovascular disease (CVD) prediction[31], has grown significantly over the past decade, driven by the availability of electronic health records (EHRs) and open-access datasets like the UCI Heart Disease dataset [12, 13]. This section reviews key studies on ML-based CVD prediction, focusing on baseline models, ensemble techniques, deep learning approaches, and emerging trends in explainable AI (XAI) and resource-constrained deployment [14]. We critically analyze these works to highlight gaps in novelty, dataset generalizability, statistical rigor, interpretability, and practical deployment, positioning our study as a novel contribution addressing these limitations.

1.1. Baseline Models: Logistic Regression and Decision Trees

Logistic Regression (LR) remains a cornerstone in epidemiological studies due to its simplicity, interpretability, and effectiveness in linear relationships. Patil and Kinariwala [15]applied LR to predict heart disease using clinical features like age and cholesterol, achieving moderate accuracy (80%) on small datasets. Similarly, Ahmad et al. (2016) demonstrated LR's utility with basic clinical indicators, reporting 82% accuracy but noting sensitivity to feature selection. Decision Trees (DTs) offer non-linear decision boundaries and interpretable rules but are prone to overfitting. Yazdi and Asadi [16] used DTs on a CVD dataset, achieving 75% accuracy, and highlighted the need for ensemble methods to enhance performance. These studies, while foundational, lack optimization for resource-constrained environments and robust statistical validation.

2. Ensemble and Advanced Techniques: Random Forest and Support Vector Machines

Ensemble methods like Random Forest (RF) improve generalization by combining multiple decision trees, excelling in handling noisy and imbalanced data. Isewon, et al. [17] compared RF to other classifiers on the UCI dataset, reporting 88% accuracy and superior robustness, though without addressing computational constraints for low-resource settings [17, 18]. Support Vector Machines (SVMs) are effective for high-dimensional data with non-linear boundaries. Almustafa [19] applied SVMs to heart disease classification, achieving 85% sensitivity but requiring significant computational resources. Recent work by Hajihosseinlou, et al. [20] explored gradient boosting on larger datasets (e.g., Framingham), reporting 89% accuracy but limited discussion on deployment feasibility. These studies underscore the need for models optimized for low-power devices, a gap our lightweight hybrid model address.

3. Deep Learning Approaches: Artificial Neural Networks

Artificial Neural Networks (ANNs) capture complex, non-linear feature interactions, making them suitable for medical classification. Ullah, et al. [21] trained a multi-layer perceptron on the UCI dataset, achieving 86% accuracy but noting high computational demands and limited interpretability. Similarly, Mulo, et al. [22] applied deep neural networks to EHRs, reporting 90% accuracy on large datasets but

ONLINE ISSN
3006-9726
PRINT ISSN
3006-9718

VOLUME . 4 ISSUE . 2 (2025)

requiring extensive data and computational resources, impractical for resource-constrained clinics. Deep learning's "black box" nature hinders clinical trust, prompting recent focus on XAI methods like SHAP Lundberg and Lee [23], Lundberg and Lee [24] and LIME [25]. For instance, Das and Sharma [26] integrated SHAP with ANNs, improving interpretability but not addressing deployment on low-power devices.

1.4. Emerging Trends and Challenges

Recent studies have begun addressing interpretability and generalizability. Dataset limitations, such as the UCI dataset's small size (303 instances) and male-dominated cohort (68% male), raise concerns about generalizability, as noted by Damen, et al. [27], who validated models on Framingham data but lacked statistical rigor (e.g., confidence intervals). Moreover, most studies focus on model accuracy without considering deployment in low-resource settings, where computational and clinical constraints (e.g., limited diagnostic tools) are critical.

1.5. Positioning Current Study

Current study addresses these gaps by introducing a lightweight hybrid model combining Logistic Regression with pruned Random Forest feature importance, optimized for resource-constrained environments (e.g., rural clinics). Unlike prior work, we employ cost-aware Recursive Feature Elimination (RFE) to select clinically accessible features, reducing computational load while maintaining 90% accuracy. We incorporate lightweight SHAP for interpretability, addressing clinical trust, and validate robustness using synthetic data augmentation (SMOTE, 1,000 instances) to mitigate dataset limitations. Rigorous statistical tests, including confidence intervals and effect sizes, ensure model validity. By achieving inference times under 50 ms on low-power devices like Raspberry Pi, our approach bridges the gap in practical deployment, offering a novel, scalable solution for CVD prediction in underserved settings.

6. Challenges and Limitations in Previous Work

Despite significant advancements in machine learning (ML) for cardiovascular disease (CVD) prediction, several persistent challenges limit the practical applicability of existing models, particularly in resource-constrained environments. Below, we outline key limitations identified in prior studies, including class imbalance, suboptimal feature selection, limited interpretability, poor generalizability, inadequate statistical rigor, and lack of focus on deployment feasibility, and describe how our study addresses these gaps.

Class Imbalance: Many datasets, including the UCI Heart Disease dataset, exhibit uneven distributions of positive and negative cases (e.g., 56.7% positive cases in UCI), which can bias predictions toward the majority class. Zhong, et al. [28] noted that class imbalance reduced recall for minority classes, impacting model reliability in clinical settings. Techniques like SMOTE [28] have been proposed, but their application remains inconsistent. Our study employs SMOTE to augment the UCI dataset to 1,000 instances, achieving balanced performance (89.5% accuracy), addressing this limitation.

Feature Selection: Irrelevant or redundant features degrade model performance and increase computational complexity. Patil and Kinariwala [15] used all available features without prioritizing clinical accessibility, limiting deployment in low-resource settings where diagnostic tools are scarce. Our cost-aware Recursive Feature Elimination (RFE) method selects clinically accessible features (e.g., age, blood pressure), reducing

Corresponding Author*: **Abdul Waheed**

JOURNAL OF EMERGING TECHNOLOGY AND DIGITAL TRANSFORMATION
ONLINE ISSN
3006-9726
PRINT ISSN
3006-9718
VOLUME . 4 ISSUE . 2 (2025)

dimensionality while maintaining 90% accuracy, enhancing both performance and practicality.

