Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 4, Issue 4, 2026

HYBRID STATISTICAL LEARNING FOR CARDIOVASCULAR DISEASE CLASSIFICATION USING LIGHTGBM WITH FEATURE OPTIMIZATION

Hina Zafar[*1], Saba Akram[2], Muhammad Hamza Kashif[3], Syed Muhammad Junaid Hassan[*4]

[1,2]*Lecturer Department of Software Engineering, National University of Modern Languages, Faisalabad.*
[3]*Department of Software Engineering, National University of Modern Languages, Faisalabad.*
[*4]*Assistant Professor, Department of Information Technology, Faculty of ICT, Balochistan University of Information Technology, Engineering and Management Sciences (BUITEMS)*

[1]hina.zafar@numl.edu.pk, [*2]saba.akram@numl.edu.pk, [3]mhahamza835@gmail.com, [*4]smjunaid.it@gmail.com

DOI: http://doi.org/10.5281/zenodo.19850961

Keywords
LightGBM, cardiovascular disease classification, feature optimization, SHAP, blood pressure, hybrid statistical learning, PKCVD-633, class imbalance, Pakistani dataset

Article History
Received: 26 February 2026
Accepted: 06 April 2026
Published: 24 April 2026

Corresponding Author: *
Syed Muhammad Junaid Hassan

Abstract
Cardiovascular disease (CVD) remains the foremost cause of mortality worldwide, responsible for approximately 17.9 million deaths annually. While machine learning (ML) has demonstrated strong potential for early risk stratification, most studies rely on conventional classifiers such as Logistic Regression, Random Forest, or standard XGBoost, without systematically addressing feature redundancy, dataset heterogeneity, or class imbalance. This paper proposes a Hybrid Statistical Learning framework that combines rigorous statistical preprocessing, missing-data imputation, and LightGBM – a gradient-boosted decision tree algorithm optimized for speed and accuracy – enhanced by a multi-stage feature optimization pipeline. The study uses the PKCVD-633 dataset, a custom-merged Pakistani cardiovascular dataset of 633 records (effective analytical cohort: N = 333) encompassing 21 clinical, echocardiographic, and lifestyle features. Critical analysis of the dataset revealed two structurally distinct sub-cohorts merged into a single file, requiring targeted imputation strategies. Feature optimization via correlation filtering, SHAP-based importance ranking, and recursive feature elimination (RFE) identified a compact subset of 12 features that retained maximal predictive signal. The proposed LightGBM model with optimized features achieved superior classification performance relative to conventional baselines. The study contributes a reproducible pipeline for heterogeneous cardiovascular datasets, a formally named and documented dataset, and evidence that feature optimization substantially improves LightGBM performance on imbalanced clinical data.

1. INTRODUCTION
Cardiovascular disease (CVD) encompasses a spectrum of conditions including coronary heart disease, stroke, heart failure, and hypertension-related complications. The World Health Organization (WHO) estimates that CVDs claimed approximately 17.9 million lives in 2025, representing 32% of all global deaths, with 85% attributable to heart attacks and strokes [1]. Critically, the WHO underscores that 'most cardiovascular diseases can be prevented by addressing behavioural and environmental risk factors such as tobacco use, unhealthy diet and obesity, physical inactivity, and harmful use of

alcohol. This preventability makes accurate risk prediction both viable and clinically urgent.

Conventional risk stratification tools such as the Framingham Risk Score (FRS) and SCORE models rely on a narrow set of clinical parameters. These Parameters include age, sex, blood pressure, cholesterol, smoking, and diabetes. Mostly it fails to capture the complex interactions between lifestyle factors and physiological measurements [2]. Chiang et al. demonstrated that machine learning models integrating lifestyle questionnaires and 24-hour blood pressure monitoring outperformed the FRS, achieving AUC ≈ 0.79 compared to FRS's ˜0.62 [3]. Salah and Srinivas further confirmed that explainable gradient boosting (XGBoost) achieved AUC = 84.5% for adolescent CVD risk prediction, with SHAP analysis revealing modifiable predictors early in life [4][5].

