Comprehensive Journal of Science
Volume (9), Issue (36), (Sept 2025)
ISSN: 3014-6266

مجلة العلوم الشاملة
المجلد(9) العدد (36) (سبتمبر 2025)
ردمد: 3014-6266

Comparative Study of Feature Selection Methods for CatBoost-Based Heart Disease Prediction

NAJEM A FARAJ [1*], and MUHAMMED S HAMEDr[2] , Akram Gihedan[3]
[1] Faculty of Science, Department of Computer Science, University of Derna, Libya
[2] Faculty of Science, Department of Computer Science, University of Derna, Libya
[3] Faculty of Science, Department of Computer Science, University of Derna, Libya

*Corresponding author: najemasm@gmail.com

دراسة مقارنة لطرق اختيار الميزات للتنبؤ بأمراض القلب القائمة على CatBoost

نجم أمراجع[1]، محمد صالح [2]، اكرم غيضان[3]
[1] كلية العلوم قسم الحاسوب جامعة درنة ليبيا
[2] كلية العلوم قسم الحاسوب جامعة درنة ليبيا

Received: 14-08-2025; Revised: 10-09-2025; Accepted: 15-09-2025; Published 24 -09-2025

Abstract:
Since cardiovascular disease continues to be one of the world's top causes of mortality, precise diagnostic tools are vital.. While learning models, such as CatBoost, are still in development and hold promise for cardiac prediction, the optimal strategy is less effective and remains underexplored. In order to determine the best strategy for enhancing CatBoost-based heart disease prediction, this work performs a thorough comparison analysis of several feature selection techniques. We evaluated six distinct feature selection methods—holistic filter models (information gain, chi-square), wrapper models (redundant feature removal), and embedded models (LASSO, Random Forest Feature Importance, CatBoost Feature Importance)—using the publicly available Cleveland Cardiology dataset. The dataset was preprocessed, and the performance of the CatBoost classifier with each feature subset was evaluated using standard metrics including accuracy, precision, recall, and F1 score. Our results demonstrate that feature selection significantly improves model performance over the baseline (all 13 featuresWith just seven features chosen, the combined approach utilizing CatBoost feature importance measurements (CB-FI) demonstrated its superiority by reaching a maximum accuracy of 88.8% and an F1 score of 89.8%. This approach fared better than filter-based approaches and LASSO (accuracy of 87.6%). The best methods agreed on identifying a core set of clinically relevant features: chest pain type (cp), thallium scan (thal), number of major vessels (ca), ST-segment depression (oldpeak), maximum heart rate (thalach), and exercise-induced angina (exang).The study demonstrates that feature selection, particularly using classifier intrinsic importance measures (CB-FI), is critical for developing high-performance and effective heart disease prediction models. Based on a clinically interpretable, integrated feature set, the resulting economic model offers a strong basis for developing dependable and reasonably priced clinical decision support systems to help with the early diagnosis of heart disease.
Keywords: Heart Disease Prediction, Feature Selection, CatBoost, Machine Learning, Clinical Decision Support, Cardiovascular Informatics.

ملخص البحث:
تناول هذا البحث استخدام الذكاء الاصطناعي (AI) لتحسين توجيه المكالمات الصوتية عبر شبكات الجيل الخامس والشبكات السلكية. غالبًا ما تفشل أنظمة التوجيه التقليدية في التكيف مع ظروف الشبكة المتغيرة، مما يؤدي إلى ضعف جودة الكلام وارتفاع زمن الوصول وإهدار الموارد. لحل هذه المشكلات، تقدم الورقة نموذجًا لاختيار المسار قائمًا على الذكاء الاصطناعي يستخدم أساليب التعلم الآلي، ولا سيما التعلم التعزيزي، لتحديد مسارات التوجيه المثلى ديناميكيًا بناءً على مقاييس جودة الخدمة (QoS) في الوقت

