Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 4, Issue 2, 2026

Evaluation of Machine Learning Models for Heart Disease Prediction under Progressive Feature Reduction: An Exploratory Study

Ayesha Khattak[1], Syed Asim Jalal [2]*, Saif ur Rehman [3], Shaukat Ali [4], Muhammad Haseeb[5]

[1] School of Computer Science and IT, Institute of Management Science, Peshawar, Pakistan
[2, 3, 4, 5] Department of Computer Science, University of Peshawar

[2]sa.jalal@uop.edu.pk

DOI: https://doi.org/

Keywords

Article History
Received on 20 Jan 2026
Accepted on 15 Feb 2026
Published on 28 Feb 2026

Copyright @Author

Corresponding Author: *

Syed Asim Jalal
Department of Computer Science, University of Peshawar

Email: sa.jalal@uop.edu.pk

Abstract

Cardiovascular disease (CVD) remains the main cause of death worldwide, which makes timely and accurate risk identification very important for reducing related risks. In constrained screening situations, there is need for predictive models that can work with fewer inputs as a decision support system. Reduced features count, however, does not necessarily means lower input features acquisition burden. This study presents a comparative evaluation of eleven commonly used machine learning classifiers on progressively reduced subsets of the commonly used UCI Cleveland Heart Disease dataset. From the full 13-feature dataset, we extracted three subsets containing 10, 8, and 6 features. This was done using Chi-Square-based feature selection technique. The aim was to examine how performance of classifier changes under progressive feature reduction and to identify models that remain robust for situations when fewer inputs are available. This paper should be read as an exploratory baseline study. It does not propose a new prediction model. Instead, it compares commonly used models under the same feature-reduction setting to understand their behaviour. Results of the study suggest that Logistic Regression demonstrated consistent performance across all feature sets, including the most reduced 6-feature dataset with little loss in classification metrics. Random Forest and Naïve Bayes also gave competitive performance. However, the reduced subsets were selected statistically and retained some clinically burdensome variables. Therefore, fewer features should not be treated as direct evidence of lower cost or reduced screening time. Cost/time implications are discussed qualitatively and were not measured. The study contributes comparative baseline evidence on classifier robustness under feature reduction and highlights the need for feasibility-aware feature selection in future work.

Introduction

Cardiovascular disease (CVD) remains one of the leading causes of death worldwide and accounts for millions of deaths each year [1]. In low- and middle-income countries (LMICs), timely intervention is often difficult because access to diagnostic testing infrastructure is limited, especially in rural or resource-constrained areas. Typical diagnostic processes involve time-consuming and costly procedures such as thallium stress tests or fluoroscopy. This has resulted in interest in predictive systems as decision support system for situations where full diagnostic features are not readily available.

Machine learning (ML) techniques have shown success in predicting heart diseases. Many studies have used classifiers like Logistic Regression [2], Naïve Bayes [1], Decision Trees [3], and Random Forests [4] for this predicting cardiovascular diseases. Recently, many studies have applied deep learning approaches [5-7] for predicting these diseases. Majority of researches in this domain have used the Cleveland Heart Disease dataset from the UCI Machine Learning Repository [8, 9]. This dataset comprises 13 clinical features and a binary disease label. Many models have achieved high performance on benchmark dataset, the majority rely on the full feature set, as more diagnostic features always improves performance. Because the Cleveland dataset is small and widely used, this study does not use it to claim a new state-of-the-art result. It is used as a common benchmark for controlled comparison. This allows us to study how different classifiers respond when the number of input features is reduced step by step.

These studies show success of ML in CVD detection, however, many studies focus on improving classification accuracy and focus less on the difficulty of collecting the required features in real life. The 13 features in the UCI dataset come from various medical measurements. Some are readily available, such as age and sex, while some, such as thallium stress tests or fluoroscopy, need specialized equipment and expensive clinical facilities. Few studies have focused on examining whether acceptable diagnostic accuracy can still be achieved with fewer features, or if statistically reduced features subsets actually correspond to easier-to-obtain in limited resource settings [7, 10].

In some situations, collecting all 13 features for each patient, for ML based screening, may not always be practical. This highlights the importance of reduced-input based modeling. However, fewer features do not always mean lower cost or faster screening as the practical burden of collect features depends on which features remain in the subset. It is therefore useful to understand how well different ML algorithms perform as the number of input features is progressively reduced.

A number of recent studies have challenged the assumption that all features are always needed. For example, Noroozi et al. [11], Khan et al. [12], Pathan et al. [13], and Bouqentar et al. [14] examined reduced-feature models for heart disease prediction. They have explored methods such as genetic algorithms, recursive feature elimination and hybrid transformer-based approaches. Some of these studies have reported accuracy above 90% even with fewer features. However, most of such studies are conducted in terms of computational efficiency, and not from the aspects of of clinical practicality or the cost of acquiring those features. Theerthagiri et al. [15] and Islam Sumon et al. [16] used deep learning and transformer-based models with fewer inputs to get comparable results. Again, they did not examine reduced features aspect from the perspective of feasibility in resource-constrained contexts.

Our study is aimed at understanding how different ML classifiers perform when trained on progressively smaller subsets of statistically ranked features from a standard benchmark dataset. We derived three reduced-feature datasets from the full 13-feature dataset. These

included 10, 8, and 6 features based datasets. We then compare classifier behavior across all four datasets. The objective is to identify which classifiers are robust as feature availability is progressively reduced.

The results of our study indicate that Logistic Regression, and to a lesser extent Random Forest, maintain strong performance even on the most reduced feature set. This makes them good comparative baselines for reduced-input based prediction in this dataset. The reduced features subsets were selected by statistical ranking and were not based on minimizing the burden of acquisition. Therefore, reduced features datasets in this study should not be interpreted as low cost or time-saving. This study therefore provides empirical baseline on how commonly used classifiers behave under progressive feature reduction. This study also highlights the gap between statistical feature reduction and real-world clinical feasibility. The purpose of this study is not to introduce a new model or to report the highest accuracy on the Cleveland dataset. The purpose of the study is to provide a simple and reproducible baseline analysis of classifier robustness under progressive feature reduction. The main contributions are fivefold:

A controlled empirical evaluation of classifier performance under systematic feature reduction.

Identification of classifiers that remain comparatively resilient as feature availability decreases.

Evidence that comparable predictive performance can be retained with fewer statistically selected input features, highlighting robustness under reduced-input conditions.

A replicable experimental pipeline using the public UCI Cleveland dataset and open-source machine learning libraries.

A cautionary finding that reduced feature count does not necessarily imply lower acquisition burden, because statistically selected subsets may still retain clinically burdensome or infrastructure-dependent variables.

We treat cost/time feasibility as exploratory. We do not quantify test cost or collection time in this study.

