Importance of Feature Selection for Multiple Disease Classification

Rio Arya Andika[1*], Christine Dewi[2]
Program Studi S1 Teknik Informatika, Fakultas Teknologi Informasi, Universitas Kristen Satya Wacana
Jl. Dr. O. Notohamidjodjo, Blotongan, Sidorejo, Kota Salatiga, Jawa Tengah, Indonesia
Email: [1]672021166@student.uksw.edu, [2]christinedewi@uksw.edu

Abstrak. Pentingnya Pemilihan Fitur untuk Klasifikasi Berbagai Penyakit. *Kinerja machine learning dalam klasifikasi penyakit sangat bergantung pada pemilihan fitur yang tepat. Penelitian ini mengeksplorasi metode seleksi fitur—Boruta dan Recursive Feature Elimination (RFE)—dengan model ensemble seperti Random Forest, Decision Tree, Gradient Boosting, LightGBM, dan XGBoost menggunakan data Electronic Health Records (EHR). Hasil menunjukkan bahwa kombinasi Boruta dan LightGBM menghasilkan akurasi tertinggi sebesar 99%. Seleksi fitur meningkatkan presisi dengan fokus pada variabel relevan dan menghapus yang tidak perlu. Analisis lebih lanjut menunjukkan fitur seperti Red Blood Cells, Insulin, Heart Rate, dan Cholesterol sangat mempengaruhi klasifikasi penyakit tertentu. Temuan ini menyoroti pentingnya seleksi fitur dalam klasifikasi multi-penyakit dan analisis data medis, serta meningkatkan efisiensi sistem machine learning. Penelitian selanjutnya disarankan untuk mengembangkan metode seleksi fitur yang lebih fleksibel dan menguji model pada dataset penyakit yang lebih beragam.*
Kata Kunci: *seleksi fitur, pembelajaran mesin, diagnosis penyakit, akurasi klasifikasi*

Abstract. *The performance of machine learning in disease classification heavily depends on effective feature selection. This study explores feature selection methods—Boruta and Recursive Feature Elimination (RFE)—with ensemble models like Random Forest, Decision Tree, Gradient Boosting, LightGBM, and XGBoost using Electronic Health Records (EHR) data. Results show that combining Boruta with LightGBM achieves the highest accuracy of 99%. Feature selection enhances precision by focusing on relevant variables and removing unnecessary ones. Further analysis reveals that features such as Red Blood Cells, Insulin, Heart Rate, and Cholesterol significantly influence the classification of specific diseases. These findings highlight the importance of feature selection in multi-disease classification and medical data analysis, improving the efficiency of machine learning systems. Future research should develop more flexible feature selection methods and test models on diverse disease datasets.*
Keywords: *feature selection, machine learning, disease diagnosis, accuracy performance*

1. Introduction

Electronic Health Records (EHR) are crucial in advancing machine learning applications in the medical field, especially in disease classification. EHRs provide detailed and complex patient data, which is essential for building accurate and reliable disease diagnosis models [1][2][3]. With the vast amount of real-time and historical health information, EHRs help machine learning systems identify key variables while eliminating unnecessary data. Recent innovations in EHRs have garnered significant research interest due to their potential to improve patient outcomes, streamline healthcare processes, and support personalized medicine [4].

EHRs enable continuous analysis of various diseases by leveraging the accumulated data over time. This capability allows healthcare providers to detect patterns, predict disease progression, and personalize treatment plans effectively [5]. Continuous monitoring is especially beneficial in the management of chronic diseases, as it facilitates the early identification of complications and permits timely interventions. Machine learning classifiers such as LogitBoost, Random Forest, XGBoost, Decision Tree, and Support Vector Machine have been proposed for disease recognition tasks. However, most existing methods focus on recognizing diseases at one point with a single feature set, which becomes challenging as the number of activities and features increases.

This research aims to evaluate and compare the performance of five machine learning classifiers—Decision Tree (DT), Random Forest Classifier (RFC), XGBoost, Gradient Boosting Classifier (GBC), and Light Gradient-Boosting Machine (LGBM)—in disease classification, focusing on their accuracy across different feature sets. The study investigates the importance of features to identify which variables are most critical for disease detection. Boruta and Recursive Feature Elimination (RFE) are employed for feature selection to enhance classification accuracy and efficiency. This approach improves model performance and provides valuable insights into the key factors influencing disease classification, contributing to the development of more interpretable and reliable healthcare models.