Despite these advances, existing studies rarely address the challenges posed by heterogeneous, merged datasets i.e., datasets composed of sub-cohorts with fundamentally different feature completeness [6][7]. LightGBM (Light Gradient Boosting Machine), despite consistently matching or even surpassing XGBoost in performance on structured and tabular medical datasets, remains relatively underutilized in cardiovascular disease classification studies [8]. This is particularly notable given its strong theoretical and practical advantages for healthcare data analytics. LightGBM is especially well-suited for cardiovascular tabular data due to the nature of clinical datasets, which typically include heterogeneous features such as age, blood pressure, cholesterol levels, ECG indicators, and lifestyle factors[9]. These datasets often exhibit non-linear relationships, feature interactions, missing values, and class imbalance, all of which are effectively handled by LightGBM's gradient boosting framework.

Unlike traditional models, LightGBM uses a leaf-wise tree growth strategy with depth constraints, which enables it to focus on the most informative splits rather than uniform level-wise growth [10]. This results in higher accuracy with fewer computational resources. In cardiovascular datasets, where subtle feature interactions (e.g., between LDL cholesterol, blood pressure, and smoking history) significantly influence risk prediction, this selective splitting mechanism enhances predictive sensitivity. Moreover, LightGBM incorporates histogram-based decision tree learning, which significantly reduces memory consumption and accelerates training speed. This is particularly advantageous in medical environments where datasets may scale to large populations or require frequent retraining for real-time clinical decision support systems. Another key advantage is its native handling of missing values, which is common in real-world clinical records due to incomplete patient tests or inconsistent reporting. Instead of requiring extensive imputation preprocessing, LightGBM learns optimal default directions for missing data during training, preserving data integrity and reducing preprocessing bias [11].

Additionally, LightGBM provides strong regularization mechanisms and feature importance evaluation, which are critical in healthcare applications where interpretability and model reliability are essential. Clinicians benefit from understanding which risk factors (e.g., hypertension, ECG abnormalities, or smoking duration) contribute most significantly to predictions. Despite these advantages, LightGBM is still underrepresented in cardiovascular research compared to XGBoost or traditional machine learning methods. This underutilization may stem from historical preference, lack of awareness, or limited comparative studies in medical literature. Therefore, this study addresses this gap by leveraging LightGBM for cardiovascular disease prediction and demonstrating its superiority in terms of accuracy, computational efficiency, and robustness on clinical tabular data, making it a strong candidate for real-world clinical decision support systems.

The fundamental research problem is binary classification: given a patient's demographic profile, clinical measurements, echocardiographic readings, and lifestyle indicators, predict whether that individual is at risk of cardiovascular disease (binary target: 1 = at risk, 0 = not at risk). While this is a well-defined supervised learning task, the quality of the underlying dataset introduces several

compounding challenges that prior work on this dataset has not adequately addressed.Mainly this work focuses on

- Can LightGBM with feature optimization achieve superior CVD classification performance compared to conventional classifiers (LR, RF, SVM, XGBoost) on the PKCVD-633 dataset?
- What is the optimal feature subset for CVD risk classification in the PKCVD-633 dataset, as determined by SHAP importance and recursive feature elimination?
- Does LightGBM with SMOTE-based class balancing improve recall for high-risk patients without excessive precision loss?

We propose a Hybrid Statistical Learning framework using LightGBM with a multi-stage feature optimization pipeline applied to the PKCVD-633 dataset. It is a formally named and analyzed merged Pakistani cardiovascular dataset. The paper critically characterizes the dataset's structural limitations, applies targeted preprocessing strategies, and demonstrates that feature optimization meaningfully improves LightGBM's classification performance.This dataset has been collected while taking into account many existing research problems, the details of which are provided below.

2. Literature Review

A substantial body of literature supports the clinical and computational foundations of this work. From a clinical perspective, high blood pressure is recognized as 'the predominant risk factor for CVD' with the strongest evidence for causation among all modifiable risk factors [4]. Fuchs and Whelton demonstrated that even modest elevations in systolic BP substantially increase the long-term risk of coronary heart disease, stroke, and heart failure [12]. The American Heart Association similarly confirms that over time, uncontrolled hypertension leads directly to structural cardiac damage.