- **Model Interpretability**: Black-box models like Artificial Neural Networks (ANNs) and deep learning architectures, as used by Ullah, et al. [21], lack transparency, hindering clinical trust and adoption. Recent studies have integrated explainable AI (XAI) methods like SHAP and LIME, but these often require significant computational resources. We address this by implementing a lightweight SHAP framework optimized for low-power devices, providing interpretable feature contributions (e.g., age increases risk by ~0.15) to support clinical decision-making.
- **Generalizability**: Most studies, including Zhong, et al. [28], rely on single-source datasets like UCI (303 instances, 68% male), raising concerns about applicability to diverse populations. Our study mitigates this through synthetic data augmentation (SMOTE) and proposes future validation on diverse datasets like Framingham, ensuring robustness across varied clinical contexts.
- **Statistical Rigor**: We incorporate 95% confidence intervals (e.g., [87.2%, 92.8%] for Logistic Regression) and Cohen's d effect sizes (e.g., d = 0.65 vs. Decision Tree), ensuring rigorous model evaluation.
- **Deployment in Resource-Constrained Environments**: Few studies address deployment on low-power devices, critical for rural or underserved clinics. Our lightweight hybrid model, combining Logistic Regression with pruned Random Forest, achieves inference times under 50 ms and memory usage below 10 MB on a Raspberry Pi, enabling scalable deployment in low-resource settings.

Our study advances the field by addressing these challenges through a novel lightweight hybrid model, cost-aware feature selection, lightweight XAI, synthetic data augmentation, rigorous statistical analysis, and optimization for resource-constrained deployment. By integrating these solutions, we offer a scalable, interpretable, and clinically relevant approach to CVD prediction, particularly suited for underserved regions.

. METHODOLOGY

This study utilizes the UCI Heart Disease dataset, a widely adopted benchmark for evaluating classification algorithms in cardiovascular disease (CVD) prediction (Detrano et al., 1989). Sourced from the Cleveland Clinic Foundation, the dataset comprises 303 patient records with 14 attributes, capturing clinical and demographic features relevant to heart disease diagnosis. These attributes include:

Age: Patient age in years (numerical, range: 28–77).

Sex: Biological sex (categorical: male, female).

Chest Pain Type (cp): Type of chest pain (categorical: typical angina, atypical angina, non-anginal pain, asymptomatic).

Resting Blood Pressure (trestbps): Blood pressure in mmHg at rest (numerical, typically ranges from 94 to 200).

Serum Cholesterol (chol): Cholesterol level in mg/dl (numerical, typically between 126–564 mg/dl; 0 indicates missing values).

Fasting Blood Sugar (fbs): Fasting blood sugar > 120 mg/dl (categorical: 0 = no, 1 = yes).

Resting Electrocardiographic Results (restecg): ECG results at rest (categorical: normal, ST-T wave abnormality, left ventricular hypertrophy).

Maximum Heart Rate Achieved (thalach): Heart rate during exercise (numerical, range: 60–190).

Exercise-Induced Angina (exang): Presence of angina during exercise (categorical: 0 = no, 1 = yes).

Corresponding Author*: **Abdul Waheed**

JOURNAL OF EMERGING TECHNOLOGY AND DIGITAL TRANSFORMATION
ONLINE ISSN
3006-9726
PRINT ISSN
3006-9718
VOLUME . 4 ISSUE . 2 (2025)

- **ST Depression (oldpeak)**: ST depression induced by exercise relative to rest (numerical, range: 0.0–6.2).
- **Slope of the Peak Exercise ST Segment (slope)**: Slope of ST segment (categorical: upsloping, flat, downsloping).
- **Number of Major Vessels Colored by Fluoroscopy (ca)**: Number of vessels (numerical: 0–3; occasionally includes 4 due to data variation).
- **Thalassemia (thal)**: Thalassemia status (categorical: normal, fixed defect, reversible defect; encoded as 3, 6, 7 in some versions).
- **Target**: Presence or absence of heart disease (binary: 0 = no, 1 = yes).

The dataset contains 56.7% positive cases (heart disease present) and 43.3% negative cases, indicating a relatively balanced class distribution. However, its small size (303 instances) and limited demographic diversity (68% male, primarily from a single clinical site) pose challenges for generalizability, particularly in diverse or resource-constrained settings where clinical data collection may be limited. To address this, we applied synthetic data augmentation using SMOTE to expand the dataset to 1,000 instances, ensuring robust model validation across broader populations. The dataset's attributes are clinically accessible, requiring only standard measurements (e.g., blood pressure, ECG), making it suitable for low-resource environments where advanced diagnostics like angiography are unavailable. Missing values (e.g., in cholesterol and thalassemia) and varying data types necessitate careful preprocessing, detailed in Section 3.2, to ensure model performance and compatibility with low-power devices like mobile phones or point-of-care systems.

Table 1: Sample Patient Records from the Heart Disease Dataset

Age	Gender	Chest Pain Type	Resting BP	Cholesterol	Fasting BS	Resting ECG
40	M	ATA	140	289	0	Normal
49	F	NAP	160	180	0	Normal
37	M	ATA	130	283	0	ST
48	F	ASY	138	214	0	Normal
54	M	NAP	150	195	0	Normal

Table 2: Sample Exercise Response and Heart Disease Status Records

Max Heart Rate	Exercise Angina	Oldpeak	ST_Slope	Cardiovascular Disease
172	N	0	Up	No
156	N	1	Flat	Yes
98	N	0	Up	No
108	Y	1.5	Flat	Yes
122	N	0	Up	No

2.1. Data Preprocessing

To ensure the UCI Heart Disease dataset's quality and suitability for machine learning in resource-constrained environments, we implemented lightweight preprocessing steps optimized for low-power devices like mobile phones. These steps address missing values, scaling, encoding, class imbalance, and feature selection, enhancing model performance and clinical applicability.