While numerous studies have explored machine learning techniques for disease detection, few have focused on detecting multiple diseases simultaneously. For instance, EHRs have been used to classify four different blood diseases, with models such as LogitBoost, Random Forest, XGBoost, Decision Tree, and Support Vector Machine being tested. The reported accuracy of these models varied, with the highest reaching 98% [6]. Another study by Arumugam et al. [7] explored disease detection for heart disease and diabetes using three distinct datasets for each condition. Among the tested models, Decision Tree outperformed the others with an accuracy of around 90%, surpassing Support Vector Machine (87%) and Naive Bayes (77%).

2. Methods

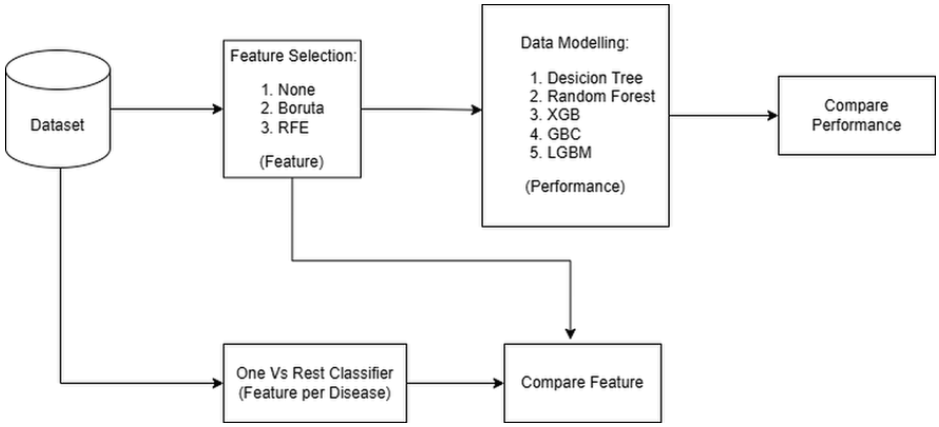

Figure 1. The Workflow of This Research

Figure 1 illustrates the workflow of this research. The process begins with the dataset, which serves as the foundation for the entire analysis. Initially, feature selection is conducted using three different approaches: no feature selection (utilizing all available features), Boruta (an advanced feature selection technique), and RFE (Recursive Feature Elimination). These methods aim to identify and retain the most relevant features while eliminating less important or redundant ones. After feature selection, the dataset trains several machine learning models, including Decision Tree, Random Forest, XGBoost, Gradient Boosting Classifier (GBC), and LightGBM. These models are evaluated using various performance metrics such as accuracy, precision, recall, and others, to determine the most effective classifier for the task. Subsequently, a One-vs-Rest (OvR) classification approach is employed to analyze each disease class individually. The features selected by Boruta and RFE are assessed for their relevance and impact on classification performance for each disease (Anemia, Diabetes, Heart Disease, Healthy, Thalassemia, and Thrombocytopenia). This step allows for a comparison of the effectiveness of the feature selection methods for each specific disease. Finally, the features identified by Boruta and RFE are juxtaposed with those identified by the One-vs-Rest classifier to provide a comprehensive understanding of how the different feature selection techniques contribute to the classification process and model performance, ensuring that the results are applicable to the specific diseases being examined.

2.1. Random Forest

Random Forests (RF) is an approach that integrates multiple decision trees to improve the performance of a single decision tree classifier. This is accomplished by employing the bootstrap aggregating (bagging) method and adding randomness when selecting data nodes for partitioning during the decision tree construction. An RF classifier combines several independent decision tree classifiers [8]. A decision tree with M leaves splits the feature space into M regions, denoted as R_m, where $1 \leq m \leq M$. For each tree, the prediction function $f(x)$ is defined as shown in Equation (1):

$$\mathrm{f(x)} = \sum_{m=1}^{M} c_m \, \pi(x, \ R_m) \tag{1}$$

In this formulation, M represents the number of regions in the feature space, R_m is the region associated with index m, c_m is the constant corresponding to m, and 1 is an indicator function, defined in Equation (2):

$$\pi(\mathrm{x}, R_m) = \begin{cases} 1, & if \ x \in R_m \\ 0, & otherwise \end{cases} \tag{2}$$

The ultimate classification decision is determined by the majority vote of all the trees.

