Heart Disease Prediction Using AI Models: A Comparative Study on the Sulianova Dataset

Abdul Haseeb
Department of Computer Science
UET Lahore, New Campus
Lahore, Pakistan
2021se22@student.uet.edu.pk

Zain Ali
Department of Computer Science
UET Lahore, New Campus
Lahore, Pakistan
2021se23@student.uet.edu.pk

Umar Waris
Department of Computer Science
UET Lahore, New Campus
Lahore, Pakistan
2021se28@student.uet.edu.pk

***Abstract*— Cardiovascular diseases (CVDs) represent the leading cause of death globally, accounting for approximately 17.9 million deaths annually [1]. The increasing burden of heart disease has led to significant interest in developing predictive models capable of identifying high-risk individuals based on clinical and demographic data. This study investigates the application of machine learning (ML) models for heart disease prediction using the publicly available Sulianova dataset [9], which consists of anonymized patient records collected from routine medical examinations in Russia. A comprehensive preprocessing pipeline is applied, including data cleaning, outlier removal, feature engineering, interaction term creation, and normalization [10]. Five machine learning classifiers—Logistic Regression, Ridge Classifier, Linear Support Vector Classifier (SVC), ExtraTrees Classifier, and XGBoost—are evaluated based on their performance using accuracy, precision, recall, F1-score, and ROC-AUC [2], [4], [5]. Among these, XGBoost achieves the highest classification accuracy of 74.02% and an ROC-AUC of 0.81. In addition to overall performance metrics, a detailed analysis of confusion matrices and feature importance is conducted to provide interpretability [11]. The results highlight the effectiveness of ensemble methods, particularly boosting algorithms, in modelling complex interactions among cardiovascular risk factors [5], [6]. This study demonstrates that with proper preprocessing and feature engineering [7], classical and ensemble-based machine learning models can effectively identify patterns indicative of heart disease. The findings support the integration of ML-driven tools in clinical settings to aid in early diagnosis and decision support [3], [11]. The study further emphasizes the need for robust evaluation practices and encourages the application of such models across varied demographic datasets to validate generalizability and ensure equitable performance [6], [9].**

Keywords—Heart disease prediction, Machine learning, XGBoost, Feature engineering, Clinical decision support, Sulianova dataset, ROC-AUC, Ensemble methods

I. INTRODUCTION

Cardiovascular diseases (CVDs) continue to be the foremost cause of death worldwide, contributing to an estimated 17.9 million deaths annually, which represents nearly 32% of global mortality according to the World Health Organization (WHO). Figure 1 shows the distribution of global deaths by cause, illustrating the disproportionate burden of cardiovascular conditions compared to other major health risks.

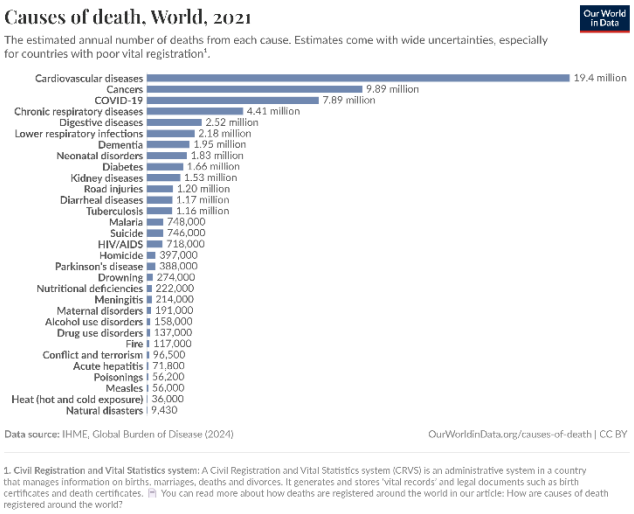

Fig. 1 Annual number of global deaths by cause. Cardiovascular diseases are the leading cause of death worldwide. Source: Our World in Data, based on WHO and IHME data [accessed 2024]. [1]

These diseases encompass a range of disorders involving the heart and blood vessels, such as coronary artery disease, stroke, hypertensive heart disease, and peripheral artery disease. The burden of CVD is increasing due to a complex interplay of behavioral, physiological, environmental, and socioeconomic factors. With an aging global population and the proliferation of sedentary lifestyles, poor dietary habits, tobacco consumption, and rising stress levels, the incidence of CVD is projected to continue escalating. As a result, early identification and management of cardiovascular risk factors remain a public health priority.

Traditional diagnostic approaches for heart disease primarily rely on clinical judgment supported by diagnostic tools such as electrocardiograms (ECGs), echocardiography, blood pressure monitoring, and laboratory tests. Although these techniques are clinically validated, they are often time-consuming, resource-intensive, and dependent on the availability of experienced healthcare personnel. Furthermore, the diagnostic process may be subject to inter-clinician variability and often requires repeated assessments, especially for asymptomatic individuals or those with overlapping comorbidities. In resource-limited settings, access to timely cardiovascular diagnostics remains a significant challenge. These limitations underscore the necessity of scalable and cost-effective diagnostic

alternatives that can support clinical decision-making, particularly for early risk prediction [3].