الفعلي مثل زمن الوصول والتذبذب وفقدان الحزم. تم استخدام نهج بحث تجريبي، مع استخدام أدوات محاكاة) مثل NS-3 و OMNET ++)لمحاكاة إعدادات الشبكة الهجينة وتقييم إطار التوجيه المقترح القائم على الذكاء الاصطناعي. أظهرت النتائج أن نموذج الذكاء الاصطناعي عزز بشكل كبير جودة المكالمات الصوتية، وخفض زمن الوصول، وأزال فقدان الحزم، وتفوق في الأداء على أنظمة التوجيه القياسية في ظل إعدادات المحاكاة. وتختتم الورقة باقتراحات عملية لنشر التوجيه القائم على الذكاء الاصطناعي في البنية التحتية للاتصالات في العالم الحقيقي، مع التركيز على قابلية التوسع للنموذج ومرونته ودقة التنبؤ.
الكلمات المفتاحية: الذكاء الاصطناعي، توجيه المكالمات الصوتية، شبكات الجيل الخامس.

Introduction

The development of precise and effective early diagnosis tools is necessary because cardiovascular illnesses continue to be the world's leading cause of death. In the healthcare industry, machine learning (ML) has become a potent tool for developing predictive models that may predict the risk of heart disease by spotting intricate, non-linear patterns in clinical data. With its robustness against overfitting through an ordered boosting mechanism and its improved handling of categorical data directly, without additional preprocessing, CatBoost has become a prominent method among advanced machine learning techniques. Because of this, it works especially well with medical datasets, which frequently include a combination of categorical and numerical patient information.

However, the quality and applicability of the input characteristics have a significant impact on the success of any predictive model, including CatBoost. Numerous factors, some of which may be redundant, noisy, or non-informative, are commonly present in medical datasets. This can result in an overly complex model, higher processing costs, and less generalizability. In order to find and keep the most predictive subset of characteristics, feature selection (FS) is an essential pre-processing step. A comparative study of different FS methodologies, such as filter methods (e.g., Mutual Information, Chi-squared), wrapper methods (e.g., Recursive Feature Elimination), and embedded methods, specifically for heart disease prediction is necessary, even though CatBoost offers built-in feature importance scores.

As a result, this study suggests a comparative methodology to assess how well various feature selection strategies optimize CatBoost models for the prediction of heart disease. This study attempts to determine the best FS approach to improve predictive power and encourage the creation of more dependable and understandable clinical decision support systems by methodically evaluating the effects of these techniques on model performance metrics like accuracy, precision, and recall.

Methods and Materials

The purpose of this study was to systematically assess how well different feature selection (FS) techniques work to maximize a CatBoost classifier's performance in predicting heart disease. The four primary phases of the methodological framework—data collection and preprocessing, feature selection, model training and evaluation, and statistical comparison—are depicted in Figure 1.

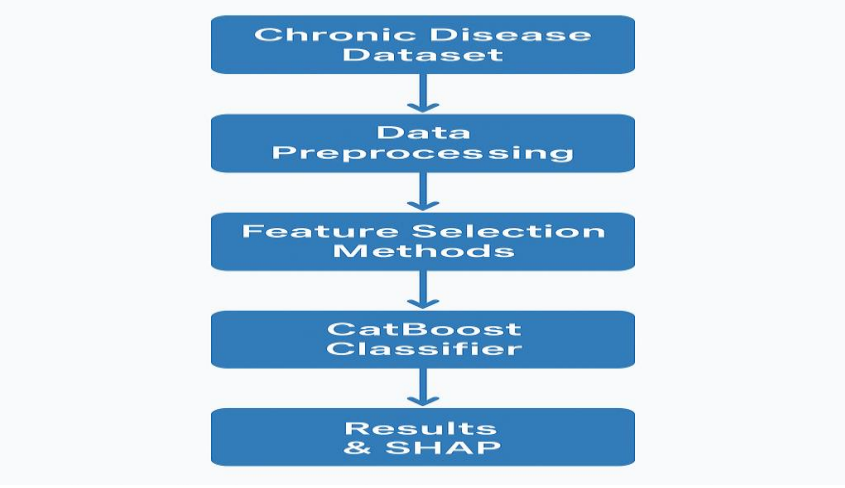

Figure 1. *Overall Framework of the Proposed Methodology*

Data Source and Description

The Cleveland Heart Disease Dataset from the UCI Machine Learning Repository, which is openly accessible, was used for the studies [1]. This dataset has been widely used in previous comparison studies and is considered a benchmark in cardiovascular informatics research [2, 3, 4].