2.2. Gradient Boosting

Gradient Boosting Classifier shows that it can be a decent model with 97% accuracy on the heart disease dataset and 73% on the cardiovascular dataset [9]. Gradient Boosting is a powerful supervised classification technique that constructs a strong predictive model by combining multiple weak learners [10]. The core principle involves iteratively adding new weak models, which are typically simple models that perform marginally better than random guessing, to the ensemble [11]. Each new model is trained to correct the residual errors made by the previous models, using gradient descent optimization to minimize these errors. Gradient Boosting is commonly utilized with tree-based models like decision trees or random forests. Many hyperparameters, such as the number of trees, the learning rate, and the maximum depth of the trees impact the model's efficacy. Recent applications of the Gradient Boosting Classifier demonstrate its robustness, achieving a 97% accuracy on a heart disease dataset and a 73% accuracy on a cardiovascular dataset [9]. These results underscore Gradient Boosting's capability to deliver high performance across diverse datasets, making it a versatile and reliable tool in machine learning.

XGBoost, an advanced extension of Gradient Boosting, is widely recognized for its superior performance in machine learning tasks. It improves upon traditional Gradient Boosting through features such as advanced tree-based models, regularization techniques to prevent overfitting, robust handling of missing values, and parallel computing for efficient training on large datasets, alongside customizable hyperparameters and support for multiple objectives [12]. Recent research highlights XGBoost's exceptional performance compared to other models; for example, one study demonstrated that XGBoost achieved an accuracy of 91%, surpassing alternative approaches [13], while another found that XGBoost performed well with a notable accuracy of 99.16% [14]. These advancements underscore XGBoost's effectiveness as a leading technique for achieving high accuracy in diverse machine-learning applications.

LightGBM is a high-performance Gradient Boosting framework that significantly enhances traditional Gradient Boosting techniques through several innovative methods [15]. LGBM consists of individual shallow decision trees that avoid overfitting problems [16][17]. LightGBM employs Gradient-Based One-Side Sampling and Exclusive Feature Bundling to reduce computational complexity and improve efficiency. Additionally, its use of Leaf-wise Tree Growth accelerates model training and enhances accuracy by focusing on the most promising splits. LightGBM supports parallel and distributed learning, which allows it to handle large-scale data processing efficiently while managing multiple objectives and optimizing memory usage

[18]. Swainn et al. [19] demonstrated the effectiveness of LightGBM in classifying Parkinson's disease, achieving an impressive accuracy of 98%. This research highlights LightGBM's capability to deliver high accuracy and speed in complex machine-learning tasks, confirming its value as a leading tool for advanced data analysis and model development. Another research by Sharma and Singh [20] shows that LGBM can reach 99% accuracy

2.3. Feature Selection Method

Boruta is a feature selection algorithm built around the random forest classification technique. It works by creating shadow features, which are randomly permuted versions of the original features, and then training a random forest classifier on the combined dataset [21]. The algorithm iteratively checks the importance of each feature by comparing it to its shadow features. Features with higher importance scores than their shadow features are deemed relevant and are retained, while those with lower scores are considered irrelevant and are removed [9][22].

Recursive Feature Elimination (RFE) is a wrapper method that recursively eliminates features and builds a model over the remaining ones [23]. It ranks features based on importance and eliminates the least important ones until the desired number of features is reached [24]. RFE is an iterative process that involves training a model on all features, ranking them, eliminating the least important feature, and repeating the process until the desired number of features is achieved. Specifically, the RFE algorithm [18] operates as follows: fit the model using all independent variables, calculate the variable importance of all variables, rank each independent variable based on its importance to the model, drop the weakest variable, and build a model using the remaining variables, then calculate model accuracy. This process is repeated until all variables have been used and ranked according to when they were dropped. For classification, accuracy and precision are used as metrics. Guyon et al. [25] introduced Recursive Feature Elimination (RFE), a method applied in cancer classification using Support Vector Machine (SVM). RFE starts by using all available features to train an SVM model. It then ranks the importance of each feature based on its contribution to the model, generating a ranked list. Irrelevant features contribute less to the model's performance and are eliminated iteratively until the desired number of features remains [26].

2.4. Dataset

The multiple disease prediction data set is publicly available on Kaggle [27]. This dataset comprises 25 features and 2837 entries. It is designed to assess the health status of individuals, determining whether a person has a specific disease or is healthy based on blood samples and various health parameters. These features are presented in Table 1, and the disease class is shown in Table 2.