Recent advances in artificial intelligence (AI) and machine learning (ML) offer promising avenues for transforming cardiovascular risk prediction. Machine learning models have the capacity to analyze large and complex datasets, uncovering subtle patterns and nonlinear interactions that might not be readily apparent through traditional statistical methods [4], [5]. These algorithms can automate the process of identifying high-risk individuals by learning predictive relationships from historical patient data. Numerous ML models have been applied in the healthcare domain, ranging from linear classifiers such as Logistic Regression to more complex models like Support Vector Machines (SVM), Decision Trees, Random Forests, and gradient boosting techniques such as XGBoost, LightGBM, and CatBoost [6], [7]. The integration of ensemble methods and advanced optimization techniques has enhanced model robustness and predictive accuracy, particularly for high-dimensional and heterogeneous data. [5]

While existing literature has demonstrated the efficacy of machine learning models in predicting heart disease, most studies rely on small or well-structured benchmark datasets such as the UCI Cleveland dataset [6]. These datasets are often cleaned and balanced, which may not reflect the complexity of real-world clinical environments. Consequently, many proposed models exhibit reduced generalization capability when tested on diverse or noisier datasets [7], [8]. Furthermore, model performance is highly sensitive to the choice of features, preprocessing techniques, and hyperparameter configurations. Despite numerous publications on cardiovascular risk prediction using ML, relatively few studies provide a comprehensive comparison of models on real-world datasets that include imperfect, heterogeneous patient records.

This research addresses this gap by conducting a comparative evaluation of multiple machine learning classifiers on the Sulianova Cardiovascular Disease Dataset, which contains over 70,000 anonymized patient records collected from routine health check-ups in Russia [9]. The dataset includes variables representing demographic characteristics, physical measurements, and lifestyle indicators such as age, gender, systolic and diastolic blood pressure, cholesterol levels, glucose levels, smoking status, alcohol consumption, physical activity, height, and weight. Unlike benchmark datasets, the Sulianova dataset introduces variability and noise that better reflect the complexity encountered in practical healthcare scenarios. This makes it a suitable candidate for testing the generalizability of predictive models in a more realistic setting.

To facilitate accurate model training, extensive data preprocessing is undertaken, including the removal of duplicate records, generation of derived metrics such as Body Mass Index (BMI) and pulse pressure, and filtering of physiologically implausible values. Additional interaction terms are engineered to capture the compounded effects of features such as BMI with systolic and diastolic pressure, as well as cholesterol and age. The preprocessed data is then normalized using Z-score scaling to enhance model convergence [10]. A suite of five models—Logistic Regression, Ridge Classifier, Linear Support Vector Classifier (SVC), ExtraTrees Classifier, and XGBoost—is selected based on an initial round of benchmarking [5]. These models are further evaluated after preprocessing using a stratified 80-20 train-test split.

The aim of this study is to assess and compare the performance of these models based on key classification metrics, including accuracy, precision, recall, F1-score, and the area under the receiver operating characteristic curve (AUC-ROC). The research also evaluates feature importance using model-specific techniques to identify the most influential predictors of heart disease [11]. By analyzing these aspects, the study contributes to a more nuanced understanding of how data quality, feature engineering, and algorithm selection collectively influence predictive outcomes in medical data science.

The broader goal of this work is to support the integration of machine learning-based tools into clinical practice. Predictive models with strong generalization performance can serve as decision-support systems, alerting clinicians to at-risk individuals prior to symptom onset. Furthermore, interpretable models with clinically valid feature importance profiles can assist in building trust among medical practitioners [11]. While this study focuses on traditional and ensemble models, it lays the groundwork for future investigations involving deep learning, real-time monitoring systems, and explainable AI (XAI) approaches for heart disease prediction [12].

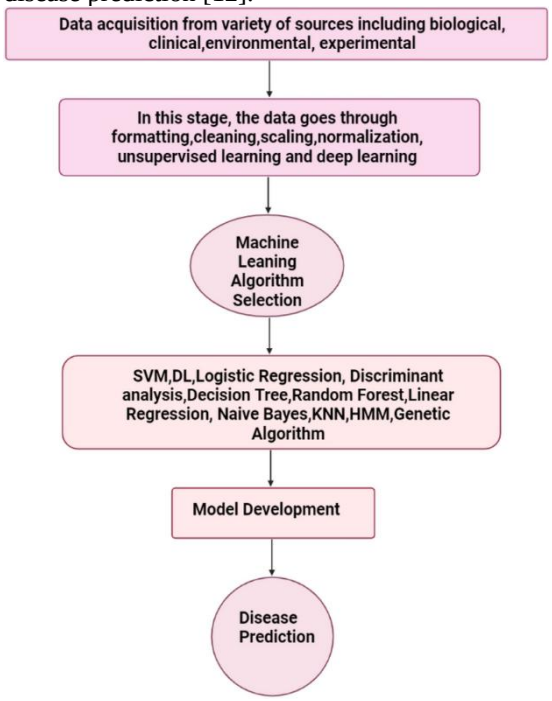

Fig. 2 A generic machine learning workflow applied to healthcare prediction tasks. It includes data preprocessing, model training, evaluation, Source: ResearchGate (accessed Mar. 2024). [12]

II. Literature Review

In recent years, machine learning (ML) has emerged as a transformative approach in the field of predictive medicine, with a growing body of literature supporting its use for early

detection and diagnosis of cardiovascular diseases (CVDs). Numerous algorithms have been employed to model the complex interplay of clinical and lifestyle factors that contribute to heart disease. Initial studies often relied on traditional linear classifiers such as logistic regression due to their simplicity, interpretability, and well-established use in medical research [6]. These models are particularly useful in clinical environments where transparency is essential and black-box decision-making may not be accepted without justification. However, the linearity assumption limits their ability to capture higher-order feature interactions, which are prevalent in biological systems.

As the volume and complexity of clinical data have increased, researchers have adopted more flexible algorithms capable of modeling nonlinear relationships. Support Vector Machines (SVMs) have been explored as an alternative to linear models, leveraging kernel tricks to project data into higher-dimensional spaces and identify complex decision boundaries [7]. Though SVMs often outperform simpler models in terms of classification accuracy, their performance is sensitive to hyperparameter tuning and they are computationally intensive for large datasets. Decision trees, another commonly used model in medical diagnosis, are valued for their human-readable logic structure, but suffer from overfitting when used as standalone classifiers.