The remainder of this paper is organized as follows. Section 2 reviews related work on ML-based heart disease prediction and feature selection. Section 3 explains the methodology, including dataset processing and model evaluation. Section 4 presents the experimental results. Section 5 discusses the implications and limitations. Section 6 concludes the paper and outlines future research directions.

Literature Review

Machine Learning for Heart disease prediction

There has been increase in the use of machine learning (ML) for predicting cardiovascular disease (CVD). This growth is the result of better access to clinical data and advances in machine learning applications. UCI Cleveland Heart Disease dataset is one of the widely used datasets in this domain. This dataset has 13 features based on the patient history, physical exams, and diagnostic tests. Early research work mostly focused on classical ML models such as Naïve Bayes, Decision Trees, K-Nearest Neighbors (KNN), Support Vector Machines (SVM), and Logistic Regression. Zriqat et al. [17] compared the performance of heart disease prediction on both the Cleveland and Statlog datasets. They noted that Decision Trees achieved up to 99% accuracy on the Statlog dataset. Similarly, Pouriyeh et al. [18] examined performance of multiple classifiers and noted that preprocessing and model selection can influence prediction performance.

Many recent studies have assessed deep learning and ensemble methods for Heart disease prediction. Alsabhan and Alfadhly [5] compared traditional ML models, like, Logistic Regression, SVM, and Decision Trees with advanced approaches such as LightGBM, CatBoost and deep neural networks. They observed that deep models outperformed the baseline classifiers. Li [6] also reported an accuracy of 93.4% using a hybrid model based

on Logistic Regression and Deep Neural Network. They observed that even classical methods perform better when used within more modern pipelines.

Some studies have explored biologically inspired and vector-based learning approached. Srinivasan et al. [7] used Learning Vector Quantization (LVQ), Naïve Bayes, and RBF Neural Networks for Heart disease prediction. They reported that LVQ model achieved accuracy score of 98.7% which was comparable to deep learning models. Moreover, hybrid methods have also been reported to improve prediction. Reddy and Khare [19] tested a hybrid model based on combination of Oppositional Firefly algorithm, BAT optimization, and fuzzy logic. Samuel et al. [20] combined Artificial Neural Networks with Fuzzy Analytic Hierarchy Process (Fuzzy-AHP) and reported improved heart disease prediction. Verma et al. [21] proposed an ensemble model that combined Decision Trees, Multi-Layer Perceptron, FURIA, and Multinomial Logistic Regression in one unified framework.

Recently heart disease ML research has started to include privacy-aware modeling and interpretability. Rodriguez and Nafea [22] used both centralized and federated learning models using the UCI dataset and applied Shapley-based methods for explaining predictions. Their SVM model achieved 83.3% accuracy in the federated setting. Meanwhile, Wan [10] offered a systematic review of recent workflows in ML-based heart disease prediction, emphasizing model explainability, and clinical integration.

Despite developments in Heart disease prediction research, most studies still focus on improving accuracy using the complete 13-feature UCI dataset. Using more input variables lead to better prediction performance remains unchallenged. Some studies, however, investigate performance changes in performance of classifiers when only a subset of features is available. In our study, we have compared performance of classifiers when the number of features were reduced based on statistical ranking of the features. Any cost/time impact is discussed qualitatively. This gap motivates the present study. In this study we evaluate a range of classifiers across progressively reduced feature sets. This is done to examine classifier robustness under statistical feature reduction and the limits of treating fewer features as practical feasibility. Therefore, the study is positioned as a baseline comparison rather than a model-development study. The value of the work is in the controlled comparison and in the interpretation of reduced features from a practical point of view.

Feature Selection in Clinical Machine Learning

Feature selection is an important step in healthcare prediction tasks. This is done not only for improving model performance but also for reducing data acquisition burden. In healthcare domain, most features often correspond to a diagnostic tests. Some of these tests may be invasive, expensive, or time-consuming. Choosing the most informative subset of features is, therefore, both a technical and practical concern.

Several studies have addressed feature selection for Heart disease prediction. Gokulnath and Shantharajah [23] used a genetic algorithm (GA) together with Support Vector Machines (SVM) to identify the most relevant features and improved accuracy from 83.70% to 88.34% compared with using all features. Vivekanandan and Iyengar [24] applied a modified Differential Evolution (DE) algorithm for optimal feature selection before classification with Feedforward Neural Networks. In a related study, Saxena and Sharma [25] proposed an efficient prediction system that reduced feature dimensionality through correlation-based methods.

In a broader review of feature selection methods, Jovic et al. [26] highlighted the usefulness of filter-based techniques such as Chi-Square, Information Gain, and Mutual Information for high-dimensional biomedical

datasets. Chi-Square-based selection has also been applied successfully in studies such as Jin et al. [27] for cancer classification and Bachri et al. [28] for academic performance prediction.

While these studies show the importance of feature selection for improving classifier performance, very few of them examine how predictive accuracy changes systematically under progressive feature reduction. Most studies have used one reduced features dataset and optimize for accuracy and do not evaluate which models remain robust while features are removed step by step.

Heart Disease Prediction with Reduced Feature Sets

Recently some studies explored the use of reduced input feature sets for heart disease prediction and demonstrated that comparable classification performance can be achieved without using all available attributes. Noroozi et al. [11] evaluated 16 feature selection methods. They used filter, wrapper, and evolutionary approaches, on the Cleveland dataset and demonstrated that certain reduced-feature setups could improve precision and F1-score, however, sensitivity varied by classifier.

Khan [12] proposed an adapted bee optimization algorithm for feature selection. Their algorithm achieved strong results using a smaller subset of features. Similarly, Pathan et al. [13], Theerthagiri et.al.[15], and Islam Sumon et al. [16] used recursive feature elimination and transformer-based models to identify fewer informative feature sets to achieve performance nearing 90% accuracy. All these approaches framed their contributions around accuracy improvement or computational efficiency. Bouqentar et al. [14] applied forward feature selection across several UCI heart datasets and identified key predictors such as chest pain type, maximum heart rate, exercise-induced angina, and thalassemia. Although the selected features often correspond to standard clinical measurements, only a few studies directly examine how model performance changes when the number of features is progressively reduced.

Taken together, these studies show that reduced-feature modeling is already an active research area. However, most of them identify one optimal subset or evaluate only a limited set of classifiers. Few studies have systematically examined how performance of ML classifier changes across dataset with varying number of features to identify which models demonstrate robustness and resilience to reduced input. Also there is lack of discussion on how reduced-features models could be useful in contexts where full diagnostic testing is not feasible.

Our study contributes to the discussion in this space. We systematically evaluate eleven widely used classifiers across four feature tiers, 13, 10, 8, and 6 features. These reduced features datasets are derived using a Chi-Square ranking method. We are not proposing any novel feature selection technique. This is a deliberate choice. The aim is to keep the feature selection method simple so that the main focus remains on model behaviour across different levels of feature reduction. This work is focused on understanding classifier resilience to input reduction. This will provide empirical evidence about which models retain reliable performance with fewer inputs. These insights provide comparative baseline evidence on reduced-input prediction. These insights also highlight the need for future feature-selection methods that explicitly incorporate clinical burden, test availability, or acquisition cost.