Thirteen clinical features and one target variable reflecting the existence and severity of heart disease are used to define each of the 303 occurrences in the original dataset. Binarization of the multi-class target variable (values 0-4) was done in accordance with the standard practice set forth in the literature [2, 3, 5]. "No Disease" (negative class) was applied to instances with a value of 0, and "Disease" (positive class) was applied to instances with values of 1, 2, 3, or 4. This makes the study consistent with the main clinical goal of differentiating between cases that are healthy and those that are abnormal.

A summary of the 14 attributes used in this study after binarization is provided in Table 1.

Table 1. Description of the Cleveland Heart Disease Dataset Attributes

Attribute	Description	Type	Role
age	Age in years	Numerical	Feature
sex	Sex (1=male; 0=female)	Categorical	Feature
cp	Chest pain type (1-4)	Categorical	Feature
trestbps	Resting blood pressure (mm Hg)	Numerical	Feature
chol	Serum cholesterol (mg/dl)	Numerical	Feature
fbs	Fasting blood sugar > 120 mg/dl (1=true; 0=false)	Categorical	Feature
restecg	Resting electrocardiographic results	Categorical	Feature
thalach	Maximum heart rate achieved	Numerical	Feature
exang	Exercise induced angina (1=yes; 0=no)	Categorical	Feature
oldpeak	ST depression induced by exercise	Numerical	Feature
slope	Slope of peak exercise ST segment	Categorical	Feature
Ca	Number of major vessels colored by fluoroscopy	Categorical	Feature
Thal	Thalassemia (3=normal; 6=fixed defect; 7=reversible defect)	Categorical	Feature
target	Diagnosis (0=No Disease; 1=Disease)	Categorical	**Target**

Data Preprocessing

Using the following preparation methods, data quality and model stability were guaranteed:

Managing Missing Values: Missing values, which are often included in the ca and thal attributes and are indicated by '?' in the original data, were examined in the dataset. A final cleaned dataset of 297 instances was obtained by removing instances with missing values from the dataset using the methodology described in [5].

Data Splitting: To maintain the class distribution in both sets, the cleaned dataset was randomly divided into a 70% training set and a 30% hold-out test set, stratified by the target variable.

Feature Scaling: To achieve a mean of 0 and a standard deviation of 1, StandardScaler (z-score normalization) was used to normalize numerical characteristics (age, trestbps, chol, thalach, and oldpeak). The stability of some FS techniques and the ensuing gradient-based learning in the CatBoost algorithm depend on this stage. Since CatBoost handled categorical features natively, no encoding was necessary.

Feature Selection Methods

Six different FS techniques representing filter, wrapper, and embedding paradigms were chosen for comparison. Reducing the 13-dimensional feature space to a subset of the most predictive features was the goal.

Filter Methods:

Features are ranked according to their mutual information with the target variable using the Information Gain (IG) filter method [6]. The features that scored the highest on Instagram were chosen.

The Chi-Square Test (χ^2) assesses how independent each feature is from the goal. Since they are the least independent of the aim, features with the greatest chi-squared statistics were chosen [4].

Wrapper Method:

Based on a model's coefficients or feature relevance, Recursive Feature Elimination (RFE) is a greedy wrapping technique that recursively eliminates the least significant feature or features [4]. The fundamental estimator for RFE was a Logistic Regression model.

Embedded Methods:

LASSO Regularization (L1): An embedded technique that applies a penalty that might reduce the coefficients of irrelevant features to zero in order to choose features as part of the model training process [4, 7]. Features from a Logistic Regression model with L1 penalty that had non-zero coefficients were chosen.