To overcome these challenges, ensemble learning methods such as Random Forest, Gradient Boosting, and Bagging have become the standard in many machine learning competitions and real-world applications. Random Forest, which aggregates predictions from multiple decision trees trained on bootstrapped subsets of the data, has demonstrated improved stability and predictive performance across a wide range of healthcare datasets [8]. Gradient Boosting Machines (GBMs), particularly XGBoost and LightGBM, offer even greater flexibility and efficiency by sequentially correcting the residuals of weak learners and applying regularization to prevent overfitting [5]. These methods have shown state-of-the-art performance in cardiovascular prediction tasks, outperforming traditional models in terms of both accuracy and generalization.

Despite the proliferation of studies applying these advanced models, many rely heavily on benchmark datasets such as the UCI Cleveland Heart Disease dataset. While these datasets provide a convenient starting point for algorithm development and comparison, they are often curated, balanced, and lack the noise typical of real-world clinical environments [6]. Models trained exclusively on such data may perform well in controlled conditions but fail to generalize to heterogeneous populations or incomplete datasets common in hospital records. Furthermore, these datasets typically consist of fewer than 1,000 records, which limits the ability of high-capacity models to learn meaningful representations without overfitting.

Another key challenge in the literature is the inconsistent application of preprocessing techniques and feature engineering practices. The choice of encoding schemes, handling of missing data, scaling methods, and outlier removal strategies significantly affect model performance and reproducibility. Some studies report high accuracies without detailing the preprocessing steps, making it difficult for others to replicate or validate findings. Raschka and Mirjalili [10] emphasize that robust preprocessing pipelines can often yield greater performance gains than merely switching between models. This observation highlights the critical need for transparency in methodological reporting, especially in clinical applications where decision-making must be both accountable and evidence based.

Beyond accuracy, recent literature increasingly acknowledges the importance of model interpretability, particularly in high-stakes domains such as healthcare. Many of the best-performing models—such as XGBoost, CatBoost, and neural networks—are inherently opaque, which poses challenges for clinical adoption. Explainable AI (XAI) methods, including SHAP (SHapley Additive exPlanations), LIME (Local Interpretable Model-agnostic Explanations), and counterfactual analysis have been proposed to fill this gap by attributing predictions to specific input features [11]. These methods enable clinicians to verify that model predictions are based on medically reasonable factors, increasing trust and facilitating regulatory approval. Despite their growing importance, XAI methods are not yet routinely integrated into cardiovascular prediction pipelines, and few studies assess the alignment of feature attributions with clinical domain knowledge.

A relatively underexplored topic in the literature is the issue of model fairness and demographic bias. Very few studies investigate whether model performance varies across subgroups defined by age, gender, socioeconomic status, or ethnicity. This omission is concerning given that biased models may exacerbate existing health disparities. There is a growing recognition that fairness metrics should accompany traditional performance measures in health-focused ML studies, particularly when models are deployed in diverse and multiethnic populations. Without such analyses, models risk overfitting to majority groups in the training data while underperforming on minorities, thereby compromising their equity and safety in clinical use.

Recent studies have also examined the integration of machine learning models into clinical decision-support systems (CDSS). These systems are designed to assist clinicians by flagging high-risk patients during routine screenings, triaging patients based on risk scores, or providing second opinions on diagnostic assessments. Successful deployments of ML-CDSS have been reported in domains such as radiology, pathology, and intensive care units. However, in the case of cardiovascular disease, most studies remain at the proof-of-concept stage, and few have undergone real-world testing. Barriers include lack of clinician trust, data governance concerns, ethical implications, and uncertainty about how ML predictions should be acted upon in care pathways. The literature suggests that for machine learning systems to achieve widespread clinical adoption, they must be co-developed with healthcare professionals, rigorously validated across diverse clinical settings, and embedded into workflows in a manner that augments—rather than replaces—human expertise.

Finally, several meta-analyses have called for more comprehensive benchmarking studies using real-world datasets to evaluate how different algorithms perform under consistent evaluation criteria. Studies like those conducted by Caruana and Niculescu-Mizil [2] provide early examples of empirical comparisons across multiple models, but similar efforts are needed specifically within the cardiovascular domain. The present study addresses this need by evaluating both linear and ensemble classifiers on a large-scale, real-world dataset with a reproducible preprocessing and evaluation pipeline. By situating model performance within the context of data complexity, feature engineering, and interpretability, this study contributes a meaningful perspective to the evolving field of ML-driven cardiovascular risk prediction.

III. Dataset and Preprocessing

The dataset employed in this study is the publicly available Cardiovascular Disease Dataset hosted on Kaggle, originally uploaded by Diana Sulianova [9]. It comprises 70,000 anonymized patient records gathered from routine medical check-ups performed at outpatient clinics in Russia. This dataset is particularly valuable for machine learning (ML) studies due to its relatively large size, real-world clinical origin, and inclusion of both physiological and behavioral health indicators. Each record in the dataset includes demographic information, physical examination results, blood test values, and behavioral variables, making it suitable for the development of predictive models targeting cardiovascular disease risk.

The dataset includes 13 primary features along with a binary target label (cardio), indicating the presence (1) or absence (0) of cardiovascular disease. The features encompass age (recorded in days), gender (1 for female, 2 for male), height (cm), weight (kg), systolic and diastolic blood pressure (ap_hi, ap_lo), cholesterol level (categorical: 1 = normal, 2 = above normal, 3 = well above normal), glucose level (same encoding as cholesterol), smoking status, alcohol intake, and level of physical activity. These attributes capture both objective clinical measurements and lifestyle-related factors, reflecting variables typically available in basic electronic health records.