Gaps and Motivation for This Study

Despite extensive research on ML-based heart disease prediction, the following lines of investigation have received limited attention.

Lack of analysis over progressive feature reduction: Most studies evaluate performance of ML models using either the full feature datasets or a single reduced subset. There is lack of studies that systematically examine model performance as features are progressively reduced.

Evaluation of classifier robustness under input constraints: There is lack of understanding on how different classifiers tolerate reduction in input features. This line of research is particularly important in resource-constrained healthcare settings.

Comparative analysis across classifiers and reduction levels: There is lack of studies that compare performance various of classifiers across multiple datasets with feature reduction.

Our study investigates these aspects through a controlled comparative evaluation using the UCI Cleveland Heart Disease dataset. Three progressively smaller subsets (10, 8, and 6 features) are extracted from the full 13-features dataset using Chi-Square feature ranking mechanism. Eleven widely-used classifiers are evaluated across all configurations to record their performance on input reduction. We identify models that retain acceptable performance on reduced features. Using this approach, we contribute insight into classifier robustness under progressive feature reduction and into the limits of interpreting statistically reduced subsets as lower-burden inputs.

Methodology

We used a comparative experimental design to evaluate the performance of popular machine learning (ML) classifiers on the UCI Heart Disease dataset under different feature subsets. Our objective is to study the effect on classification performance of reducing the number of statistically selected features. We also wanted to identify the algorithms that remain robust as features are reduced. The overall experimental framework is presented in Figure 1.

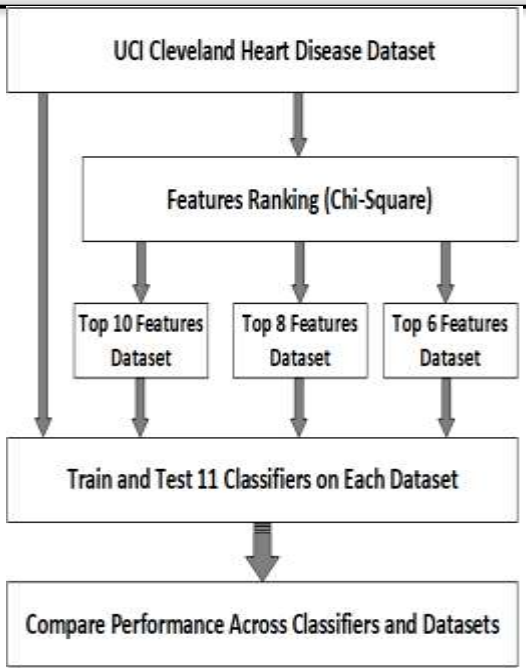


Figure 1: Experimental Framework

Dataset Description

We used the Cleveland Heart Disease dataset from the UCI Machine Learning Repository, which is a commonly studied benchmark dataset in Heart disease prediction research. It contains 303 instances with 13 clinical features and one binary target label to represent the presence (1) or absence (0) of disease. The features in the dataset, which represent common diagnostic measurements, include age, sex, resting blood pressure, cholesterol level, fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved, and chest pain type. The dataset has no missing values. The dataset is small and old, and this limits the generalization of the findings. However, it remains a useful benchmark because it has been widely used in earlier heart disease prediction studies. In this paper, the dataset is used mainly for controlled comparison, not for clinical validation. All categorical features were encoded for use with ML algorithms. The class distribution is moderately balanced. It has 165 instances of heart disease and 138

without. Table 1 provides a brief description of the 13 features included in the dataset.

Datasets Generations

In order to extract other datasets with fewer features, we applied Chi-Square (χ^2) feature selection approach to rank all features and then select the most top ranked variables. Chi-Square is a univariate filter-based method that measures the statistical dependence between each feature and the target class. It is widely used in biomedical classification tasks for its simplicity and computational efficiency [26-28]. As chi-square feature ranking is suitable for categorical or non-negative features, the feature selection step is treated as a simple heuristic rather than as an optimal feature engineering approach.

Using the SelectKBest method from the scikit-learn library with the chi2 scoring function, we identified the top 10, 8, and 6 ranked features and generate the three reduced-feature datasets accordingly. These reduced features datasets, Reduced-10, Reduced-8, and Reduced-6, were used to evaluate the eleven machine learning classifiers, alongside the original Full-13 dataset. The reduced sets were created only from statistical ranking. They were not designed as low-cost clinical screening panels. Therefore, the reduced feature sets should be understood as experimental variants of the dataset. All datasets retained the same number of patient instances. This allowed us to systematically assess how classification performance changes as statistically selected input is reduced from 13 to 6 features. Table 2 presents top selected features for each dataset.

Table 1: Features in the UCI Heart Disease Dataset

Featur	Description	Feature	Descriptio
age	Age of the patient	thalach	Maximum
sex	Sex of the patient	exang	Exercise-
cp	Chest pain type	oldpea	ST
trestbp	Resting blood	slope	Slope of
chol	Serum cholesterol	CA	Number of
fbs	Fasting blood sugar	THAL	Thallium
restecg	Resting		

Table 2: Description of reduced features based datasets

Dataset	Number of Features	Selected Features
Full-13	13	All
Reduced-	10	Age, Sex, Cp, Trestbps,
Reduced-	8	Age, Cp, Trestbps, Chol,
Reduced-	6	Cp, Chol, Thalach, Exang,

Machine Learning Classifiers

In this study, we evaluate eleven machine learning classifiers that are frequently used in existing research work. The models include Logistic Regression; three Naïve Bayes variants, namely Gaussian, Bernoulli, and Multinomial; K-Nearest Neighbor (KNN); Decision Tree; Support Vector Machine with linear and RBF kernels; Random Forest; Gradient Boosting; AdaBoost; and a Artificial Neural Network implemented as a single-hidden-layer Multi-Layer Perceptron (MLP).

These subsets should be interpreted as statistically ranked features, not as any clinically optimized selections based on cost etc. Default scikit-learn parameters were used for all classifiers to keep comparisons uniform. Hyperparameter tuning was not used because the study was not designed to identify the best performance of each classifier by optimizing models. The aim was to compare and study model behaviour under feature reduction.

Experimental Setup and Evaluation

We implemented the experiments using Python and the scikit-learn machine learning library. Each dataset was randomly split into 80% training and 20% testing sets. To account for variance due to data splitting, all results were

averaged over five independent runs using different random seeds.

Performance of the classifiers were assessed using four standard evaluation metrics i.e. **accuracy**, **precision**, F1-score and **recall**. Accuracy measures the overall proportion of correctly classified instances. Precision captures the proportion of true positive among all positive predictions. Recall represents the proportion of actual positives correctly predicted. The F1-score gives a balance between false positives and false negatives. These standard metrics provides a useful view of classifiers predictive performance.