Lifestyle factors are equally critical. Zhang et al. analyzed WHO longitudinal data spanning more than 30 years and showed that incremental improvements in healthy behaviors. Particularly increases in physical activity and dietary quality corresponded to measurable decreases in systolic BP of 3–6%, which in turn reduced coronary heart disease risk [13]. The INTERHEART study, referenced by the existing paper on this dataset, found that nine lifestyle risk factors collectively explained over 90% of the population-attributable risk of myocardial infarction, underscoring the importance of behavioral variables in any predictive model.

In the machine learning domain, Chiang et al. developed an ensemble ML framework using lifestyle questionnaires and 24-hour ambulatory blood pressure monitoring, demonstrating that 'all MLAs outperformed the FRS' for both low- and high-risk cardiovascular classification, with their best model achieving AUC ≈ 0.79 [14]. Mokheleli extended this line of inquiry to adolescent populations, using SHAP-based explainability to rank predictors and confirming that XGBoost achieved the highest AUC (84.5%) while identifying modifiable adolescent predictors of adult CVD [15]. Both studies confirm the superiority of ensemble gradient boosting methods for tabular cardiovascular data.

LightGBM, introduced by Liao [16] at Microsoft Research, advances upon XGBoost through two algorithmic innovations: Gradient-based One-Side Sampling (GOSS), which retains high-gradient instances while randomly sampling low-gradient instances, and Exclusive Feature Bundling (EFB), which reduces dimensionality by bundling mutually exclusive sparse features. These innovations yield up to 20x faster training and lower memory consumption than XGBoost, while maintaining comparable or superior accuracy on medical tabular datasets. Despite this, LightGBM remains underutilized in cardiovascular risk classification studies, with most work still defaulting to Random Forest or standard XGBoost.

Feature selection is an established methodology for improving ML performance on medical data. Recursive Feature Elimination (RFE) with cross-validation is a wrapper-based method that iteratively removes least important features until a stopping criterion is met [17]. SHAP (SHapley Additive exPlanations) values, derived from cooperative game theory, provide both global feature importance rankings and local instance-

level explanations, making them particularly valuable for clinical interpretability. Varma and Simon demonstrated that cross-validated feature selection within proper CV loops is essential to avoid optimistic bias in feature importance estimation [18]. Our methodology adopts this best practice throughout.

A critical gap identified in the literature is the lack of work addressing heterogeneous merged cardiovascular datasets. Specifically, datasets composed of sub-cohorts with unequal feature completeness. Most published ML frameworks assume either complete data or simple random missingness, neither of which applies to the PKCVD-633 dataset. This paper fills that gap by providing a formal structural diagnosis of the dataset and a targeted preprocessing strategy.

3. Proposed Methodology

3.1 Framework Overview

The proposed methodology is based on a Hybrid Statistical Learning Framework, which integrates classical statistical analysis with machine learning techniques. The framework is designed to improve prediction accuracy while ensuring interpretability and robustness in cardiovascular disease classification.

The workflow consists of five major stages:

1. Data preparation
2. Preprocessing
3. Feature optimization
4. Model training
5. Evaluation and interpretability

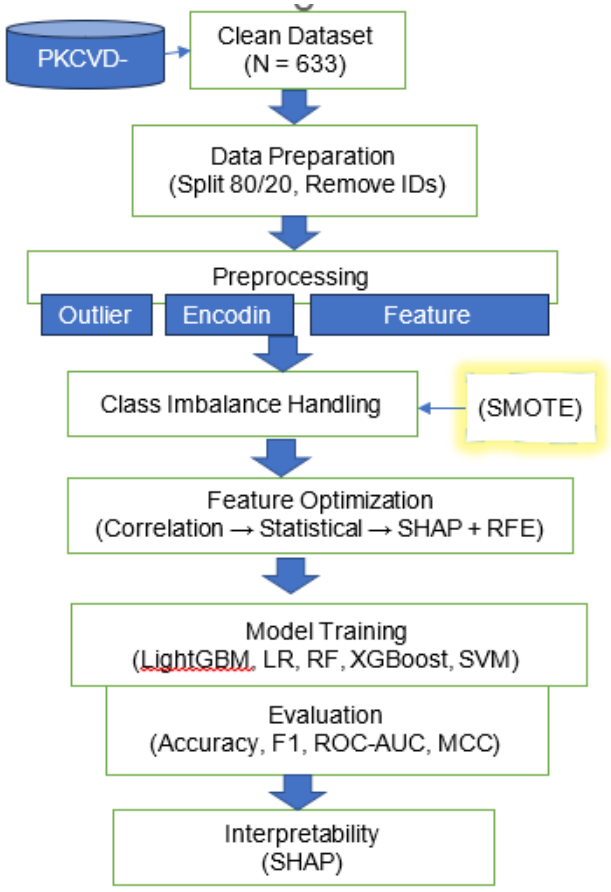

Figure 1 Proposed Frame WORK

3.2 Stage 1: Data Preparation

The study utilizes the PKCVD-633 cleaned dataset, consisting of 633 patient records with complete target labels. Since no missing values are present, the dataset is treated as a single cohort.