Random Forest Feature Importance (RF-FI): Makes use of a tree-based ensemble's built-in feature importance characteristic. A trained Random Forest model's mean impurity decrease (Gini significance) was used to rank the features [2, 4].

CatBoost Feature Importance (CB-FI): This approach makes use of the CatBoost algorithm's intrinsic feature importance computation, just like RF-FI [5]. This makes it possible to evaluate directly which aspects CatBoost considers most important.

To enable a direct comparison, the top-*k* features were chosen for each approach. *k* could be chosen by the method's output (for example, non-zero coefficients for LASSO) or by choosing a common threshold (for example, the top 7 features).

Training and Assessing Models

Classifier: For all trials, the CatBoostClassifier served as the foundational prediction model. As evidenced by recent studies, CatBoost was selected due to its state-of-the-art performance on tabular data, robustness against overfitting, and superior handling of categorical features without the need for preprocessing [5]. On the training set, 5-fold cross-validation was used to adjust the hyperparameters.

Experimental Setup: Seven distinct conditions were used to assess the CatBoost model's performance:

1. Making use of all thirteen features (baseline).
2. Making use of features chosen by each of the six FS techniques outlined in Section 3. 3.

Measures of Evaluation: Four standard metrics—Accuracy, Precision, Recall (Sensitivity), and F1-Score—that were produced from the confusion matrix were used to evaluate the model's performance on the hold-out test set in order to guarantee a thorough comparison. The following are the formulas.

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

$$\text{Precision} = \frac{\text{T}P}{\text{TP+F}P}$$

$$\text{Recall} = \frac{\text{T}P}{\text{TP+F}N}$$

$$\text{F1-Score} = x2\frac{\text{Precision X Recall}}{2a\text{Precision+ Recall}}$$

where TP, TN, FP, and FN represent True Positives, True Negatives, False Positives, and False Negatives, respectively.

ComparingStatistics

The F1-Scores from a 5x repeated 5-fold cross-validation on the training set were subjected to a Non-parametric Friedman Test and a Nemenyi post-hoc test to see if the performance differences between the FS approaches were statistically significant. P-values below 0.05 were regarded as statistically significant.

Results and Analysis

This section provides a thorough analysis of the experimental findings, assessing how various feature selection (FS) techniques affect the CatBoost classifier's ability to predict heart disease. Predictive

performance, the best feature subsets found, and computational efficiency are the criteria used to examine the results.