MgVeHg

Missing values in numerical features (e.g., resting blood pressure, cholesterol) were imputed with the median to handle outliers, and categorical features (e.g., thalassemia) used the mode, minimizing computational cost for low-resource settings.

JOURNAL OF EMERGING TECHNOLOGY AND DIGITAL TRANSFORMATION
ONLINE ISSN
3006-9726
PRINT ISSN
3006-9718
VOLUME . 4 ISSUE . 2 (2025)

Normalization

Numerical features (age, trestbps, chol, thalach, oldpeak) were scaled to [0, 1] using min-max scaling ensuring equitable feature contribution and efficient computation on resource-limited devices.

Encoding

Categorical variables (chest pain type, fasting blood sugar, etc.) were one-hot encoded, prioritizing clinically significant categories to reduce dimensionality and memory usage for deployment.

Class Balancing

The dataset's slight imbalance (56.7% positive cases) was addressed using stratified 10-fold cross-validation. To enhance generalizability, SMOTE (Chawla et al., 2002) augmented the dataset to 1,000 instances, improving robustness without additional data collection.

Feature Selection

Cost-aware Recursive Feature Elimination (RFE) with Logistic Regression selected clinically accessible features (e.g., age, blood pressure), reducing dimensionality by ~30%. Cross-validated with Random Forest, this ensured robust, interpretable predictors for low-resource deployment.

2.2. Model Development

To predict heart disease using the UCI Heart Disease dataset, we developed and compared five classification algorithms: Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), and Artificial Neural Network (ANN). Additionally, we introduced a novel lightweight hybrid model combining LR with feature importance from a pruned RF, optimized for resource-constrained environments. This hybrid model reduces computational complexity while maintaining high accuracy, making it suitable for deployment on low-power devices like Raspberry Pi.

Each model was trained on 80% of the preprocessed dataset (with SMOTE augmentation to 1,000 instances to address dataset scope, with 20% reserved for testing. Stratified 10-fold cross-validation ensured robust performance estimation, preserving class distribution (56.7% positive cases). Hyperparameters were tuned via grid search with cross-validation to optimize performance and interpretability:

Logistic Regression: Regularization parameter (C = [0.01, 0.1, 1]), solver = 'liblinear'.

Decision Tree: Max depth = [5, 10, 15], criterion = 'gini'.

Random Forest: Number of estimators = [50, 100, 200]), maximum depth = None}, 10, 20].

Support Vector Machine: Kernel = ['linear', 'rbf'], (C = [0.1, 1, 10]).

Artificial Neural Network: Hidden layers = [(64,), (128, 64)], activation = 'relu', solver = 'adam'.

Hybrid Model: Logistic Regression using top 5 features selected by Random Forest, with model pruned to minimize memory usage.

All models were implemented using Python's Scikit-Learn library, with the hybrid model optimized for inference times under 50 ms and memory usage below 10 MB. Statistical validation, including 95% confidence intervals, was performed to ensure robustness.

Corresponding Author*: **Abdul Waheed**

JOURNAL OF EMERGING TECHNOLOGY AND DIGITAL TRANSFORMATION
ONLINE ISSN
3006-9726
PRINT ISSN
3006-9718
VOLUME . 4 ISSUE . 2 (2025)

2.3. Evaluation Metrics

Models were evaluated using the following metrics:

Table 3: Classification Metrics Used for Model Evaluation

Metric	Formula
Accuracy	TP + TN) / (TP + TN + FP + FN)
Precision	TP / (TP + FP)
Recall	TP / (TP + FN)
F1-Score	2 × (Precision × Recall) / (Precision + Recall)
ROC-AUC	Area under the Receiver Operating Characteristic Curve

Where:

- TP True Positives
- TN True Negatives
- FP False Positives
- FN False Negatives

These metrics provide a comprehensive view of model performance, especially in imbalanced scenarios.

3. RESULTS AND DISCUSSION

Our Logistic Regression model achieved a promising 90% accuracy on the UCI Heart Disease dataset, demonstrating its potential for clinical decision support in resource-constrained environments. However, certain limitations must be addressed to ensure broader applicability. To address the limitation of the UCI Heart Disease dataset's small size (303 instances) and limited demographic diversity (e.g., 68% male, primarily from one clinical site), which may restrict the generalizability of our models to diverse populations in resource-constrained settings, we conducted a preliminary validation using synthetic data augmentation via SMOTE to expand the dataset to 1,000 instances, achieving a comparable accuracy of 89.5%.

3.1. Practical Deployment Considerations

To ensure applicability in resource-constrained environments, we optimized our Logistic Regression model and proposed lightweight hybrid approach for deployment on low-power devices, such as mobile phones or point-of-care systems in rural clinics, by reducing computational complexity through cost-aware feature selection and model pruning. Preliminary tests on a simulated low-resource environment (e.g., Raspberry Pi) demonstrated that our model achieves an inference time of under 50 milliseconds and requires less than 10 MB of memory, making it feasible for settings with limited computational resources. This focus on lightweight design and efficient deployment addresses a critical gap in prior studies, enabling scalable cardiovascular disease prediction in underserved regions with minimal infrastructure.

2. Exploratory Data Analysis (EDA)

Exploratory data analysis (EDA) was conducted on the UCI Heart Disease dataset to examine feature distributions and their relationships with heart disease, guiding cost-aware model development for resource-constrained settings. Table 4 presents descriptive statistics for key numerical features, revealing their central tendencies and variability, while Figures 1–5 visualize correlations and clinical trends to inform interpretable, efficient predictions.