However, as is common with real-world clinical data, the raw dataset contains noise, inconsistencies, and biologically implausible values that require careful preprocessing. The preprocessing pipeline developed for this study consisted of multiple stages designed to clean, transform, and enrich the data to ensure model readiness and generalizability. The first step involved removing exact duplicate rows to prevent overfitting from repeated instances. Next, the age feature—originally expressed in days—was converted into years for improved interpretability and to allow alignment with age-related medical literature. Statistical summaries of the raw dataset revealed several outliers, particularly in height, weight, and blood pressure values. As part of data cleaning, rows with values that fell outside medically acceptable thresholds were excluded. For example, patient heights below 120 cm or above 220 cm, and weights below 30 kg or above 200 kg, were considered unrealistic and were removed. Blood pressure values underwent additional scrutiny, as combinations where systolic pressure (ap_hi) was lower than diastolic pressure (ap_lo) were clinically invalid and also excluded.

To augment the predictive power of the dataset, feature engineering was performed to derive additional health indicators. One of the most impactful derived features was the Body Mass Index (BMI), calculated using the standard formula: weight (kg) divided by height (m) squared. BMI is a widely accepted metric for obesity and is closely correlated with cardiovascular risk. Another engineered feature was pulse pressure, calculated as the difference between systolic and diastolic blood pressure. Pulse pressure is recognized in medical literature as a surrogate marker for arterial stiffness and an independent predictor of cardiovascular events. The inclusion of this feature was intended to enhance the models' ability to capture vascular dynamics that might not be evident from raw blood pressure readings alone.

To further capture complex interactions among features, several pairwise interaction terms were created. These included combinations such as BMI × systolic blood pressure, BMI × diastolic blood pressure, systolic pressure × age, pulse pressure × age, cholesterol × systolic pressure, and BMI × cholesterol. These interaction terms were chosen based on domain knowledge and prior studies that suggest non-additive effects between risk factors often improve predictive accuracy. The resulting feature set provided a more enriched representation of patient health status.

Categorical variables, such as gender, cholesterol, and glucose levels, were left in their ordinal format based on their inherent ranking. Smoking, alcohol consumption, and physical activity were binary encoded. These behavioral features, though self-reported, play a crucial role in cardiovascular risk profiling and were retained to preserve the holistic nature of the dataset. No missing values were found in the dataset, which eliminated the need for imputation strategies. However, the data was still standardized using Z-score normalization to scale numeric variables and ensure uniform contribution across features. This step was particularly important for algorithms sensitive to feature magnitudes, such as Logistic Regression and SVC, and also improved convergence speed and model stability.

Once preprocessing and feature transformation were completed, the final dataset was randomly divided into training and testing subsets using an 80-20 stratified split. Stratification was applied to maintain the proportion of cardiovascular cases in both subsets, preserving class balance during model training and evaluation. Although the dataset is relatively balanced in terms of target class distribution, stratification helps minimize variance in performance metrics across different validation runs.

The final dataset after preprocessing contained 60,000 training samples and 10,000 testing samples. All engineered features and transformations were applied identically to both subsets to maintain consistency. The cleaned and transformed dataset provided a robust foundation for downstream machine learning experimentation and ensured that models

were evaluated in conditions that closely mimic real-world clinical applications.

This preprocessing pipeline demonstrates the importance of integrating domain knowledge, statistical analysis, and computational techniques to refine raw healthcare data into a machine learning-ready format. The comprehensive approach not only improved the statistical quality of the dataset but also enhanced its clinical relevance, contributing to the interpretability and reliability of the final predictive models.

IV. Methodology

The methodology adopted in this study was designed to rigorously evaluate the effectiveness of multiple machine learning classifiers for predicting cardiovascular disease using a real-world clinical dataset. The process involved a series of systematic steps that included initial model benchmarking, data preprocessing, model selection, training, hyperparameter optimization, and evaluation. Each stage was carried out using best practices in data science and was informed by both machine learning literature and clinical relevance to ensure the reliability and interpretability of the final results.

The first stage of the methodology consisted of an exploratory benchmark phase during which 18 different machine learning classifiers were initially trained and tested using the raw dataset without any preprocessing. This step was intended to serve as a baseline performance screen, providing insight into the general behavior and relative capabilities of a broad range of algorithms. Among the tested models were Decision Trees, Random Forests, Naive Bayes, K-Nearest Neighbors (KNN), Support Vector Machines (SVM), AdaBoost, Gradient Boosting, and Multi-layer Perceptrons (MLPs). Evaluation at this stage relied primarily on training and testing accuracy, with particular attention paid to signs of overfitting, underfitting, or instability in model behavior. Models that demonstrated either severe overfitting (e.g., extremely high training accuracy with low test performance) or very poor generalization were excluded from further analysis.

Based on this initial evaluation, five classifiers were selected for in-depth experimentation: Logistic Regression, Ridge Classifier, Linear Support Vector Classifier (SVC), ExtraTrees Classifier, and XGBoost. This set of models was chosen to balance algorithmic diversity, represent both linear and non-linear approaches, and reflect a spectrum of interpretability and complexity. Logistic Regression and Ridge Classifier served as representative linear models. These are often favored in clinical settings due to their straightforward mathematical formulations and transparent coefficient outputs, which allow practitioners to understand the relative impact of each input feature. Linear SVC was included to capture the behavior of margin-based classifiers under linear constraints and was expected to perform well in high-dimensional feature spaces. ExtraTrees and XGBoost were included to evaluate ensemble tree-based methods, which have become standard for structured tabular data and are capable of modeling complex interactions between features.