Results

Classifier Performance across All Feature Sets

The results are reported as exploratory baseline results. They should not be read as final clinical performance estimates. The main aim is to compare relative model stability across feature sets. In this we present comparative analysis of performance of the selected classifiers across the four datasets. We consider four standard metrics: accuracy, precision, recall, and F1-score. The performance values are average of five independent random train-test splits.

Accuracy Comparison

Accuracy is the overall proportion of correctly classified instances among all. It is used as a general indicator of model performance. Table 3 reports the accuracy of classifiers on all four datasets. Figure 2 illustrates these results as a grouped bar chart to indicate which models retain accuracy better under feature reduction.

Logistic Regression and Random Forest consistently achieved high accuracy on all four datasets. Their performance decreased by small margin as the number of features got reduced. Logistic Regression achieved the highest accuracy, 0.826, on 10 features. This difference is not large and should not be treated as an evidence that the Reduced-10 set is generally better than the full set. It only shows stability of Logistic Regression in this experimental setting. This was interestingly better than its performance on the full 13-feature set. This highlights its robustness to feature reduction.

In contrast, classifiers like K-Nearest Neighbor (KNN) and Support Vector Machines (SVM) proved to be more sensitive to feature reduction. KNN plateaued around 66–67% accuracy across all variants. It shows its limited capacity to benefit from more features or improve with fewer. SVM's accuracy remained around 65% for most feature sets. Naive Bayes variants (Gaussian, Bernoulli, and Multinomial) also demonstrated stable performance. Particularly, Bernoulli Naive Bayes maintained accuracy close to 80% across all feature sets.

Table 3: Accuracy of Classifiers Across Full and Reduced Feature Sets

Classifier	Full-13	Reduced-10	Reduced-8	Reduced-6
Logistic	0.81	0.826	0.810	0.818
Random	0.81	0.822	0.797	0.802
Naive	0.81	0.805	0.805	0.797
Bernoulli	0.80	0.789	0.797	0.797
Multinomia	0.75	0.768	0.764	0.760
AdaBoost	0.77	0.768	0.739	0.743
ANN	0.73	0.671	0.802	0.756
Linear	0.70	0.748	0.623	0.706
Decision	0.74	0.739	0.723	0.719
KNN	0.66	0.669	0.669	0.649
SVM	0.64	0.648	0.652	0.669

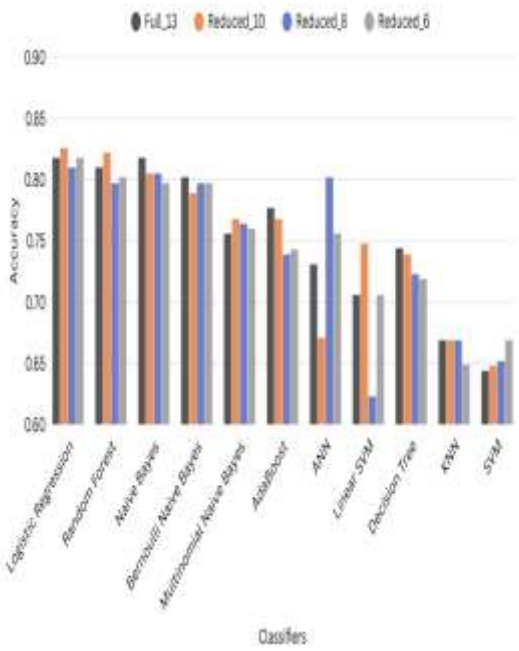


Figure 2: Accuracy Comparison of Classifiers Across Feature Sets

Precision Comparison

Precision is the proportion of correctly predicted positive cases out of all cases predicted as positive. Precision is particularly important in medical diagnosis as false positives may lead to unnecessary anxiety, tests, or treatments for patients. Table 4 presents the precision values for all classifiers for all four datasets. Figure 3 shows illustrates the variations in the results to help identify the models that are robust even with fewer inputs.

For precision values, Logistic Regression and Naive Bayes again appear among the top-performing classifiers. These have high values across all feature sets. Naive Bayes (0.816) and Random Forest (0.825) achieved the highest precision on the full 13-feature dataset. Both also kept strong scores when the feature set was reduced to 6 and 8 features.

Random Forest shows a small decline in precision as the number of features decreases. Its precision drops from 0.825 on Full-13 to 0.779 on Reduced-6. The model still remains competitive overall. Bernoulli Naive Bayes and Multinomial Naive Bayes remain stable. This suggests that probabilistic models can still give better results even when the dimensionality of the data is reduced.

SVM and Linear SVM show the lowest precision on most subsets. Linear SVM has lowest precision of 0.476 on the 8-feature dataset. This instability may indicate a mismatch between the decision boundaries of these models. ANN also gives moderate precision values. ANN has, however, with clear variation across feature sets. Its precision ranges from 0.659 on Full-13 to 0.749 on Reduced-10. This suggest suggests that its performance is sensitive to feature configuration.

Overall, the precision results support the earlier findings. Logistic Regression, Naive Bayes, and Random Forest performs as the most reliable classifiers even when the input features are reduced. This means that they remain comparatively reliable even when the number of input features is reduced.

Table 4: Precision of Classifiers Across Full and Reduced Feature Sets

Classifier	Full-13	Reduced-10	Reduced-8	Reduced-6
Logistic	0.80	0.813	0.798	0.801
Random	0.82	0.797	0.777	0.779
Naive	0.81	0.808	0.812	0.806
Bernoulli	0.79	0.793	0.784	0.784
Multinomia	0.77	0.782	0.784	0.780
AdaBoost	0.79	0.780	0.748	0.775
ANN	0.65	0.749	0.720	0.736
Decision	0.76	0.783	0.757	0.753
KNN	0.68	0.684	0.684	0.667
SVM	0.63	0.637	0.642	0.648
Linear	0.59	0.523	0.476	0.631

Recall Comparison

Recall, also referred to as sensitivity, measures a model's ability to correctly identify positive cases, that is, identify patients who actually have heart disease. In clinical diagnosis, this metric is particularly important because it reduces false negatives. It ensures that patients with the condition are not missed. Table 5 reports the recall scores of all models across all the datasets. Figure 4 presents the same results in the form of a grouped bar chart.

Logistic Regression again shows the strongest recall performance among the classifiers. Its recall values remain above 0.87 for all datasets. These also include an identical value of 0.885 on both the Full-13 and Reduced-10 datasets. This supports its role as a sensitive classifier that stays reliable under reduced number of diagnostic features.

Random Forest and Naive Bayes has strong recall results. Random Forest has its highest recall 0.864 with 10 features. Gaussian Naive Bayes is above 0.83 even on the 6-feature dataset. Bernoulli Naive Bayes shows its best recall of 0.863 on Reduced-6, which is equal to or better than its result on the full dataset. These pattern indicates that a set of statistically selected features may provide strong predictive value for detecting CVD.