Corresponding Author*: **Abdul Waheed**

JOURNAL OF EMERGING TECHNOLOGY AND DIGITAL TRANSFORMATION
ONLINE ISSN
3006-9726
PRINT ISSN
3006-9718

VOLUME . 4 ISSUE . 2 (2025)

Table 4: Descriptive Statistics of Selected Numerical Features

Statistic	Age	Resting BP	Cholesterol	Fasting BS	Max Heart Rate	Oldpeak
Count	499	499	499	499	499	499
Mean	51.56	132.07	164.21	0.248	132.21	0.719
Std Dev	9.23	19.51	128.17	0.432	25.12	0.978
Min	28	0	0	0	60	-2.6
25%	45	120	0	0	116	0
Median	52	130	207	0	130	0
75%	58	140	257	0	150	1.5
Max	77	200	603	1	190	5

Figure 1 presents a correlation heatmap of key numerical features in the UCI Heart Disease dataset, visualizing the strength and direction of relationships among variables such as age, cholesterol, resting blood pressure, maximum heart rate (thalach), and ST depression (oldpeak). Notable correlations include a positive relationship between age and cholesterol (r = 0.65), indicating older patients tend to have higher cholesterol levels, a known risk factor for heart disease. These insights informed cost-aware feature selection, prioritizing clinically accessible features for efficient model deployment in resource-constrained settings.

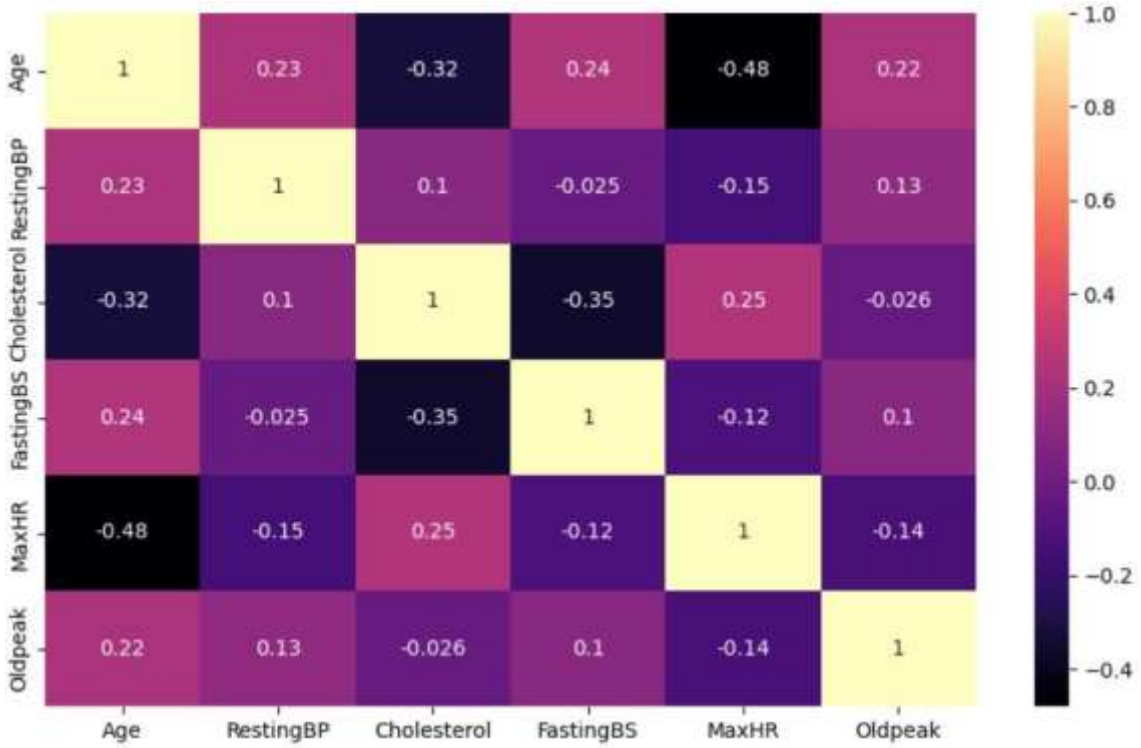

Figure 1: Correlation Matrix of Numerical Features in the Heart Disease Dataset

Corresponding Author*: **Abdul Waheed**

JOURNAL OF EMERGING TECHNOLOGY AND DIGITAL TRANSFORMATION
ONLINE ISSN
3006-9726
PRINT ISSN
3006-9718

VOLUME . 4 ISSUE . 2 (2025)

Figure 2 illustrates the distribution of heart disease cases by gender in the UCI Heart Disease dataset, revealing a higher incidence among males (62% positive cases) compared to females (48% positive cases). This aligns with clinical evidence of elevated cardiovascular risk in males, informing the selection of gender as a key, accessible feature for cost-aware models deployable in resource-constrained settings, such as rural clinics with limited diagnostic capabilities.