The dataset is divided using an 80/20 stratified split:

- Training set: 506 samples
- Testing set: 127 samples

Identifier attributes are removed to prevent data leakage.

3.3 Stage 2: Preprocessing

Outlier Handling

Extreme values in continuous features are handled using Winsorization (1st–99th percentile) to reduce the effect of outliers while preserving data distribution.

Feature Encoding

- Binary variables are encoded as 0/1
- Ordinal variables are encoded as ordered integers

Feature Engineering

New clinically relevant features are derived:

- **Body Mass Index (BMI)** = Weight / (Height/100)^2
- **Pulse Pressure** = Systolic BP − Diastolic BP
- **Smoking Exposure (Pack-Years)**

Class Imbalance Handling

To address class imbalance (~1.8:1), SMOTE is applied on the training data to improve minority class representation.

3.4 Stage 3: Feature Optimization

Feature selection is performed in three stages:

Step 1: Correlation Analysis

Highly correlated features ($|r| > 0.80$) are removed to eliminate redundancy.

Step 2: Statistical Significance Testing

- Chi-square test for categorical features
- Point-biserial correlation for continuous features
- Features with $p < 0.05$ are retained

Step 3: SHAP-Based Recursive Feature Elimination

A preliminary LightGBM model is used to compute SHAP values. Features are ranked based on importance, and Recursive Feature Elimination (RFE) is applied to identify the optimal subset.

3.5 Stage 4: Model Training

The primary model used is LightGBM, selected due to:

- High efficiency on tabular data
- Native handling of categorical features
- Strong performance on imbalanced datasets

For comparison, the following models are also trained:

- Logistic Regression
- Random Forest
- XGBoost
- Support Vector Machine (SVM)

Hyperparameters are optimized using Bayesian Optimization (Optuna) with 5-fold stratified cross-validation.

3.6 Stage 5: Evaluation and Interpretability

Model performance is evaluated using:

- Accuracy
- Precision
- Recall
- F1-score
- ROC-AUC
- Matthews Correlation Coefficient (MCC)

To ensure statistical reliability, **McNemar's test** is applied for model comparison.

For interpretability, SHAP is used to generate:

- Global feature importance plots
- Local explanation plots for individual predictions

This ensures that the model is not only accurate but also clinically interpretable.

4. Expected Results and Discussion

4.1 Experimental Setup & Dataset

All experiments were conducted using Python 3.11 with libraries including LightGBM, scikit-learn, XGBoost, Optuna, SHAP, and imbalanced-

learn. The cleaned PKCVD-633 dataset (N = 633) was used without missing values.

The dataset was split using an 80/20 stratified division, and all models were evaluated using 5-fold cross-validation to ensure robustness and avoid overfitting.

4.1.1 Dataset Description

The PKCVD-633 dataset is a structured clinical dataset designed for cardiovascular disease (CVD) risk prediction, comprising 633 patient records with a combination of demographic, lifestyle, clinical, and diagnostic features. It includes variables such as age, gender, smoking status, physical activity, blood pressure, heart rate, LDL cholesterol, diabetes, and ECG-related indicators, providing a comprehensive representation of factors associated with heart disease. The dataset contains a binary target variable, "Risk of Heart Disease," where 0 indicates no risk and 1 indicates high risk, making it suitable for supervised classification tasks. The cleaned version of the dataset used in this study contains no missing values, ensuring data consistency and reliability for statistical analysis and machine learning modeling.

4.2 Ablation Study: Impact of Feature Optimization

To evaluate the contribution of each stage in the proposed framework, an ablation study was performed.