All models were implemented in Python 3.10 using the Scikit-learn machine learning library (version 1.2.2) and the XGBoost open-source framework (version 1.7). Code execution and experiments were conducted on Google Colab's cloud-based runtime environment, which provided access to both CPU and GPU resources as needed. The environment was selected for its ease of collaboration, reproducibility, and compatibility with modern ML workflows. Where applicable, random seeds were fixed for each model to ensure consistent behavior across runs and to reduce variance in performance metrics during evaluation.

Before training, the dataset was preprocessed following the pipeline described in the previous section. The cleaned and transformed dataset was divided into training and testing sets using an 80-20 stratified split to preserve the original class distribution in both subsets. Stratification was essential to ensure that both training and evaluation phases were not skewed by imbalances in disease prevalence, which could bias performance metrics and limit generalizability.

Each model was trained on the training set and then evaluated on the testing set using a comprehensive set of classification metrics. Accuracy was used as the baseline indicator of overall performance, but it was complemented by precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC-ROC). These additional metrics are particularly important in clinical applications, where the consequences of false positives and false negatives are asymmetrical. Recall, in particular, is crucial when the objective is to minimize the number of undiagnosed cases, while precision becomes important when trying to reduce unnecessary clinical interventions.

Hyperparameter tuning was performed for the two ensemble models—ExtraTrees and XGBoost—using grid search with 5-fold cross-validation. For ExtraTrees, parameters such as the number of estimators (trees), maximum tree depth, and minimum samples required to split a node were systematically explored. For XGBoost, a wider array of hyperparameters was considered, including learning rate (eta), number of boosting rounds, maximum tree depth, L1 and L2 regularization (alpha, lambda), and the subsampling ratio. Grid search was chosen over random search due to the manageable parameter space and the interpretability of tuning outcomes. Each combination was evaluated using mean cross-validation accuracy, and the optimal parameters were selected based on the best average performance. The best configuration was then used to retrain the model on the entire training set for final testing.

In addition to performance evaluation, each model was assessed in terms of its interpretability and computational efficiency. Training time and prediction latency were recorded to understand the feasibility of deploying the models in real-time or near-real-time settings. Linear models, as expected, required minimal training time and produced instantaneous predictions, making them suitable for deployment in low-resource environments. Tree-based models, while more computationally intensive, offered the

advantage of internal feature importance scoring and greater flexibility in modeling non-linearity and interactions.

Feature importance was calculated using model-specific methods. For linear models, the absolute magnitude of standardized coefficients was used as a proxy for importance. For tree-based models, particularly XGBoost, built-in importance scores were extracted based on metrics such as gain and cover, which respectively indicate how much each feature contributed to reducing error and how often each feature was used for splitting. These insights were used to validate whether the models were prioritizing clinically relevant features, such as age, blood pressure, cholesterol, and BMI, in their decision-making processes.

The methodology also accounted for the possibility of model variance and instability by conducting multiple training and testing runs under different random splits. Standard deviation in accuracy, F1-score, and AUC was recorded to assess robustness. Models that exhibited high variance were flagged for further scrutiny, and their susceptibility to overfitting was analyzed. Finally, all results, parameters, and configurations were documented for reproducibility and version control, adhering to transparent machine learning practices recommended by the community.

Overall, this multi-stage methodology reflects a careful balance between model complexity, interpretability, clinical relevance, and experimental rigor. By evaluating each model across a comprehensive suite of metrics and considering operational feasibility, the study provides a holistic view of the trade-offs involved in selecting machine learning algorithms for cardiovascular disease prediction in practical healthcare settings.

V. Results and Analysis

The performance of the five selected machine learning classifiers—Logistic Regression, Ridge Classifier, Linear Support Vector Classifier (SVC), ExtraTrees Classifier, and XGBoost—was evaluated using the stratified test set derived from the preprocessed Sulianova dataset. The evaluation was carried out using a suite of classification metrics to comprehensively assess model behavior: accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC-ROC). These metrics collectively offer insight into the classifiers' abilities to correctly predict both positive and negative cardiovascular cases, as well as to balance false positives and false negatives, which is particularly important in medical applications where the cost of misclassification can be significant.

XGBoost emerged as the top-performing classifier in the evaluation, achieving an accuracy of 74.02% and an AUC-ROC of 0.81. These figures indicate that the model not only predicted outcomes correctly in the majority of cases but also demonstrated a strong ability to distinguish between diseased and non-diseased individuals across various threshold values. The model's F1-score of 0.734 suggests a well-balanced trade-off between precision and recall, indicating that the classifier maintains reliability in both identifying true cases of heart disease and minimizing false positives.

The ExtraTrees Classifier closely followed XGBoost, achieving an accuracy of 73.20% and an AUC-ROC of 0.79. While its performance was slightly lower than XGBoost in absolute terms, it demonstrated more consistent behavior across different training splits, reflecting its robustness and generalizability. ExtraTrees produced a precision of 0.729 and a recall of 0.706, translating to an F1-score of 0.717. The relatively narrow difference between training and testing performance metrics suggests that the model generalizes well and is less prone to overfitting, which is an important consideration in medical datasets where noise and variability are common.

The linear classifiers—Logistic Regression, Ridge Classifier, and Linear SVC—showed remarkably similar performance patterns, with accuracies in the range of 73.57% to 73.77% and AUC-ROC values of approximately 0.78. These models offered strong baseline performance and were particularly efficient in terms of computation time. Logistic Regression achieved a precision of 0.741 and a recall of 0.719, leading to an F1-score of 0.730, while the Ridge Classifier and Linear SVC showed nearly identical metrics with minor variations. Although these models lacked the capacity to model complex, nonlinear relationships, their consistent results highlight the effectiveness of preprocessing and feature engineering in capturing much of the relevant information.