ANN shows a more unstable pattern. Its recall is 0.747 on the full feature set, but rises to 0.909 on the 6-feature dataset. This high recall should be interpreted carefully because the dataset is small and the model also shows variation in other metrics. A high value in one metric does not mean that the model is the best overall choice. This suggests that the neural model may be highly influenced by specific inputs or possible overfitting. This behavior needs further investigation.

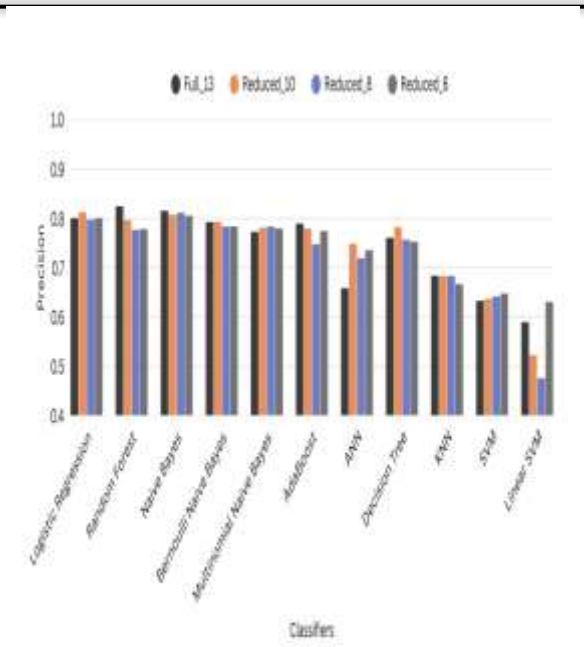


Figure 3: Precision Comparison of Classifiers Across Feature Sets

Table 5: Recall of Classifiers Across Full and Reduced Feature Sets

Classifier	Full-13	Reduced-10	Reduced-8	Reduced-6
Logistic	0.88	0.885	0.871	0.886
Random	0.84	0.864	0.863	0.848
Naive	0.86	0.840	0.841	0.833
Bernoulli	0.85	0.825	0.863	0.863
Multinomia	0.78	0.803	0.788	0.780
AdaBoost	0.81	0.803	0.788	0.750
ANN	0.74	0.849	0.856	0.909
SVM	0.83	0.832	0.825	0.863
Decision	0.74	0.696	0.704	0.765
KNN	0.73	0.734	0.734	0.720
Linear	0.67	0.962	0.553	0.665

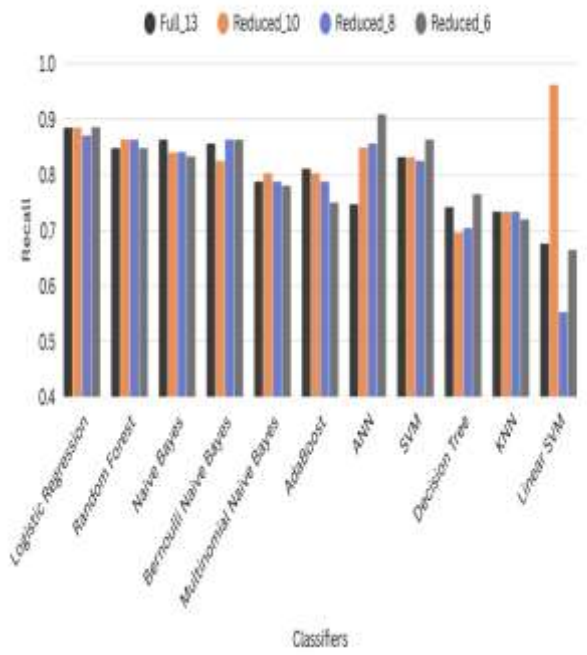


Figure 4: Recall Comparison of Classifiers Across Feature Sets

The performance of Linear SVM was not consistent, achieving a recall of 0.962 on Reduced-10, better than other models. But its recall drops to 0.553 on Reduced-8. This drop represents big inconsistency for generalization across feature set and is considered not reliable for real-world applications. Logistic Regression, Random Forest, and the Naive Bayes variants demonstrated stable recall on all datasets. Such stability is desired for healthcare applications to avoid false negatives.

F1-Score Comparison

The F1-score shows the ability of classifiers of handling both false positives and false negatives. In healthcare predictive applications F1-score is considered very useful metric, especially when the datasets has class imbalance. Table 6 shows the F1-scores of different classifiers on all four datasets. Figure 5 shows the same results in a bar chart format for comparison of performance variations.

Logistic Regression demonstrated the highest and most stable F1-scores across on all datasets. Interestingly, it performance on the Reduced-10 dataset was better than compared to the Full-13. This result supports the view that fewer features may still give strong performance.

Random Forest and Naïve Bayes gave competitive F1-scores between 0.81 and 0.83, even on Reduced-6 dataset. This stability suggests that they may be good considered for settings where to only partial data is accessible. Bernoulli Naive Bayes F1-scores remained around 0.82 on the reduced datasets.

ANN achieved its best F1-score of 0.837 on the Reduced-10 dataset. Its F1-score dropped to 0.729 on the Reduced-6 dataset. This shows that ANN was sensitive to feature reduction. The model performed better with a moderate number of features, but its performance decreased when more important features were removed. Linear SVM and KNN also showed weaker performance on reduced-feature datasets.

Overall, Logistic Regression, Random Forest, and Naive Bayes were the most stable classifiers across all datasets. These models maintained good performance on all evaluation metrics. Their consistent results suggest that they may be suitable in situations where all input features are not available.

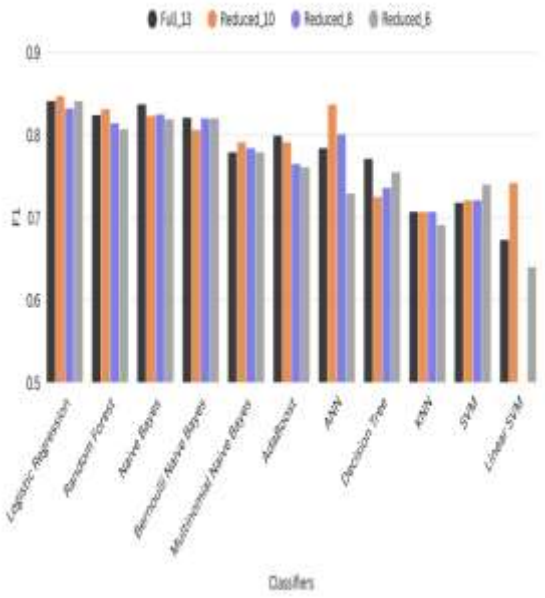


Figure 5: F1 Comparison of Classifiers Across Feature Sets

Table 6: F1 of Classifiers Across Full and Reduced Feature Sets

Classifier	Full-13	Reduced-10	Reduced-8	Reduced-6
Logistic	0.84	0.847	0.832	0.841
Random	0.82	0.831	0.814	0.807
Naive	0.83	0.823	0.825	0.819
Bernoulli	0.82	0.806	0.820	0.820
Multinomia	0.77	0.791	0.784	0.779
AdaBoost	0.79	0.791	0.765	0.761
ANN	0.78	0.837	0.801	0.729
Decision	0.77	0.725	0.736	0.755
KNN	0.70	0.707	0.707	0.691
SVM	0.71	0.721	0.721	0.740
Linear	0.67	0.742	0.493	0.640

Top Performing Classifier–Feature Set Combinations

These rankings are used only to summarize the observed results. They should not be interpreted as a general ranking of models for heart disease prediction. The rankings are specific to this dataset, feature selection method, and experimental setup.