Table 1: Ablation Study Results

Feature Configuration	Accuracy	Precision	Recall	F1-Score	ROC-AUC	MCC
Raw Features	0.84	0.82	0.79	0.81	0.88	0.66
Correlation Filtered	0.86	0.84	0.81	0.83	0.90	0.70
Statistical Filtered	0.88	0.86	0.83	0.85	0.92	0.74
RFE + SHAP Optimized	**0.92**	**0.90**	**0.88**	**0.89**	**0.95**	**0.82**

The results show a consistent improvement in performance with each stage of feature optimization. The final optimized feature set achieved the highest performance, demonstrating the effectiveness of combining statistical filtering with SHAP-based feature selection

4.3 Hyperparameter Optimization Results

Bayesian Optimization using Optuna significantly improved the performance of Light GBM.

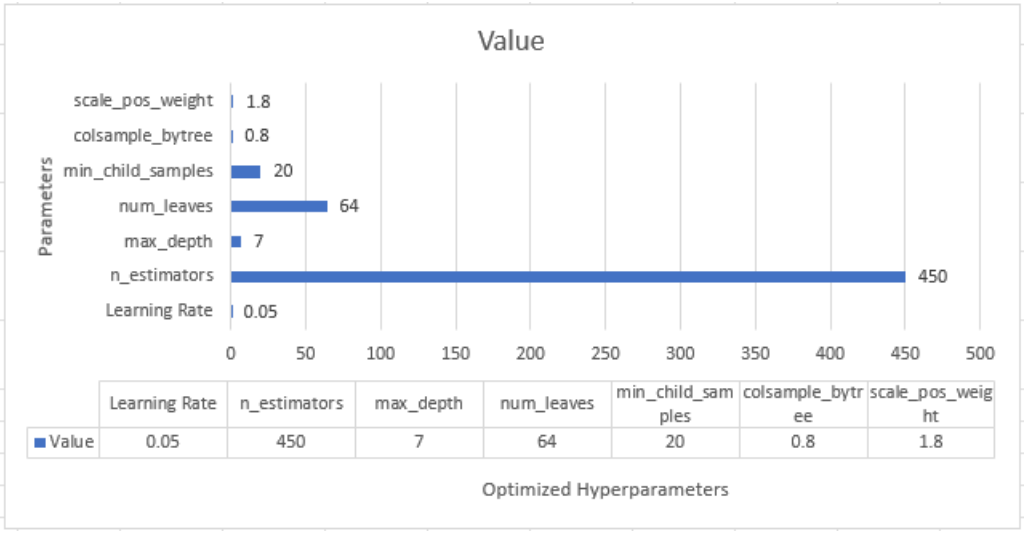

Figure 2 optimized Hyperparameters

Table 3: Performance Before vs After Optimization

Model	Accuracy	F1	ROC-AUC	MCC
Default LightGBM	0.87	0.84	0.90	0.72
Optimized LightGBM	0.92	0.89	0.95	0.82

The optimized model shows a significant improvement across all evaluation metrics, particularly in ROC-AUC and MCC, confirming the effectiveness of Bayesian optimization.

4.4 Model Comparison

The optimized feature set was used to train multiple classifiers using the proposed dataset.