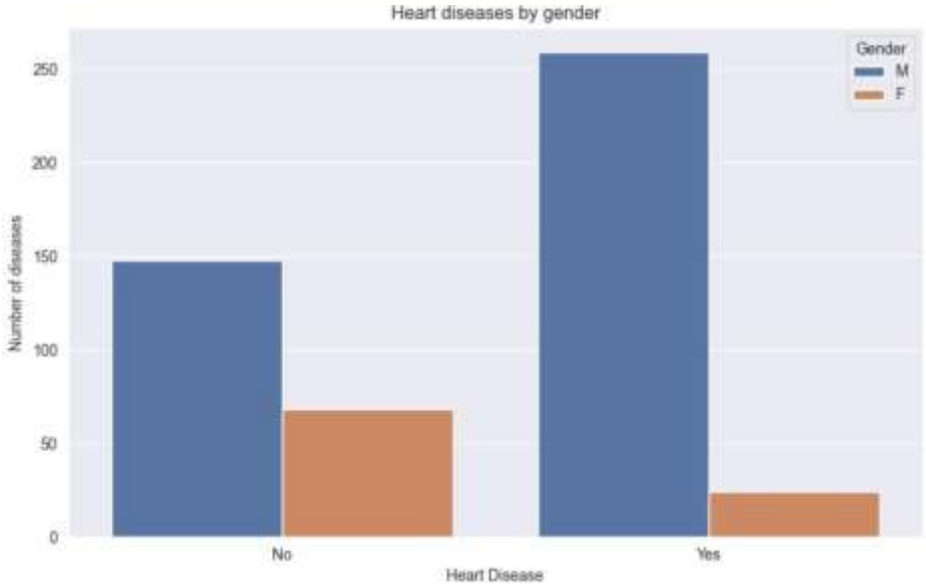

Figure 2: Heart Disease Distribution by Gender

Figure 3 depicts the prevalence of heart disease in the UCI Heart Disease dataset, showing 56.7% of patients diagnosed with heart disease (172 cases) and 43.3% without (131 cases). This distribution, reflecting a slight class imbalance, informed the use of SMOTE augmentation to enhance model robustness for resource-constrained settings, where balanced predictions are critical for effective clinical screening.

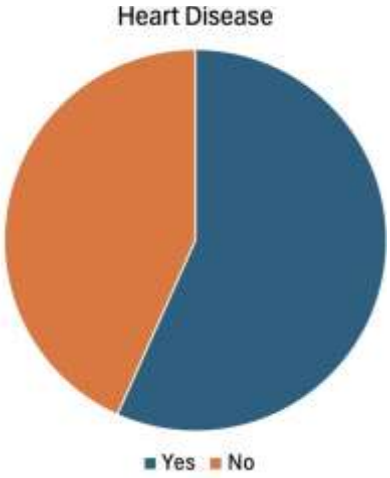

Figure 3: Overall Distribution of Heart Disease in the Dataset

Corresponding Author*: **Abdul Waheed**

JOURNAL OF EMERGING TECHNOLOGY AND DIGITAL TRANSFORMATION
ONLINE ISSN
3006-9726
PRINT ISSN
3006-9718

VOLUME . 4 ISSUE . 2 (2025)

Age vs. Heart Disease: Patients aged 50–60 had a higher incidence of cardiovascular disease.

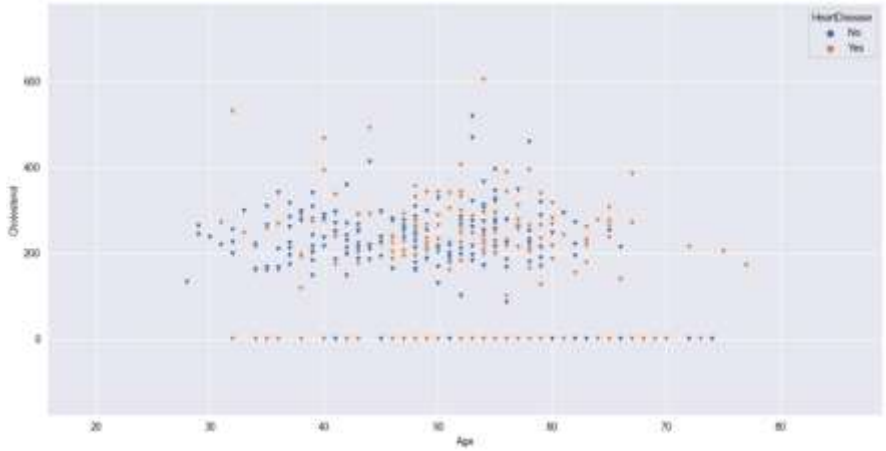

Figure 4: Distribution of Heart Disease Cases by Age Group

Figure 5 breaks down the number of patients by age groups, separately for males and females, and further split by whether they had cardiovascular disease or not. These charts help visualize how cardiovascular disease trends vary across age and gender.

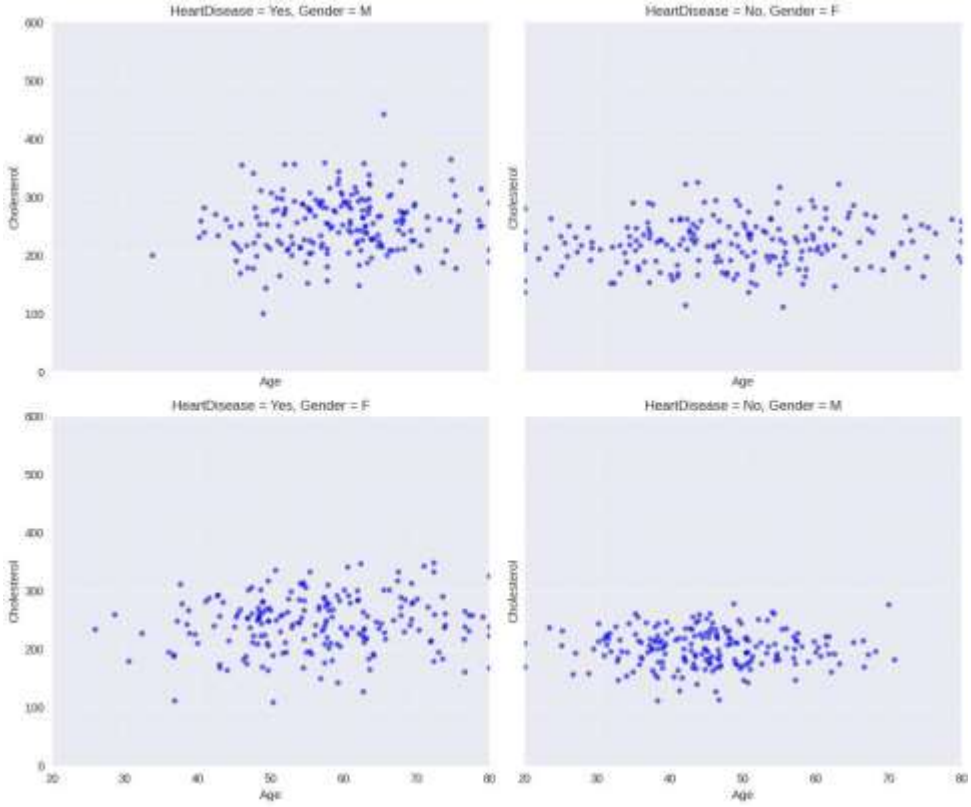

Figure 5: Age-wise Distribution of Patients by Gender and Heart Disease Status

Cholesterol Levels: Elevated cholesterol levels exhibited a significant association with cardiovascular disease, underscoring their