In this section, we identify the top-performing configurations. We rank the top 10 model-feature set combinations for Accuracy, Precision and Recall. This ranking gives a clear overview of which classifiers and feature-sets combination gives better results.

Accuracy Ranking

Table 7 presents the top 10 classifier-feature set combinations ranked by accuracy values. The results suggests that Logistic Regression consistently performed well. It achieved the highest accuracy of 82.6% on the Reduced-10 feature among all classifiers-feature set combinations. It also appears on 3rd, 5th, and 7th position in the table on Reduced-6, Full-13, and Reduced-8, respectively. It, however, showed limited decline in performance on other reduced input. Random Forest has the second highest accuracy of 82.2%, which was again, on the Reduced-10 dataset. It appears on position 6th for the Full-13 dataset, with accuracy value of 0.818, indicating its adaptability under feature reduction.

Logistic Regression accuracy on both full-13 and Reduced-6 dataset is the same (81.8%). This shows that accuracy is maintained with only six top contributing inputs features and removal of features had little impact on it.

Overall, this rankings of results suggests that comparable prediction is possible with reduced feature sets. Logistic Regression and Random Forest were overall top-performer across all datasets including Reduced-6. This robustness in performance supports their consideration for scenarios where all features are not available.

Table 7: Top 10 Classifier–Dataset Combinations Ranked by Accuracy

Rank	Feature Set	Classifier	Accuracy
1	Reduced-10	Logistic	0.826
2	Reduced-10	Random	0.822
3	Reduced-6	Logistic	0.818
4	Full-13	Naive Bayes	0.818
5	Full-13	Logistic	0.818
6	Full-13	Random	0.810
7	Reduced-8	Logistic	0.810
8	Reduced-8	Naive Bayes	0.805
9	Reduced-10	Naive Bayes	0.805
10	Reduced-8	ANN	0.802

Table 8: Top 10 Classifier–Dataset Combinations Ranked by Precision

Rank	Feature Set	Classifier	Precision
1	Full-13	Random Forest	0.825
2	Full-13	Naive Bayes	0.816
3	Reduced-10	Logistic	0.813
4	Reduced-8	Naive Bayes	0.812

Rank	Feature Set	Classifier	Precision
5	Reduced-10	Naive Bayes	0.808
6	Reduced-6	Naive Bayes	0.806
7	Full-13	Logistic	0.801
8	Reduced-6	Logistic	0.801
9	Reduced-8	Logistic	0.798
10	Reduced-10	Random Forest	0.797

Precision Ranking

Table 8 presents the ranking of top 10 classifier-feature set combinations by precision values. Random Forest has highest precision value of 85.5% on the Full-13 dataset. Naïve Bayes on Full-13 dataset is at 2nd position with 81.6%. Logistic Regression on the Reduced-10 dataset comes at third position with a precision value 81.3%. Logistic Regression also appears on seventh place with a precision values of 80.1% on both the Full-13 and Reduced-6 datasets. Its precision value on the Reduced-6 dataset is slightly below the highest ranked score observed on Full-13. It indicates only minor performance loss as the number of features are reduction.

Overall, these results show that Random Forest, Naive Bayes, and Logistic Regression can provide strong precision even with fewer input features. Due to robustness of these classifiers these can be useful in settings where reducing false positives is important with fewer input features.

Recall Ranking

Table 9 shows ranking of classifier-feature set combinations based on the recall values. Linear SVM has highest recall value of 96.2% on the Reduced-10 feature set, however, it does not perform well based on accuracy or precision rankings.

ANN performs with second highest recall value of 90.9% on the Reduced-6 dataset. This indicates that ANN may be considered situations when only fewer features are available. Logistic Regression on the Reduced-6 dataset has 88.6% recall. However, its recall on Reduced-10 dataset is slightly lower on 88.5% while its performance on Full-13 is 88.5%. This indicates good generalization on all datasets. Random Forest and SVM has high recall values of above 86% on both Reduced-10 and Reduced-6.

This results based on ranking recall suggest that Logistic Regression and ANN can maintain high recall with reduced number of features. This should be interpreted as comparative evidence of model behavior under reduced-input conditions, where minimizing false negatives is especially desired.

Table 9: Top 10 Classifier–Dataset Combinations Ranked by Recall

Rank	Feature Set	Classifier	Recall
1	Reduced-10	Linear SVM	0.962
2	Reduced-6	ANN	0.909
3	Reduced-6	Logistic	0.886
4	Reduced-10	Logistic	0.885
5	Full-13	Logistic	0.885
6	Reduced-8	Logistic	0.871
7	Reduced-10	Random Forest	0.864
8	Reduced-6	SVM	0.863
9	Reduced-8	Random Forest	0.863
10	Reduced-6	Bernoulli Naive	0.863

Discussion

The findings should be interpreted with caution. The study uses a small benchmark dataset and common classifiers. Therefore, the main value of the study is not in model novelty, but in showing how classifier performance changes under progressive feature reduction. In this section, we discuss the results of the comparative evaluation of the machine learning classifiers under progressively reduced feature datasets. We highlight the differences in classifier robustness, sensitivity to feature reduction, and the implications of these findings for interpreting reduced-feature results in constrained-input scenarios.

Classifier Robustness and Stability

In this section, we synthesis the response of the classifiers to feature reduction. We compare

their accuracy, precision, recall, and F1-score. Logistic Regression is observed to be the most robust and stable model. Its highest accuracy was recorded on the Reduced-10 dataset (82.6%). It also retained comparable performance on the Reduced-6 dataset (81.8%). Its F1-score for all datasets remained above 84%. Its recall values remained between 87.1% and 88.6%. This consistency under feature reduction indicates that Logistic Regression can generalize well on different inputs. This makes it a strong reduced-input baseline model when complete input features are not available.

Random Forest also demonstrated resilience, especially for precision. It achieved the highest precision score of 82.5% on Full-13 dataset. It also performed well on Reduced-10 dataset (82.2% accuracy, 86.4% recall) and Reduced-6 (80.2% accuracy, 84.8% recall). Its F1-score decreased slightly as the number of input features were reduced. Overall Random Forest maintained strong performance.