Table 1. Multiple Disease Prediction Dataset Features

No	Features	No	Features
1	Glucose	14	Diastolic Blood Pressure (DBP)
2	Cholesterol	15	Triglycerides
3	Hemoglobin	16	HbA1c (Glycated Hemoglobin)
4	Platelets	17	LDL (Low-Density Lipoprotein) Cholesterol
5	White Blood Cells (WBC)	18	HDL (High-Density Lipoprotein) Cholesterol
6	Red Blood Cells (RBC)	19	ALT (Alanine Aminotransferase)
7	Hematocrit	20	AST (Aspartate Aminotransferase)
8	Mean Corpuscular Volume (MCV)	21	Heart Rate
9	Mean Corpuscular Hemoglobin (MCH)	22	Creatinine
10	Mean Corpuscular Hemoglobin Concentration (MCHC)	23	Troponin
11	Insulin	24	C-reactive Protein (CRP)
12	BMI (Body Mass Index)	25	Disease
13	Systolic Blood Pressure (SBP)		

Table 2. Disease Class

No	Disease
1	Cholesterol
2	Hemoglobin
3	Platelets
4	White Blood Cells (WBC)
5	Red Blood Cells (RBC)
6	Hematocrit

2.5. Environment

This research used Jupyter Notebook as the primary development environment for building and evaluating classification models. The dataset used in this study comprises 2,837 records, each corresponding to one of six health conditions: Anemia, Diabetes, Heart Disease, Healthy, Thalassemia, and Thrombocytopenia. These conditions were selected due to their clinical significance and the need for accurate, early detection methods. Table 2 provides a detailed breakdown of the dataset distribution.

To evaluate classifier performance, this study utilizes four key metrics: accuracy, precision, recall, and F1-score, which are essential for validating predictive models in healthcare applications. Accuracy, as shown in Equation (3), measures the overall correctness of predictions, while precision, as shown in Equation (4), evaluates the proportion of true positive predictions among all predicted positives, thereby minimizing the impact of false positives. Recall, as shown in Equation (5), measures the model's ability to correctly identify all relevant positive instances. The F1-score, as shown in Equation (6), provides a harmonic mean between precision and recall, particularly beneficial when dealing with imbalanced datasets. These metrics collectively ensure a comprehensive and balanced assessment of model performance [28].

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{3}$$

$$Precision = \frac{TP}{TP+FP} \tag{4}$$

$$\text{Recall} = \frac{TP}{TP+FN} \tag{5}$$

$$\text{F1} = \frac{2\times P\times R}{P+R} \tag{6}$$

where: TP = True positive; FP = False positive; TN = True negative; FN = False negative; P = Precision; R = Recall.

2.6. K-Fold Cross-Validation

To assess the model's accuracy more objectively and to mitigate bias caused by imbalanced data splitting, this study employs the K-Fold Cross-Validation technique with Stratified K-Fold. This method divides the dataset into K equal-sized parts (folds). Each fold is used once as a test set, while the remaining K-1 folds are used for training. This process is repeated for each fold, and the final evaluation results are averaged across all folds based on performance metrics such as accuracy, precision, and recall. This method helps provide a more reliable estimate of the model's performance by ensuring that every data point is used for training and testing.

The use of Stratified K-Fold ensures that the class distribution within each fold is representative of the overall distribution of the classes in the dataset. This is especially important in imbalanced datasets, where certain classes may be underrepresented. By maintaining similar class distributions across all folds, Stratified K-Fold allows the model to learn from all classes proportionally, reducing the risk of the model being biased toward the majority class. This is particularly beneficial in healthcare datasets, where imbalanced class distributions are common and ensures a more accurate model evaluation [29].

In this study, the Random Forest and XGBoost models are selected due to their ability to handle complex, high-dimensional data and their robustness against overfitting, especially when used with cross-validation. Both models are trained using the features selected through Boruta

and Recursive Feature Elimination (RFE) methods, and their performance is evaluated based on the accuracy, precision, recall, and other relevant metrics. These models are particularly well-suited for this task due to their capacity for learning non-linear relationships and handling missing or noisy data.