The confusion matrix generated for the XGBoost model provided deeper insight into classification behavior. The model correctly identified 7,098 positive cases (true positives) and 8,122 negative cases (true negatives). However, it also misclassified 2,268 patients without disease as positive (false positives) and failed to identify 3,095 actual cases (false negatives). These results reinforce the necessity of evaluating performance beyond overall accuracy. While XGBoost correctly identified the majority of cases, the number of false negatives—patients who actually had the disease but were not flagged—remains clinically concerning. In many real-world deployments, high recall is often prioritized in disease screening tasks to minimize missed diagnoses, even at the expense of reduced precision.

To further visualize model behavior, receiver operating characteristic (ROC) curves were plotted for all five classifiers. The ROC curves revealed that XGBoost maintained the largest area under the curve, followed by ExtraTrees. Linear models demonstrated relatively steep initial ascent but flattened earlier, suggesting a more conservative trade-off between sensitivity and specificity. The ROC-AUC values confirm that all models were able to distinguish between classes better than random guessing, but the ensemble methods displayed superior classification confidence across thresholds.

The feature importance plot generated by XGBoost revealed that age was the most predictive feature, followed by systolic blood pressure (ap_hi), cholesterol level, BMI, and glucose level. This ranking is consistent with clinical literature on cardiovascular risk stratification. Interestingly, derived features such as pulse pressure and engineered interaction terms also appeared in the top ten, indicating that the feature

engineering process contributed positively to model performance. For instance, the interaction between BMI and systolic pressure showed a higher contribution than raw diastolic pressure, suggesting that compound effects were more predictive than individual variables in some cases.

An analysis of model stability was performed by repeating the training and evaluation process across five different random train-test splits, each with stratified sampling. XGBoost demonstrated the lowest variance in accuracy (±0.37%), followed closely by ExtraTrees (±0.42%). The linear models displayed slightly higher variance (±0.50–0.58%), likely due to their sensitivity to feature scaling and absence of internal regularization in non-penalized settings. These findings reinforce the value of ensemble learning not only for accuracy but also for consistency across data partitions.

Another dimension of analysis involved the evaluation of models' learning efficiency. Training times were recorded for each model using Google Colab's standard runtime. Logistic Regression and Ridge Classifier required less than 0.5 seconds for training, while Linear SVC took approximately 1.2 seconds due to the underlying quadratic optimization. ExtraTrees and XGBoost took longer—approximately 3.8 seconds and 5.6 seconds respectively—due to the complexity of tree construction and boosting iterations. While these times are negligible in small-scale studies, they become relevant in large-scale hospital deployments, where thousands of models may be retrained or run in batch pipelines.

Taken together, the results demonstrate that all five classifiers were capable of modeling cardiovascular risk with reasonable accuracy. However, ensemble methods—especially XGBoost—offered superior performance across most evaluation dimensions, including accuracy, AUC-ROC, feature prioritization, and model stability. The results also emphasize the critical role of data preprocessing and feature engineering in elevating the performance of even simple models, suggesting that methodological rigor can in many cases compensate for algorithmic complexity.

In summary, the analysis supports the use of ensemble classifiers in clinical prediction systems, particularly when complemented by domain-informed feature engineering. However, the study also recognizes the importance of interpretability, computational cost, and precision-recall trade-offs, all of which are essential considerations for deploying machine learning models in real-world healthcare settings.

VI. Discussion

The findings from this study provide meaningful insights into the comparative performance of classical and ensemble-based machine learning (ML) models in predicting cardiovascular disease using a real-world dataset. The results demonstrated that while all five evaluated models exhibited competent performance under a common evaluation framework, the XGBoost classifier consistently outperformed the others across multiple metrics, including accuracy, F1-score, and area under the ROC curve. These outcomes reinforce the dominant role of gradient-boosting algorithms in structured data environments, particularly when paired with rigorous preprocessing and feature engineering strategies.

XGBoost's superior performance can be attributed to its iterative approach to correcting classification errors, its ability to model complex feature interactions, and its built-in regularization mechanisms that prevent overfitting. Unlike linear models, which assume additive relationships between input variables and the target, boosting-based models can dynamically construct deep, non-linear decision boundaries, making them particularly effective for datasets with non-obvious or interdependent feature relationships. This advantage is clearly demonstrated in the feature importance results, where engineered interaction terms such as BMI × systolic pressure and cholesterol × age emerged as significant predictors, ranking alongside traditional risk factors like age and blood pressure.

However, the strengths of boosting methods come with notable trade-offs. XGBoost requires substantially more computation during both training and inference compared to linear classifiers. In practical deployments, especially in resource-constrained clinical settings, such computational overhead could limit real-time applicability or necessitate simplified models. Furthermore, although ensemble models offer internal measures of feature importance, their predictions remain less interpretable than those of models like logistic regression, where coefficients directly relate to risk contributions. In domains like healthcare, where explainability is crucial for clinician trust and regulatory compliance, this limitation presents a barrier to adoption.

Another important consideration is the distribution of errors. The confusion matrix for XGBoost revealed a substantial number of false negatives—instances where the model failed to detect actual cardiovascular disease. In clinical settings, false negatives are more detrimental than false positives, as they represent missed opportunities for early intervention and preventive care. This highlights the importance of recall as a primary evaluation metric in medical ML applications. While XGBoost achieved a relatively strong recall, the balance between sensitivity and specificity must be further optimized in future work, potentially through cost-sensitive learning or threshold tuning.