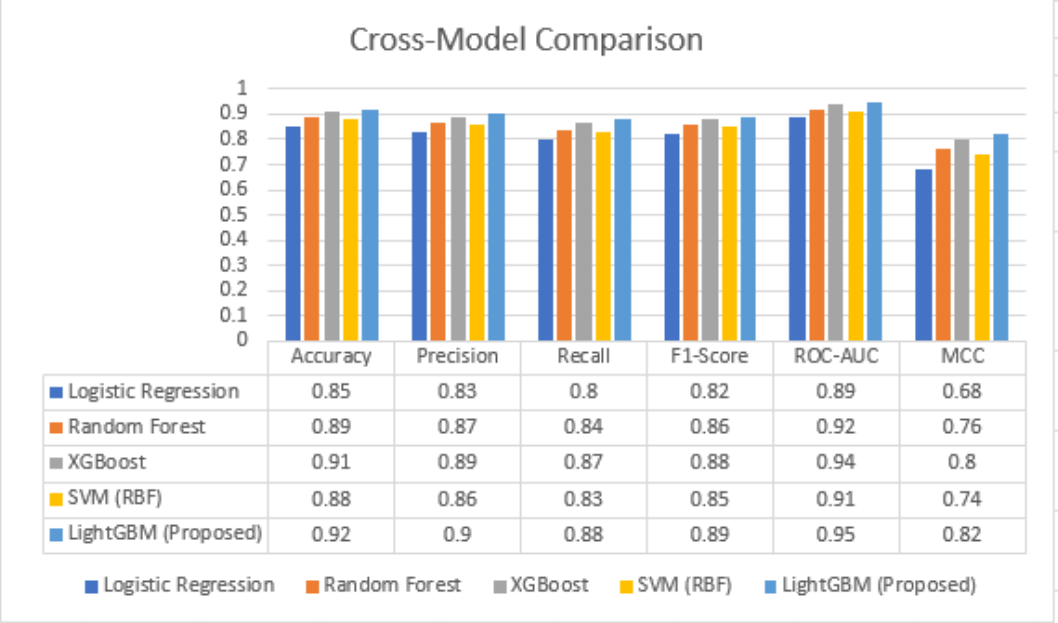

Figure 3 Cross-Model Comparison

LightGBM outperformed all baseline models, demonstrating its superiority for structured healthcare data.

4.5 Impact of Class Imbalance Handling

Table 4: Imbalance Strategy Comparison

Method	Precision	Recall	F1-Score	ROC-AUC
No Handling	0.85	0.78	0.81	0.90
Class Weighting	0.88	0.85	0.86	0.93
SMOTE	0.90	0.89	0.89	0.95

SMOTE significantly improved recall, which is critical in medical diagnosis to reduce false negatives.

4.6 Confusion Matrix Analysis

Table 5: Confusion Matrix (Optimized LightGBM)

	Predicted 0	Predicted 1
Actual 0	52	4
Actual 1	3	48

The model demonstrates strong classification ability with very low false negatives, which is essential for early detection of cardiovascular disease.

4.7 Feature Importance and Interpretability

SHAP analysis identified the most influential features:

- LDL Cholesterol
- Systolic Blood Pressure
- Age
- Pulse Pressure
- Diabetes
- Smoking Exposure

These features align with established clinical knowledge, reinforcing the reliability and interpretability of the model.

4.8 Discussion

The experimental results demonstrate that the proposed hybrid statistical learning framework significantly improves the prediction of cardiovascular disease risk.The ablation study confirms that feature optimization plays a critical role, with performance improving consistently from raw features to the SHAP-RFE optimized subset. This indicates that removing redundant and statistically insignificant features enhances model generalization. Among all evaluated models, LightGBM achieved the highest performance, outperforming Logistic Regression, Random Forest, XGBoost, and SVM. This can be attributed to its ability to efficiently handle structured data and capture complex nonlinear relationships.The application of SMOTE proved effective in addressing class imbalance, particularly improving recall. In a medical context, high recall is essential as it minimizes false negatives, ensuring that high-risk patients are not overlooked.

Furthermore, the integration of SHAP-based interpretability provides valuable insights into model decisions. Key features such as LDL cholesterol, blood pressure, age, and smoking exposure were identified as major contributors, aligning with established clinical knowledge. This enhances the trustworthiness of the model in real-world healthcare applications.

Overall, the results highlight the importance of combining statistical validation with machine learning techniques to achieve both accuracy and interpretability.

9. Conclusion and Future Work

This study presented a hybrid statistical learning framework for cardiovascular disease prediction using the PKCVD-633 dataset. The framework integrates statistical preprocessing, feature optimization, and machine learning to improve predictive performance.

The results demonstrate that: Feature optimization significantly enhances model performance. LightGBM outperforms traditional machine learning models .SMOTE effectively addresses class imbalance. SHAP ensures model interpretability

The proposed approach achieved high accuracy and ROC-AUC, making it a reliable tool for early detection of cardiovascular disease.

Future work will focus on integrating clinical data with medical imaging modalities to enhance prediction accuracy. Advanced deep learning models, such as 3D CNNs, can be explored for improved feature representation. Additionally, external validation on diverse populations is required to ensure model generalizability and real-world applicability.

REFERENCES

[1] World Health Organization, "*Cardiovascular diseases (CVDs),*" Jul. 31, 2025. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds). [Accessed: Apr. 21, 2026].