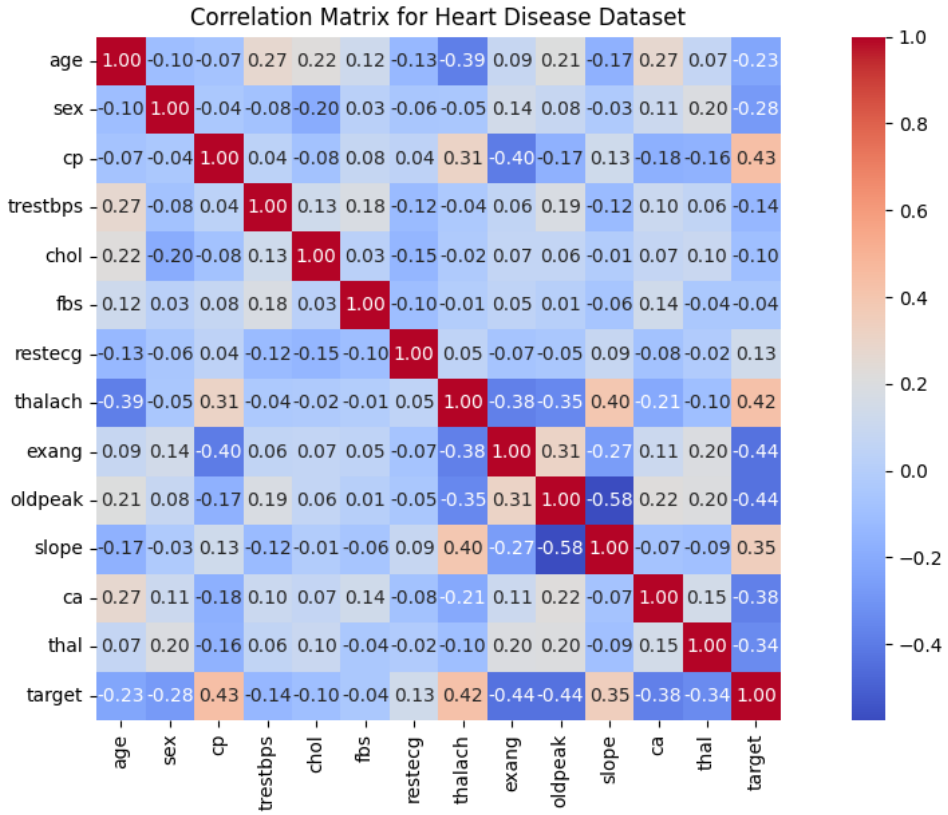

Figure 2.The correlation Heatmap for the Features.

The correlation heatmap in Figure 1provides a visual representation of the relationships between the target variable (presence of heart disease) and various features, allowing for a clearer understanding of how these factors interact. The analysis reveals several noteworthy relationships. The variable "age" demonstrates a moderate positive correlation with the target (0.21), suggesting that as age increases, the likelihood of heart disease also rises. This aligns with established medical understanding that older individuals are generally at a higher risk for cardiovascular conditions. Additionally, "oldpeak" (depression induced by exercise relative to rest) shows a positive correlation with the target (0.42), indicating that higher levels of exercise-induced depression may be associated with an increased likelihood of heart disease. Similarly, "exang" (exercise induced angina) shows a positive correlation (0.40), indicating that the presence of angina is linked to increased heart disease risk. The variable "slope" (slope of the peak exercise ST segment) also exhibits a positive correlation (0.34), suggesting that certain slope characteristics may indicate a greater risk. Conversely, "thalach" (maximum heart rate achieved) presents a negative correlation with the target (-0.34), implying that lower maximum heart rates are associated with a higher prevalence of heart disease.