Naïve Bayes classifiers (especially the Gaussian variant) also demonstrated respectable performance. Its accuracy stayed around 80.5–81.8% on the Full-13, Reduced-10, and Reduced-8 datasets. Its F1-score remained in the 82–83% range. This performance suggests that Naïve Bayes remains competitive in contexts where complete input features are not always available.

K-Nearest Neighbor and Support Vector Machine (SVM) demonstrated instability. ANN produced a high recall of 90.9% on the Reduced-6 dataset. Its precision and F1-score, however, fluctuated on other datasets. This points to a strong sensitivity of ANN to reduced input.

Overall, Logistic Regression, Random Forest, and Naïve Bayes demonstrated a good balance of stability and efficiency while the input features were reduced. Their robustness supports their use as comparative baselines in situations where it is difficult to collect the full feature set. This does not mean that these models are ready for clinical use. It only means that they are useful baseline models for further reduced-input studies.

Performance Degradation under Feature Reduction

We compared performance of Logistic Regression, the top-performing classifier, across the Full-13 dataset and the three reduced subsets, to understand its performance variation over fewer features. Figure 6 summarizes the overall pattern across all four datasets and presents how Logistic Regression changes across the four dataset. The pattern highlights its robustness under reduced-input conditions.

Interestingly, Logistic Regression performance on the Reduced-10 dataset performs slightly better than the Full-13 dataset in accuracy (82.6% vs. 81.8%) as well as F1-score (84.7% vs. 84.1%). This indicate that removing the three lowest ranked features has little impact on performance loss and may even show some improvement. The hints that the model may be benefiting from a leaner input.

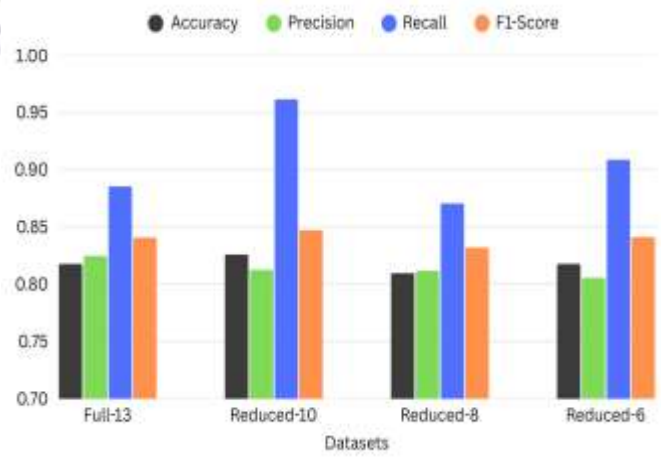


Figure 6: Logistic Regression Performance Across Full and Reduced Feature Sets

There was, however, decline in performance on Reduced-8 dataset. There was decline in scores of all metrics. However, the model still maintained higher recall of 87.1% and F1-score of 83.2%. This again indicates that acceptable performance may be achieved if nearly 40% of

the input features are removed. The values of precision (81.2%) and accuracy (80.9%) are, however, on the lower side on Reduced-10.

Performance was not unstable even at the Reduced-6 dataset. Accuracy value of 81.8% and F1-score value of 84.1% was noted, which is almost the same as on the full-13 dataset. This suggests with Logistic Regression the predictive performance does not decline with less than half of the original features removed. Recall is highest on Reduced-6 (88.6%) suggesting strong sensitivity even with severe input constraints.

These findings show that strong predictive performance can be maintained under statistical feature reduction. However, this should not be interpreted as proof that the resulting subsets are lower-cost or faster to obtain in practice. In this study, reduced subsets still retained variables that may require more advanced clinical infrastructure.

The results therefore point to a mismatch between predictiveness-driven reduction and resource-aware reduction. A model may remain accurate with fewer features, yet the retained subset may still include clinically burdensome measurements. The present findings support classifier robustness under reduced inputs, not direct claims of lower-burden screening. Cost/time savings require explicit cost-aware feature selection. A future study should include a feature-burden score. Such a score may consider cost, test availability, patient effort, and time needed to collect each feature.

Limitations and Future Directions

The study has several limitations because it was designed as an exploratory baseline study. First, the study uses the UCI Cleveland Heart Disease dataset, which is a widely used dataset in this domain, however, it is a relatively small and homogeneous dataset. This makes the generalizability of the results to larger populations is uncertain. Second, this study did not implement hyperparameter tuning or optimization techniques. The selected classifiers were applied with default configurations so that experimental control is maintained. This may have constrained the performance some models. Models like ANN and SVM normally benefit from tuning. Third, the study is motivated by deployment in resource-limited settings, however, it has not explicitly modeled cost, acquisition time, or feature accessibility burden. The reduced features datasets were derived based on statistical importance rather than real-world resource constraints. For example, a reduced subset may still contain a feature that is difficult or costly to collect. Therefore, feature count alone is not enough to define a low-burden screening model. Future work may examine directly the resource requirements in feature selection strategies. Future research can also extend this work in several ways. One direction could be the development of hybrid architectures that combine the interpretability of linear models, such as Logistic Regression, with the predictive robustness of ensemble methods like Random Forest.

Conclusion

This study evaluated the performance of eleven machine learning classifiers on the UCI Cleveland Heart Disease dataset under progressive feature reduction. The study was exploratory in nature. It did not aim to develop a new classifier or claim state-of-the-art performance. Three reduced feature datasets were created by selecting subsets of 10, 8, and 6 features after ranking all features using Chi-Square technique. Our aim was to assess robustness of classifiers and identify those that can maintain acceptable predictive performance with fewer features. This study was motivated by the assumption that reduced-feature models may be more feasible in constrained screening contexts. The selected reduced features-sets, however, show that statistical feature reduction did not necessarily removed clinically burdensome variables.

Overall, he results suggests that Logistic Regression consistently maintained high

accuracy, precision, and recall across original as well as reduced features datasets. Even on the most reduced 6-feature dataset Logistic Regression performed without any significant loss. Random Forest and Naive Bayes also showed some resilience when features were reduced. On the other hand, KNN and SVM were more sensitive and performed poorly on reduced features datasets.

Our findings suggest that Heart disease prediction is possible with fewer numbers of statistically selected features. The reduced sets exclude some features, such as Thal, but still retain other higher-tier features, such as Ca. For this reason, cost or time savings claims cannot be assumed. The main contribution of this study is to provide comparative evidence on classifier robustness under progressive feature reduction. Future work should examine cost-aware feature selection in order to investigate reduced-input prediction under real resource constraints. Overall, the study provides a useful baseline for future work on reduced-input heart disease prediction. Future studies should move beyond statistical feature reduction and include practical feature burden in the selection process.

References:

[1] E. J. Benjamin, P. Muntner, and A. Alonso, "Heart Disease and Stroke Statistics - 2019 Update: A Report From the American Heart Association," *Circulation*, vol. 139, no. 10, 2019.