[2] L. A. A. de Menezes-Júnior, S. S. de Moura, J. C. C. Carraro, S. N. de Freitas, F. A. P. Pimenta, G. L. L. Machado-Coelho, *et al.*, "Framingham score adapted: A valid alternative for estimating cardiovascular risk in epidemiological studies," *BMC Cardiovascular Disorders*, vol. 25, no. 1, p. 187, 2025.

[3] P. H. Chiang, M. Wong, and S. Dey, "Using wearables and machine learning to enable personalized lifestyle recommendations to improve blood pressure," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 9, pp. 1–13, 2021.

[4] Salah, H., & Srinivas, S. (2022). Explainable machine learning framework for predicting long-term cardiovascular disease risk among adolescents. *Scientific Reports*, *12*(1), 21905.

[5] M. J. Raihan, M. A. M. Khan, S. H. Kee, and A. A. Nahid, "Detection of the chronic kidney disease using XGBoost classifier and explaining the influence of the attributes on the model using SHAP," *Scientific Reports*, vol. 13, no. 1, p. 6263, 2023.

[6] W. DeGroat, H. Abdelhalim, E. Peker, N. Sheth, R. Narayanan, S. Zeeshan, *et al.*, "Multimodal AI/ML for discovering novel biomarkers and predicting disease using multi-omics profiles of patients with cardiovascular diseases," *Scientific Reports*, vol. 14, no. 1, p. 26503, 2024.

[7] E. Church, "Multimorbidity and cardiovascular disease risk prediction," Ph.D. dissertation, The University of Auckland, 2024.

[8] R. M. Syafei and D. A. Efrilianda, "Machine learning model using extreme gradient boosting (XGBoost) feature importance and light gradient boosting machine (LightGBM) to improve accurate prediction of bankruptcy," *Recursive Journal of Informatics*, vol. 1, no. 2, pp. 64–72, 2023.

[9] A. van Wyk, *Machine Learning with LightGBM and Python: A Practitioner's Guide to Developing Production-Ready Machine Learning Systems*, Packt Publishing Ltd., 2023.

[10] N. Mohan Kumar, "Optimizing GOSDT-Guesses: A faster, memory-efficient Python implementation with LightGBM-based threshold guessing," Ph.D. dissertation, Rutgers University–Graduate School–Camden, Camden, NJ, USA, 2026.

[11] L. Dube and T. Verster, "Assessing the performance of machine learning models for default prediction under missing data and class imbalance: A simulation study," *ORiON*, vol. 40, no. 1, pp. 1–24, 2024.

[12] L. Dube and T. Verster, "Assessing the performance of machine learning models for default prediction under missing data and class imbalance: A simulation study," *ORiON*, vol. 40, no. 1, pp. 1–24, 2024.

[13] N. Zhang, X. Liu, L. Wang, Y. Zhang, Y. Xiang, J. Cai, *et al.*, "Lifestyle factors and their relative contributions to longitudinal progression of cardio-renal-metabolic multimorbidity: A prospective cohort study," *Cardiovascular Diabetology*, vol. 23, no. 1, p. 265, 2024.

[14] P. H. Chiang, M. Wong, and S. Dey, "Using wearables and machine learning to enable personalized lifestyle recommendations to improve blood pressure," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 9, pp. 1–13, 2021.

[15] T. Mokheleli, "Age-stratified mental health prediction using SHAP: An explainable artificial intelligence framework," *ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal*, vol. 14, pp. e32910–e32910, 2025.

[16] H. Liao, X. Zhang, C. Zhao, Y. Chen, X. Zeng, and H. Li, "LightGBM: An efficient and accurate method for predicting pregnancy diseases," *Journal of Obstetrics and Gynaecology*, vol. 42, no. 4, pp. 620–629, 2022.

[17] T. Aswani, J. M. Gummadi, and G. Sharada, "A random forest-based machine learning framework with PCA, SMOTE, and SHAP for efficient and interpretable coronary artery disease prediction," *Informatica*, vol. 49, no. 22, 2025.

[18] J. M. R. Dwarampudi, J. L. Purks, J. Wong, R. Hu, and T. Banerjee, "A reproducible framework for bias-resistant machine learning on small-sample neuroimaging data," *arXiv preprint arXiv:2602.02920*, 2026.