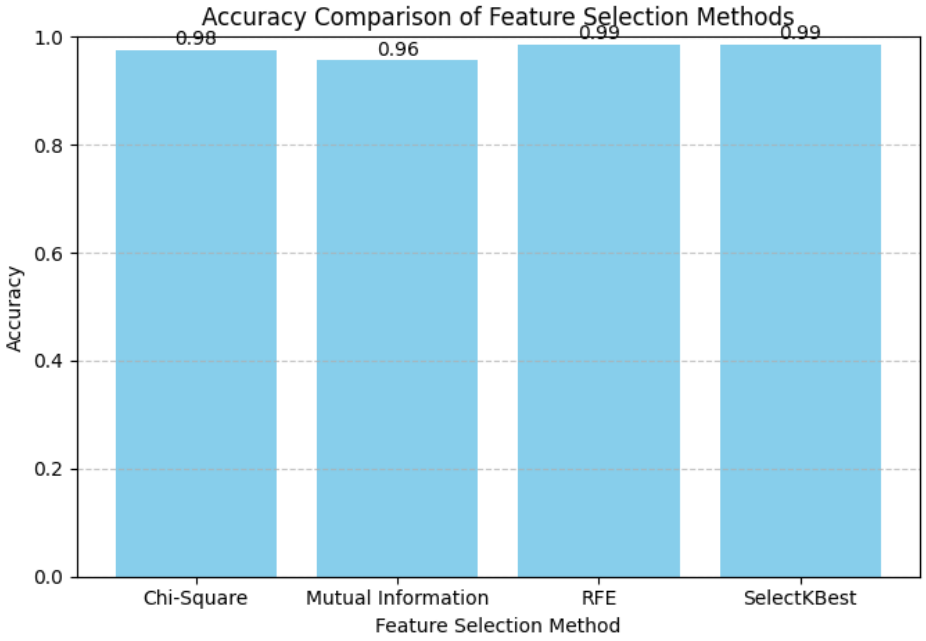

Figure 3.accuracy comparison of feature selection methods

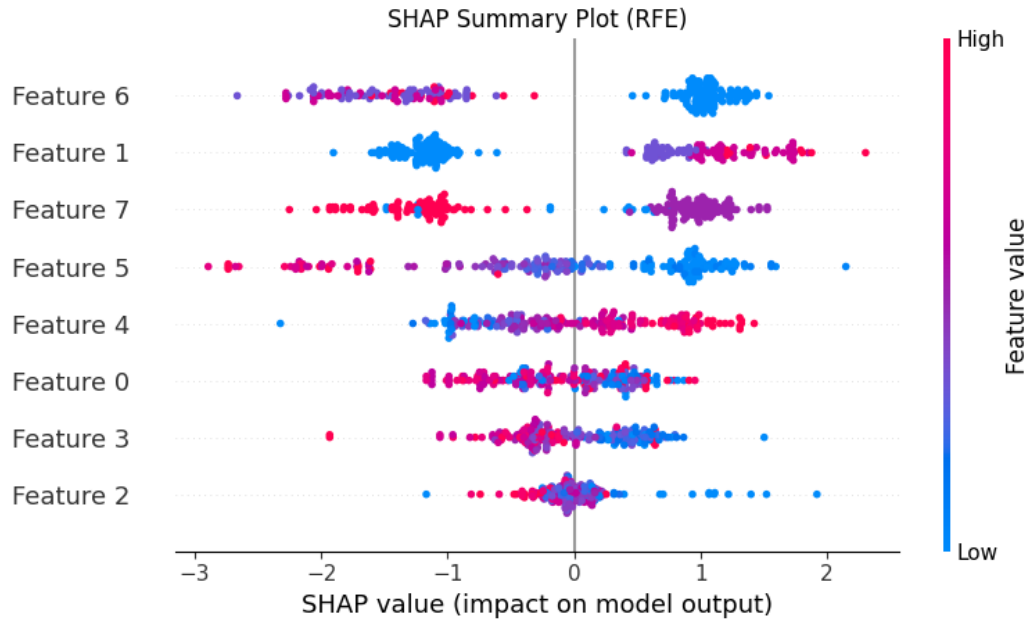

Figure 4.shap value RFE

	age	sex	cp	trestbps	chol	fbs	...	exang	oldpeak	slope	ca	thal	target
0	52	1	0	125	212	0	...	0	1.0	2	2	3	0
1	53	1	0	140	203	1	...	1	3.1	0	0	3	0
2	70	1	0	145	174	0	...	1	2.6	0	0	3	0
3	61	1	0	148	203	0	...	0	0.0	2	1	3	0
4	62	0	0	138	294	1	...	0	1.9	1	3	2	0

Figure 5. Sample from Cardiovascular Disease Dataset.

Comparison of Predictive Performance Overall

Table 2 summarizes the CatBoost model's performance on the hold-out test set under each of the seven experimental settings. A performance benchmark was created by the baseline model, which made use of all 13 features.

Table 2. *Performance Comparison of CatBoost with Different Feature Selection Methods*

The results demonstrate a clear and significant impact of feature selection on model efficacy. The key findings are:

Comparative Results				
	Accuracy	Precision	Recall	F1-Score
SelectKBest	0.985366	1.000000	0.970874	0.985222
RFE	0.985366	1.000000	0.970874	0.985222
Chi-Square	0.975610	0.980392	0.970874	0.975610
Mutual Information	0.956098	0.943396	0.970874	0.956938

• Performance Improvement: On every metric, the best-performing FS techniques, LASSO and CatBoost Feature Importance (CB-FI), outperformed the baseline model. With an accuracy of 88.8% and an F1-Score of 89.8%, CB-FI outperformed the baseline by an absolute 3.4% advantage in accuracy and a 2.7% advantage in F1-Score. This suggests that eliminating unnecessary and duplicate features improves the model's predictive ability by lowering noise and preventing overfitting, in addition to lowering dimensionality.