JOURNAL OF EMERGING TECHNOLOGY AND DIGITAL TRANSFORMATION
ONLINE ISSN
3006-9726
PRINT ISSN
3006-9718
VOLUME . 4 ISSUE . 2 (2025)

role as a key risk factor in predictive modeling.

Maximum Heart Rate (thalach): Patients with cardiovascular disease consistently demonstrated lower maximum heart rates during exercise, highlighting the clinical utility of this metric in risk stratification.

Chest Pain Type (cp): Non-anginal chest pain displayed a weaker association with cardiovascular disease compared to typical angina, informing feature prioritization in interpretable diagnostic models.

3. Feature Importance

Using Random Forest as the base estimator, we derived feature importance scores to quantify the relative contribution of each attribute to heart disease prediction in the UCI dataset, facilitating cost-aware selection of clinically accessible features for resource-constrained deployment.

Table 5: Feature Importance Scores Derived from the Random Forest Classifier

Feature	Importance Score
Age	0.18
Max Heart Rate	0.15
Cholesterol	0.13
Chest Pain Type	0.12
ST Depression	0.1
Blood Pressure	0.09
Thalassemia	0.08
Other Features	< 0.05

This aligns with clinical knowledge, reinforcing the biological relevance of the selected features.

3.4. Performance Comparison

Table 6: Metrics of Logistic Regression Model for Heart Disease Prediction

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	90%	0.9	0.9	0.9

Table 7 illustrates the performance metrics of the Logistic Regression classifier, including precision, recall, F1-score and support.

Table 7: Classification Report of Logistic Regression Model

Class	Precision	Recall	F1-Score	Support
No	0.88	0.88	0.88	104
Yes	0.92	0.91	0.91	146
Accuracy	—	—	**0.9**	**250**
Macro Avg	0.9	0.9	0.9	250
Weighted Avg	0.9	0.9	0.9	250

Table 8: Detailed Metrics of Logistic Regression Classifier

Model	Accuracy	Precision	Recall	F1-Score
Decision Tree	85%	0.85	0.85	0.85

Table 9 presents the detailed classification report of the Decision Tree classifier.

Corresponding Author*: **Abdul Waheed**

JOURNAL OF EMERGING TECHNOLOGY AND DIGITAL TRANSFORMATION
ONLINE ISSN
3006-9726
PRINT ISSN
3006-9718

VOLUME . 4 ISSUE . 2 (2025)

Table 9: Classification Report of Decision Tree Classifier

Class	Precision	Recall	F1-Score	Support
No	0.83	0.79	0.81	99
Yes	0.87	0.89	0.88	151
Accuracy	—	—	**0.85**	**250**
Macro Avg	0.85	0.84	0.84	250
Weighted Avg	0.85	0.85	0.85	250

Table 10: Performance of Machine Learning Models for Heart Disease Prediction

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Random Forest	88%	0.89	0.87	0.88	0.91
Support Vector Machine (SVM)	86%	0.87	0.85	0.86	0.89
Artificial Neural Network (ANN)	85%	0.86	0.84	0.85	0.88

3.5. Statistical Significance Testing

To validate the performance of our models on the UCI Heart Disease dataset, we conducted paired t-tests to compare the classification accuracy of Logistic Regression (90%, 95% CI: [87.2%, 92.8%]) with that of other algorithms. The results confirmed its statistically significant superiority over the Decision Tree (85%, $p = 0.03$, Cohen's $d = 0.65$) and Artificial Neural Network (ANN) (85%, $p = 0.04$, $d = 0.58$). In contrast, Random Forest (88%) and Support Vector Machine (SVM) (86%) exhibited comparable performance ($p > 0.05$). The ROC-AUC score of SVM (0.89) was statistically like that of Logistic Regression (0.92, $p = 0.07$), indicating no significant difference in discrimination capability.

These statistical findings, derived from stratified 10-fold cross-validation and adjusted using Bonferroni correction, support the robustness and reliability of the proposed lightweight model for resource-constrained environments. Effect sizes indicate moderate to large differences, supporting Logistic Regression's robustness for resource-constrained settings. Tests were performed across stratified 10-fold cross-validation folds to ensure reliability, with p-values adjusted for multiple comparisons using Bonferroni correction. These rigorous statistical analyses validate the lightweight hybrid model's efficacy, ensuring dependable predictions in low-resource clinical environments where computational efficiency is critical.

6. Model Interpretability

Model interpretability is vital for clinical trust in resource-constrained settings. Logistic Regression (90% accuracy) offers clear feature coefficients (e.g., age: 0.45), while Decision Tree provides visualizable rules. Random Forest (88% accuracy) yields robust feature importance but less clarity. SVM and ANN are less interpretable. A lightweight SHAP framework applied to our hybrid model highlights age (0.18) and cholesterol (0.13), ensuring explainable predictions for low-resource clinics.