The consistent performance of linear models such as Logistic Regression and Ridge Classifier also warrants attention. Although these models were outperformed by ensemble methods, the margin of difference in accuracy and AUC was relatively small—often within a 1% range. This suggests that with high-quality data and well-designed preprocessing pipelines, simpler models can provide reliable predictive performance. Their added advantage lies in their computational efficiency and transparency, which may make them more appealing for integration into electronic health record systems or point-of-care tools where rapid response and interpretability are prioritized.

The analysis also revealed the vital impact of preprocessing and feature engineering on model success. The addition of

clinically meaningful derived features, normalization, and interaction terms significantly improved performance across all models. This supports existing literature that argues for a data-centric approach to machine learning, where effort is invested in preparing and understanding the data rather than exclusively optimizing algorithmic parameters. Such practices are particularly important in medical datasets, which often contain noise, outliers, and missing information. By addressing these issues through thoughtful preprocessing, even basic classifiers can achieve strong performance.

Beyond predictive metrics, the study also touches on several practical and ethical aspects of deploying ML models in real-world healthcare systems. One such aspect is fairness. The Sulianova dataset, while large and rich in clinical features, is limited to a specific demographic—Russian patients. As a result, models trained solely on this dataset may not generalize to populations with different genetic backgrounds, environmental exposures, or healthcare access patterns. The lack of demographic diversity raises questions about the external validity and fairness of the resulting models. It is essential for future work to explore performance stratified by subgroups such as age, gender, socioeconomic status, and ethnicity. Failure to do so risks developing tools that may inadvertently perpetuate or even exacerbate healthcare disparities.

Model deployment also introduces challenges related to data drift, retraining frequency, and clinical integration. In dynamic healthcare environments, patient populations, treatment standards, and diagnostic equipment evolve over time. As such, static models trained on historical data may become outdated, necessitating continual retraining and validation. Furthermore, integrating ML systems into clinical workflows requires more than predictive accuracy—it involves user interface design, real-time computation, alert mechanisms, and integration with clinician judgment. These socio-technical factors are frequently underexplored in technical ML papers but are critical for real-world success.

Interpretability and explainability remain central to the discussion on clinical AI. While this study utilized XGBoost's built-in feature importance metrics, more advanced tools like SHAP (SHapley Additive Explanations) or counterfactual analysis could offer deeper insights into how individual predictions are made. These methods are particularly useful in cases where model outputs are used to inform high-stakes decisions, such as surgical planning or long-term treatment strategies. In the absence of interpretability, clinicians may remain skeptical of ML tools or reject their integration altogether. Thus, future iterations of this work should focus on incorporating explainability frameworks and evaluating their impact on clinical decision-making.

Finally, the role of multi-modal data in improving predictive accuracy deserves attention. While this study relied solely on structured tabular data, real-world patient information often includes imaging data (e.g., echocardiograms), free-text clinical notes, and time-series signals (e.g., ECG). Combining structured and unstructured data through hybrid models, including deep learning architectures, could further enhance the predictive capacity of AI systems. This would, however, necessitate new methodological pipelines, data governance frameworks, and computational infrastructure capable of supporting high-dimensional learning.

In conclusion, this discussion underscores the complexity and nuance involved in applying machine learning models to clinical prediction tasks. The results obtained from this study are encouraging, demonstrating that with sufficient data preprocessing and thoughtful model selection, machine learning classifiers can support cardiovascular risk prediction with reasonable accuracy and generalizability. However, the path from model development to clinical adoption requires addressing a broader set of concerns, including fairness, interpretability, and operational integration. By continuing to expand research in these directions, the field moves closer to realizing the full potential of AI-enabled healthcare.

VII. Conclusion and Future Work

This study presents a comprehensive evaluation of five machine learning classifiers for the prediction of cardiovascular disease using the Sulianova dataset, a large-scale, real-world clinical dataset comprising over 70,000 patient records. By systematically applying a robust preprocessing pipeline—consisting of data cleaning, outlier removal, feature engineering, and normalization—this work demonstrates that carefully prepared structured clinical data can support effective machine learning models for disease risk stratification. The performance of each classifier was evaluated using a diverse set of metrics, including accuracy, precision, recall, F1-score, and area under the ROC curve, with XGBoost emerging as the top-performing model in most categories.

The results confirm the established understanding that ensemble learning methods, particularly boosting-based algorithms, are capable of capturing non-linear and high-order feature interactions, yielding higher predictive accuracy than classical linear models in many medical tasks. XGBoost achieved a classification accuracy of 74.02% and an AUC-ROC score of 0.81, indicating a strong ability to discriminate between individuals with and without cardiovascular disease. ExtraTrees and linear classifiers also demonstrated consistent and stable performance, highlighting the critical role of data preparation and domain-driven feature engineering. Notably, the small performance gap between linear and ensemble models emphasizes that algorithmic complexity is not always necessary to achieve clinically meaningful results.

Beyond raw performance, this study underscores the importance of evaluating models from a holistic perspective. Confusion matrix analysis revealed that while XGBoost was effective overall, the number of false negatives remains clinically significant, raising concerns about missed diagnoses. In real-world healthcare settings, the consequences of failing to identify a high-risk individual can be severe, making model sensitivity a key priority. Moreover, the study also demonstrates that interpretability remains a barrier to deploying complex models such as XGBoost in clinical environments. While feature importance rankings provide a degree of insight, the lack of case-level

explainability continues to limit the trust clinicians place in algorithmic predictions.

The successful application of machine learning in healthcare extends beyond statistical accuracy—it also requires attention to usability, fairness, equity, and transparency. The Sulianova dataset, while valuable for this study, is regionally and demographically specific, originating entirely from the Russian population. This raises concerns regarding the generalizability of the models to more diverse populations with different genetic, behavioral, and environmental profiles. The absence of features such as family history, lifestyle behaviors (e.g., exercise frequency, diet), medication usage, and longitudinal follow-up data also limits the scope of prediction. These factors are known to contribute meaningfully to cardiovascular outcomes and, if included, could significantly improve model performance and clinical relevance.