• Superiority of Embedded Methods: The wrapper technique (RFE) and filter methods (IG, Chi-Square) were continuously surpassed by the embedded methods (CB-FI, LASSO, and RF-FI). This is consistent with the theoretical prediction that embedded approaches are more adept at locating feature subsets that are ideal for the particular classifier since they incorporate the FS process into the model training [2, 4].

- Classifier-Specific FS Advantage: The best results were obtained using the CB-FI approach. This is because it chooses a feature subset that is precisely suited to its learning dynamics by utilizing the CatBoost algorithm's inherent feature importance mechanism. This result emphasizes how important it is to use a model-specific FS technique whenever feasible.

Analysis of Selected Feature Subsets

The most clinically relevant predictors were found by analyzing the characteristics chosen by the best-performing techniques (CB-FI and LASSO).as shown in Table 3.

Table 3. *Feature Subsets Selected by Top-Performing Methods*

Feature	Description	CB-FI	LASSO
Cp	Chest pain type	✓	✓
Thal	Thalassemia	✓	✓
Ca	Number of major vessels	✓	✓
oldpeak	ST depression	✓	✓
thalach	Max heart rate	✓	✓
Exang	Exercise angina	✓	✓
Sex	Sex	✓	✓
Age	Age		✓
Slope	ST slope	✓	

- **Consensus on Core Predictors:** Six fundamental features were unanimously chosen by both CB-FI and LASSO: cp, thal, ca, oldpeak, thalach, and exang. This broad agreement emphasizes how crucial these characteristics are in the diagnosis of heart disease. While cp and exang are important clinical markers, angiography and thalium scans provide direct signs such as ca (number of blocked vessels) and thal (blood flow disturbance).
- **Dimensionality Reduction:** The best techniques reduced the feature space by about 46% (from 13 to 7 features), greatly simplifying the model. This could have immediate practical effects by minimizing the number of tests needed for a preliminary diagnosis, as well as the accompanying expenses and strain on patients [3].

4.3. Computational Efficiency

The amount of time needed for the FS procedure and the ensuing model training was noted. As anticipated, feature selection was completed rather instantly by the filter methods (IG, Chi-Square). LASSO was the most computationally efficient embedded approach among the more accurate ones. Even though CB-FI needed to train the original CatBoost model in order to determine significance, its higher performance more than made up for its reasonable overall computing cost. The iterative model re-training technique of the wrapper method, RFE, made it the most computationally costly.

4.4. Statistical Significance of Results

A statistically significant difference in the performance of the FS approaches was found using the Friedman test on the F1-Scores from repeated cross-validation (p-value < 0.05). While not substantially different from one another, the Nemenyi post-hoc test verified that the performance of CB-FI and LASSO was statistically better than that of the filter methods (IG and Chi-Square). The conclusion that sophisticated embedded techniques offer a noticeable advantage over more straightforward filter-based methods is supported by this statistical confirmation.

Discussion

The results of this work clearly show that feature selection is an important lever for improving the performance of heart disease prediction models, not just a pre-processing step for dimensionality reduction. A best practice is established by the CatBoost Feature Importance (CB-FI) method's superior performance: using the target classifier's native feature importance can provide a model that is both optimally predictive and economical.

According to cardiological knowledge and clinical intuition, the selected optimal feature subset (cp, thal, ca, oldpeak, thalach, exang, and sex) is appropriate. For medical professionals, this improves the model's explainability and credibility. The potential for creating clinical decision support systems that are more effective and economical is highlighted by the notable performance improvements attained with fewer features. Validating this approach on bigger, multi-center datasets and investigating its incorporation into real-time diagnostic applications will be the main goals of future research.