[2] A. Gavhane, G. Kokkula, I. Pandya, and P. K. Devadkar, "Prediction of Heart Disease Using Machine Learning," in Proceedings of the 2nd International Conference on Electronics, Communication and Aerospace Technology (ICECA 2018), 2018, pp. 1275-1278.

[3] S. Ismaeel, A. Miri, and D. Chourishi, "Using the Extreme Learning Machine (ELM) technique for heart disease diagnosis," in IEEE Canada International Humanitarian Technology Conference (IHTC), 2015, pp. 1-3.

[4] M. Gjoreski, A. Gradišek, M. Gams, M. Simjanoska, A. Peterlin, and G. Poglajen, "Chronic heart failure detection from heart sounds using a stack of machine-learning classifiers," in Proceedings of the 13th International Conference on Intelligent Environments (IE), 2017, pp. 14-19.

[5] W. Alsabhan, and A. Alfadhly, "Effectiveness of machine learning models in diagnosis of heart disease: a comparative study," *Scientific Reports*, vol. 15, no. 1, 2025.

[6] H. W. Li, "Enhancing Cardiovascular Disease Prediction with Machine Learning: A Comparative Study Using the UCI Heart Disease Dataset," in Proceedings of the 2nd International Conference on Data Analysis and Machine Learning (DAML 2024), 2024, pp. 295-300.

[7] S. Srinivasan, and et al., "An active learning machine technique based prediction of," *Nature*, 2023.

[8] S. Ghwanmeh, A. Mohammad, and A. Al-Ibrahim, "Innovative artificial neural networks-based decision support system for heart diseases diagnosis," *Journal of Intelligent Learning Systems and Applications*, vol. 5, no. 3, pp. 176-183, 2013.

[9] J. Maroco, D. Silva, A. Rodrigues, M. Guerreiro, I. Santana, and A. de Mendonca, "Data mining methods in the prediction of dementia: A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests," *BMC Research Notes*, vol. 4, no. 1, pp. 299, 2011.

[10] S. Wan, F. Wan, and X.-j. Dai, "Machine learning approaches for

cardiovascular disease prediction: A review," *Archives of Cardiovascular Diseases*, vol. 118, no. 10, pp. 554-562, 2025.

[11] Z. Noroozi, A. Orooji, and L. Erfannia, "Analyzing the impact of feature selection methods on machine learning algorithms for heart disease prediction," *Scientific Reports*, vol. 13, pp. 22588, 2023.

[12] M. A. Khan, T. Mazhar, M. Mateen Yaqoob, M. Badruddin Khan, A. K. Jilani Saudagar, Y. Y. Ghadi, U. F. Khattak, and M. Shahid, "Optimal feature selection for heart disease prediction using modified Artificial Bee colony (M-ABC) and K-nearest neighbors (KNN)," *Scientific Reports*, vol. 14, 2024.

[13] M. S. Pathan, A. Nag, M. M. Pathan, and S. Dev, "Analyzing the impact of feature selection on the accuracy of heart disease prediction," *Healthcare Analytics*, vol. 2, pp. 100060, 2022.

[14] M. A. Bouqentar, O. Terrada, and S. Hamida, "Early heart disease prediction using feature engineering techniques," *Heliyon*, vol. 10, no. 19, 2024.

[15] P. Theerthagiri, and J. Vidya, "Cardiovascular disease prediction using recursive feature elimination and gradient boosting classification techniques," *Expert systems*, vol. 39, no. 9, pp. e13064, 2022.

[16] M. S. I. Sumon, M. S. B. Islam, M. S. Rahman, M. S. A. Hossain, A. Khandakar, A. Hasan, M. Murugappan, and M. E. Chowdhury, "CardioTabNet: a novel hybrid transformer model for heart disease prediction using tabular medical data," *Health Information Science and Systems*, vol. 13, no. 1, pp. 44, 2025.

[17] I. A. Zriqat, A. M. Altamimi, and M. Azzeh, "A comparative study for predicting heart diseases using data mining classification methods," *International Journal of Computer Science and Information Security (IJCSIS)*, vol. 14, no. 12, 2016.

[18] S. Pouriyeh, S. Vahid, G. Sannino, G. De Pietro, H. Arabnia, and J. Gutierrez, "A comprehensive investigation and comparison of Machine Learning Techniques in the domain of heart disease," in 2017 IEEE Symposium on Computers and Communications (ISCC), 2017, pp. 204-207.

[19] G. T. Reddy, and N. Khare, "An efficient system for heart disease prediction using hybrid OFBAT with rule-based fuzzy logic model," *Journal of Circuits, Systems and Computers*, vol. 26, no. 4, pp. 1750061, 2017.

[20] O. W. Samuel, G. M. Asogbon, A. K. Sangaiah, P. Fang, and G. Li, "An integrated decision support system based on ANN and Fuzzy_AHP for heart failure risk prediction," *Expert Systems with Applications*, vol. 68, pp. 163-172, 2017.

[21] L. Verma, S. Srivastava, and P. C. Negi, "A hybrid data mining model to predict coronary artery disease cases using non-invasive clinical data," *Journal of Medical Systems*, vol. 40, no. 7, pp. 178, 2016.

[22] M. P. Rodriguez, and M. Nafea, "Centralized and Federated Heart Disease Classification Models Using UCI Dataset and their Shapley-value Based Interpretability," *arXiv preprint arXiv:2408.06183*, 2024.

[23] C. B. Gokulnath, and S. P. Shantharajah, "An optimized feature selection based on genetic approach and support vector machine for heart disease," *Cluster Computing*, vol. 22, no. 6, pp. 14777-14787, 2019.

[24] T. Vivekanandan, and N. C. S. N. Iyengar, "Optimal feature selection using a modified differential evolution

algorithm and its effectiveness for prediction of heart disease," *Computers in Biology and Medicine*, vol. 90, pp. 125-136, 2017.

[25] K. Saxena, and R. Sharma, "Efficient heart disease prediction system," in Procedia Computer Science, 2016, pp. 962-969.

[26] A. Jović, K. Brkić, and N. Bogunović, "A review of feature selection methods with applications," in 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2015, pp. 1200-1205.

[27] X. Jin, A. Xu, R. Bie, and P. Guo, "Machine Learning Techniques and Chi-Square Feature Selection for Cancer Classification Using SAGE Gene Expression Profiles," *Data Mining for Biomedical Applications (BioDM 2006)*, J. Wang and M. Zaki, eds., Berlin, Heidelberg: Springer, 2006.

[28] O. S. Bachri, M. Hatta, and O. D. Nurhayati, "Feature Selection Based On Chi Square In Artificial Neural Network To Predict The Accuracy of Student Study Period," *International Journal of Civil Engineering and Technology*, vol. 8, no. 8, pp. 731-739, 2017.