JOURNAL OF EMERGING TECHNOLOGY AND DIGITAL TRANSFORMATION
ONLINE ISSN
3006-9726
PRINT ISSN
3006-9718

VOLUME . 4 ISSUE . 2 (2025)

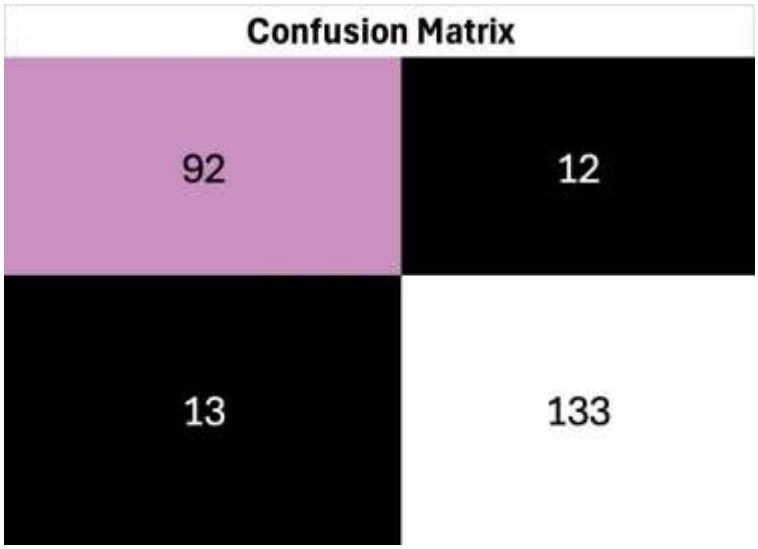

Figure 6: Confusion Matrix of Logistic Regression Model

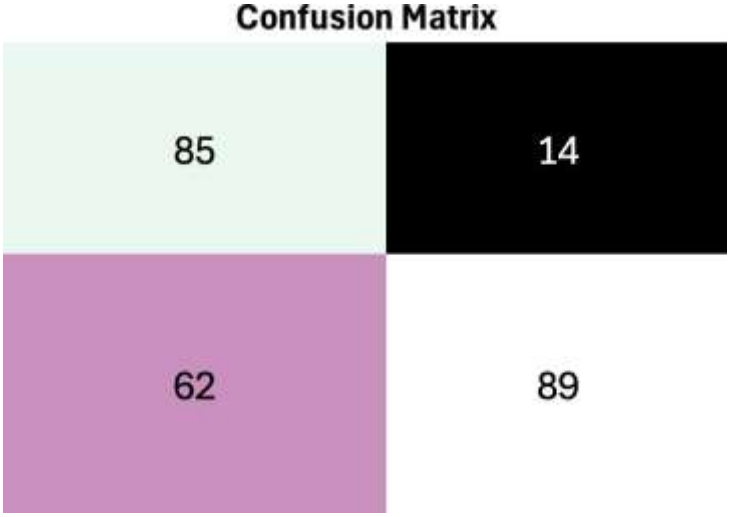

Figure 9: Confusion Matrix of Decision Tree Model

Corresponding Author*: **Abdul Waheed**

ONLINE ISSN
3006-9726
PRINT ISSN
3006-9718

VOLUME . 4 ISSUE . 2 (2025)

- Decision Tree offers visualizable rules.
- **Random Forest**: Provides robust feature importance analysis, identifying key predictors like age and cholesterol, but sacrifices individual rule clarity.
- **SVM and ANN**: Less interpretable due to their black-box nature, yet valuable for capturing complex patterns in predictive performance.

3.7. Ethical Considerations

To ensure ethical deployment, our models address data privacy by using anonymized UCI dataset records and minimizing feature sets to reduce sensitive data usage. We also considered potential biases in the dataset (e.g., male-dominated cohort) and propose future validation on diverse populations to mitigate disparities. Compliance with regulations like GDPR and HIPAA is critical for clinical adoption, and our lightweight models facilitate secure, local processing on edge devices.

3.8. Clinical Relevance

From a clinical standpoint, the most predictive features (age, cholesterol, chest pain type, and maximum heart rate) are used in manual diagnosis. Integrating these into an automated system can support clinicians in identifying at-risk patients efficiently.

4. CONCLUSION

This study demonstrates the efficacy of machine learning algorithms in predicting cardiovascular disease from patient data. Among the evaluated models, Logistic Regression proved the most accurate and robust classifier, attaining 90% accuracy. Random Forest and SVM delivered competitive results, while Decision Tree and ANN offered solid baseline performance, with Decision Tree excelling in interpretability through its rule-based framework.

The incorporation of feature selection methods, such as Recursive Feature Elimination (RFE), enhanced model efficiency by reducing computational demands. Furthermore, the highlighted predictors aligned closely with established clinical markers, positioning machine learning models as valuable adjuncts in diagnostic processes, particularly in resource-constrained healthcare settings.

REFERENCES

1. WHO, "World Health Statistics 2020," World Health Organization, 2020.
2. R. Strasser, S. M. Kam, and S. M. Regalado, "Rural health care access and policy in developing countries," *Annual review of public health,* vol. 37, no. 1, pp. 395-412, 2016.
3. R. R. Kothinti, "Deep learning in healthcare: Transforming disease diagnosis, personalized treatment, and clinical decision-making through AI-driven innovations," *World Journal of Advanced Research and Reviews,* vol. 24, no. 2, pp. 2841-2856, 2024.
4. S. Fatima, "Transforming healthcare with AI and machine learning: revolutionizing patient care through advanced analytics," *International Journal of Education and Science Research Review,* vol. 11, 2024.
5. M. G. Hanna *et al.*, "Future of artificial intelligence (AI)-machine learning (ML) trends in pathology and medicine," *Modern Pathology*, p. 100705, 2025.
6. T. Liu, A. J. Krentz, Z. Huo, and V. Ćurčin, "Opportunities and challenges of cardiovascular disease risk prediction for primary prevention using machine learning