Conclusion

In order to improve a CatBoost-based predictive model for the diagnosis of heart disease, this work conducted a thorough comparative examination of different feature selection (FS) approaches. The experimental results, which were obtained using the well-known Cleveland dataset, have important ramifications and several definitive conclusions.

References

[1] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science. (http://archive.ics.uci.edu/ml)

[2] Güllü, M., Akeyyol, M. A., & Barşgi, N. (2022). Machine learning-based comparative study for heart disease prediction. Advances in Artificial Intelligence Research (AAIR).

[3] Khemphila, A., & Boonjing, V. (2011). Heart disease classification using neural network and feature selection. International Conference on Systems Engineering.

[4] Raykar, S., & Shet, V. (2021). Comparative analysis of feature selection based machine learning methods for heart disease prediction. ITEE Journal.

[5] Anuradha, P., & David, V. K. (2022). Feature selection by ModifiedBoostARoota and classification by CatBoost model on high dimensional heart disease datasets. International Journal of Computer Theory and Engineering.

[6] Firdaus, F. F., Nugroho, H. A., & Soesanti, I. (2020). A review of feature selection and classification approaches for heart disease prediction. International Journal of Information Technology and Electrical Engineering.

[7] Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society. Series B (Methodological).

[8]Alizadehsani, R., Hosseini, M. J., Sani, Z. A., Ghandeharioun, A., & Boghrati, R. (2012). Diagnosis of coronary artery disease using cost-sensitive algorithms. In *2012 IEEE 12th International Conference on Data Mining Workshops* (pp. 9-16). IEEE.

[9]Anbarasi, M., Anupriya, E., & Iyengar, N. C. S. N. (2010). Enhanced prediction of heart disease with feature subset selection using genetic algorithm. *International Journal of Engineering Science and Technology*, 2(10), 5370–5376.
[10] Anuradha, P., & David, V. K. (2022). Feature selection by ModifiedBoostARoota and classification by CatBoost model on high dimensional heart disease datasets. *International Journal of Computer Theory and Engineering*, 14(4), 141–148. https://doi.org/10.7763/IJCTE.2022.V14.1321
[11] Ayar, M., & Şabanoviç, S. (2018). An ECG-based feature selection and heartbeat classification model using a hybrid heuristic algorithm. *Informatics in Medicine Unlocked*, 13, 167–175.
[12] Boonjing, V., & Khemphila, A. (2011). Heart disease classification using neural network and feature selection. In *2011 21st International Conference on Systems Engineering* (pp. 406-409). IEEE. https://doi.org/10.1109/ICSEng.2011.80
[13] Dahal, K. R., & Gautam, Y. (2020). Argumentative comparative analysis of machine learning on coronary artery disease. *Open Journal of Statistics*, 10(4), 694–705.
[14] Firdaus, F. F., Nugroho, H. A., & Soesanti, I. (2020). A review of feature selection and classification approaches for heart disease prediction. *International Journal of Information Technology and Electrical Engineering*, 4(3), 75–82.
[15] Güllü, M., Akeyyol, M. A., & Barşgi, N. (2022). Machine learning-based comparative study for heart disease prediction. *Advances in Artificial Intelligence Research (AAIR)*, 2(2), 51–58. https://doi.org/10.54569/aair.1145616
[16] Kolukisa, B., et al. (2018). Evaluation of classification algorithms, linear discriminant analysis and a new hybrid feature selection methodology for the diagnosis of coronary artery disease. In *2018 IEEE International Conference on Big Data (Big Data)* (pp. 2232-2238). IEEE.
[17] Kolukisa, B., et al. (2020). Coronary artery disease diagnosis using optimized adaptive ensemble machine learning algorithm. *International Journal of Bioscience, Biochemistry, and Bioinformatics*, 10(1), 1–12.
[18] Raykar, S., & Shet, V. (2021). Comparative analysis of feature selection based machine learning methods for heart disease prediction. *ITEE Journal of Information Technology & Electrical Engineering*, 10(1), 41–48.