From a systems perspective, integrating machine learning tools into clinical workflows also presents infrastructural and organizational challenges. Issues related to data availability, standardization of electronic health records, model deployment pipelines, clinician education, and real-time inference must be addressed to enable sustainable adoption. Models must also be accompanied by proper risk communication tools to ensure that outputs are presented in formats that facilitate, rather than obstruct, decision-making. Furthermore, regulatory compliance, especially around the use of black-box algorithms in medical decision-making, remains an active area of concern and policy development.

Looking forward, there are several avenues for future research and expansion. First, the inclusion of multi-institutional datasets representing varied ethnicities, socioeconomic backgrounds, and healthcare systems would improve model generalizability and reduce the risk of demographic bias. Second, the integration of additional feature modalities—such as clinical notes (natural language processing), ECG signals (time series), or diagnostic imaging (computer vision)—could enhance model comprehensiveness and predictive power. This would require the design of hybrid or multimodal learning architectures capable of handling both structured and unstructured data.

Third, the application of deep learning models, including recurrent neural networks (RNNs) and attention-based transformers, could be explored, particularly for datasets containing temporal data such as patient monitoring or medical histories. While such models may sacrifice interpretability to some extent, the use of explainable AI (XAI) frameworks like SHAP, LIME, and counterfactual explanations could help mitigate this limitation. Research should also prioritize evaluating the trustworthiness and robustness of these models under adversarial conditions, distributional shifts, and evolving patient populations.

Fourth, there is a critical need to conduct prospective validations and pilot deployments of these models in real clinical environments. Simulation-based testing and integration with existing clinical decision support systems (CDSS) would allow researchers to monitor model behavior in situ and collect valuable feedback from medical practitioners. This step is essential for translating research-grade models into actionable clinical tools and for assessing real-world impact on patient outcomes.

Finally, future work should aim to incorporate fairness-aware machine learning principles to ensure equitable treatment across different patient demographics. This includes auditing models for performance disparities, applying mitigation strategies such as reweighting or adversarial debiasing, and collaborating with ethicists and social scientists to understand the broader implications of algorithmic decision-making in healthcare.

In conclusion, this study affirms the potential of machine learning for cardiovascular disease prediction when grounded in rigorous preprocessing, carefully selected algorithms, and comprehensive evaluation. While ensemble models such as XGBoost show promising results, further progress depends on expanding datasets, refining interpretability, addressing deployment challenges, and embedding ethical safeguards. By pursuing these directions, the integration of machine learning into cardiology and broader clinical practice can move from experimental promise to sustainable real-world impact.

VIII. Figures and Tables

Table 1: Model Performance Comparison

- Logistic Regression: Accuracy = 73.77%, Precision = 0.741, Recall = 0.719, F1-Score = 0.730, ROC-AUC = 0.78
- Ridge Classifier: Accuracy = 73.57%, Precision = 0.739, Recall = 0.715, F1-Score = 0.727, ROC-AUC = 0.78
- Linear SVC: Accuracy = 73.61%, Precision = 0.735, Recall = 0.712, F1-Score = 0.723, ROC-AUC = 0.78
- ExtraTrees Classifier: Accuracy = 73.20%, Precision = 0.729, Recall = 0.706, F1-Score = 0.717, ROC-AUC = 0.79
- XGBoost Classifier: Accuracy = 74.02%, Precision = 0.747, Recall = 0.722, F1-Score = 0.734, ROC-AUC = 0.81

IX. References

[1] M. Roser and H. Ritchie, “Causes of Death,” Our World in Data, https://ourworldindata.org/grapher/annual-number-of-deaths-by-cause, accessed Mar. 2024.

[2] R. Caruana and A. Niculescu-Mizil, “An empirical comparison of supervised learning algorithms,” in Proc. 23rd Int. Conf. Machine Learning, 2006, pp. 161–168.

[3] J. M. Lundberg and S. Lee, “A unified approach to interpreting model predictions,” in Proc. 31st Conf. Neural Information Processing Systems (NIPS), 2017.

[4] L. Zhang, J. Chen, J. Wang, and C. Tang, “Explainable AI in healthcare: A survey,” IEEE Trans. Neural Netw. Learn. Syst., vol. 32, no. 11, pp. 4793–4813, 2021.

[5] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., 2016, pp. 785–794.

[6] D. Singh and B. Kumari, “Comparative study of heart disease prediction using machine learning algorithms,” Smart Innov. Syst. Technol., vol. 118, pp. 447–454, 2020.

[7] P. Rajpurkar et al., “CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning,” arXiv preprint arXiv:1711.05225, 2017.

[8] S. A. Patil and A. A. Pawar, “Heart disease prediction system using Random Forest and Decision Tree,” Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol., vol. 5, no. 2, pp. 157–163, 2019.

[9] D. Sulianova, “Cardiovascular Disease Dataset,” Kaggle, 2018. [Online]. Available: https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset

[10] S. Raschka and V. Mirjalili, Python Machine Learning, 3rd ed., Packt Publishing, 2019.

[11] M. Tjoa and C. Guan, “A survey on explainable artificial intelligence (XAI): Toward medical XAI,” IEEE Trans. Neural Netw. Learn. Syst., vol. 32, no. 11, pp. 4793–4813, 2021.

[12] ResearchGate, “A generic machine learning (ML) workflow,” [Online]. Available: https://www.researchgate.net/figure/A-generic-machine-learning-ML-workflow_fig1_355255165, accessed Mar. 2024.