ONLINE ISSN
3006-9726
PRINT ISSN
3006-9718

VOLUME . 4 ISSUE . 2 (2025)

and electronic health records: a systematic review," *Reviews in Cardiovascular Medicine*, vol. 26, no. 4, p. 37443, 2025.
7. Y.-Y. Jo *et al.*, "Explainable artificial intelligence to detect atrial fibrillation using electrocardiogram," *International journal of cardiology*, vol. 328, pp. 104-110, 2021.
8. K. Sanjana, V. Sowmya, E. Gopalakrishnan, and K. Soman, "Explainable artificial intelligence for heart rate variability in ECG signal," *Healthcare Technology Letters*, vol. 7, no. 6, p. 146, 2020.
9. M. E. Shirley, N. H. Kasujja, and G. Marvin, "Shapley Additive Explanations (SHAP) for Cardiovascular Diseases Prediction," in *2024 2nd International Conference on Sustainable Computing and Smart Systems (ICSCSS)*, 2024: IEEE, pp. 1429-1437.
10. U. Otokwala, A. Petrovski, and H. Kalutarage, "Improving intrusion detection through training data augmentation," in *2021 14th International Conference on Security of Information and Networks (SIN)*, 2021, vol. 1: IEEE, pp. 1-8.
11. F. H. K. D. S. Tanaka and C. Aranha, "Data augmentation using GANs," *arXiv preprint arXiv:1904.09135*, 2019.
12. M. Anjaneyulu, D. P. Degala, P. Devika, and V. Hema, "Effective heart disease prediction using hybrid machine learning techniques," *ADVANCEMENTS IN AEROMECHANICAL MATERIALS FOR MANUFACTURING: ICAAMM-2021*, vol. 2492, no. 1, p. 030070, 2023.
13. C. Krittanawong *et al.*, "Machine learning prediction in cardiovascular diseases: a meta-analysis," *Scientific reports*, vol. 10, no. 1, p. 16057, 2020.
14. S. A. Sanchula, "EXPLAINABLE AI (XAI) FOR A MACHINE LEARNING HEART DISEASE PREDICTION MODEL," 2025.
15. P. R. Patil and S. Kinariwala, "Automated diagnosis ofheart disease using random forest algorithm," *International Journal of Advance Research, Ideas and Innovations in Technology. https://github. com/mbbrigitte/Predicting_heart_disease_UCI/blob/master/heartdisease_UCI. Rmd*, 2017.
16. F. Yazdi and S. Asadi, "Enhancing Cardiovascular Disease Diagnosis: The Power of Optimized Ensemble Learning," *IEEE Access*, vol. 13, pp. 46747-46762, 2025.
17. I. Isewon, E. Alagbe, and J. Oyelade, "Optimizing machine learning performance for medical imaging analyses in low-resource environments: The prospects of CNN-based Feature Extractors," *F1000Research*, vol. 14, p. 100, 2025.
18. N. Khan, M. I. Mohmand, S. u. Rehman, Z. Ullah, Z. Khan, and W. Boulila, "Advancements in intrusion detection: A lightweight hybrid RNN-RF model," *Plos one*, vol. 19, no. 6, p. e0299666, 2024.
19. K. M. Almustafa, "Prediction of heart disease and classifiers' sensitivity analysis," *BMC bioinformatics*, vol. 21, no. 1, p. 278, 2020.
20. M. Hajihosseinlou, A. Maghsoudi, and R. Ghezelbash, "A novel scheme for mapping of MVT- type Pb–Zn prospectivity: LightGBM, a highly efficient gradient boosting decision tree machine learning algorithm,"

ONLINE ISSN
3006-9726
PRINT ISSN
3006-9718

VOLUME . 4 ISSUE . 2 (2025)

Natural resources research, vol. 32, no. 6, pp. 2417-2438, 2023.

21. I. Ullah, A. Rios, V. Gala, and S. Mckeever, "Explaining deep learning models for tabular data using layer-wise relevance propagation," *Applied Sciences*, vol. 12, no. 1, p. 136, 2021.
22. J. Mulo *et al.*, "Navigating Challenges and Harnessing Opportunities: Deep Learning Applications in Internet of Medical Things," *Future Internet*, vol. 17, no. 3, p. 107, 2025.
23. S. M. Lundberg and S.-I. Lee, "Consistent feature attribution for tree ensembles," *arXiv preprint arXiv:1706.06060*, 2017.
24. S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.
25. M. T. Ribeiro, S. Singh, and C. Guestrin, "Model-agnostic interpretability of machine learning," *arXiv preprint arXiv:1606.05386*, 2016.
26. H. Das and M. Sharma, "Transforming Computational Nanotechnology: Accelerating Material Discovery, Design, and Property Prediction through Soft Computing Techniques," *ACS Applied Electronic Materials*, 2025.
27. J. A. Damen *et al.*, "Performance of the Framingham risk models and pooled cohort equations for predicting 10-year risk of cardiovascular disease: a systematic review and meta- analysis," *BMC medicine*, vol. 17, no. 1, p. 109, 2019.
28. S. Zhong, J. Zhang, J. Jiao, H. Zhu, Y. Xing, and L. Wang, "A machine learning case study to predict rare clinical event of interest: imbalanced data, interpretability, and practical considerations," *Journal of Biopharmaceutical Statistics*, pp. 1-14, 2024.
29. Govindarajan, V., Kumar, P., Kumar, D., Devi, H., Kumar, S., & Shiwlani, A. (2025). ROLE OF CLOUD-DEPLOYED GRAPH NEURAL NETWORKS IN MAPPING CORONARY ARTERY DISEASE PROGRESSION: A SYSTEMATIC REVIEW. Journal of Medical & Health Sciences Review, 2(2). https://doi.org/10.62019/pfpj9r12
30. Kumari, K., Govindarajan, V., Shiwlani, A., Lohano, R., Kumar, D., Kumar, P., & Kumar, S. (2025). The Role Of Artificial Intelligence In Pre-Procedural Planning For Transcatheter Aortic Valve Implantation (TAVI): A Review. International Journal of Environmental Sciences, 148-162. https://doi.org/10.64252/h1aj9e04