By employing K-Fold Cross-Validation with Stratified K-Fold, the study ensures a thorough evaluation of the model's performance, considering the potential for bias in imbalanced datasets. This approach enhances the generalizability and reliability of the findings, ensuring that the models' performance metrics are not overly optimistic or influenced by an unrepresentative data split.

2.7. One Vs Rest Classifier

The One-vs-Rest (OvR) classifier is a multi-class classification approach where a separate binary classifier is trained for each class in the dataset. For each classifier, one class is treated as the positive class, while all other classes are combined into a single negative class. This method enables the model to focus on distinguishing one class from the rest. In the context of disease classification, OvR helps to evaluate the relevance of specific features for each disease condition separately, allowing for a more detailed and accurate analysis of how well each feature contributes to the classification of different diseases [30].

2.8. Parameter Setting

To evaluate the performance of various classification models, experiments were conducted using 70% of the dataset for training and the remaining 30% for testing. StratifiedKFold (with n_splits=10, shuffle=True, random_state=42) was employed to ensure a balanced distribution of the target variable in each fold, particularly important for datasets with imbalanced classes. Each machine learning model was configured with specific parameters for consistency and reproducibility: the DecisionTreeClassifier, RandomForestClassifier, and GradientBoostingClassifier all used a fixed random_state value of 42, ensuring that results could be consistently reproduced across multiple runs. The XGBClassifier was configured with use_label_encoder=False to prevent unnecessary warning messages and eval_metric='mlogloss' to optimize the model for multi-class classification by minimizing logarithmic loss.

Boruta was used with n_estimators='auto' for feature selection, dynamically adjusting the number of trees based on the dataset's complexity, ensuring an adaptive feature selection process. Additionally, Recursive Feature Elimination (RFE) was employed with RandomForestClassifier as the estimator, selecting the top 10 features based on their importance. The choice to retain 10 features was made to balance sufficient feature retention while minimizing the risk of overfitting, which also helped simplify the model for better interpretability and computational efficiency. This combination of Boruta and RFE provided a robust, efficient, and reproducible approach to disease classification, enhancing model performance and interpretability.

3. Result

The classification performance of five machine learning models—Decision Tree (DT), Random Forest (RFC), XGBoost (XGB), Gradient Boosting Classifier (GBC), and LightGBM (LGBM)—was evaluated across three configurations: using all features, using features selected by Boruta, and using features selected by Recursive Feature Elimination (RFE).

Table 3. Accuracy and Precision of 25 Feature Selection

Model	Mean			
	Accuracy	Precision	Recall	F1-Score
DT	95.23%	95.29%	95.23%	95.19%
RFC	97.83%	98.00%	97.83%	97.85%
XGB	99.43%	99.44%	99.43%	99.43%
GBC	99.37%	99.39%	99.37%	99.37%
LGBM	99.66%	99.66%	99.66%	99.66%

Table 3 displays model performance using all 24 features. LightGBM achieved the highest mean scores across all metrics (accuracy, precision, recall, and F1-score at 99.66%), followed closely by XGBoost (99.43%) and GBC (99.37%). Decision Tree was the lowest-performing model in this setting but maintained above 95% in all metrics.

Table 4. Accuracy And Precision with Boruta Feature Selection

Model	Mean			
	Accuracy	Precision	Recall	F1-Score
DT	92.12%	92.19%	92.12%	92.02%
RFC	97.12%	97.20%	97.12%	97.11%
XGB	99.34%	99.35%	99.34%	99.34%
GBC	99.43%	99.45%	99.43%	99.43%
LGBM	99.63%	99.63%	99.63%	99.63%

Boruta selected all 24 features, indicating that each feature contributed meaningfully to the classification task. As shown in Table 4, the performance of the models with Boruta-selected features remained comparable to the full-feature configuration, with only minor variations. For instance, LightGBM again achieved 99.63% across all metrics, nearly identical to the full-feature result, while Decision Tree saw a slight decline to 92.12% accuracy. These findings suggest that the dataset contains no redundant features, and Boruta's all-inclusive selection validates the global relevance of each feature.

Table 5. Accuracy And Precision with RFE Feature Selection

Model	Mean			
	Accuracy	Precision	Recall	F1-Score
DT	91.58%	91.73%	91.58%	91.48%
RFC	96.26%	96.38%	96.26%	96.26%
XGB	97.35%	97.39%	97.35%	97.33%
GBC	97.32%	97.35%	97.32%	97.30%
LGBM	97.35%	97.38%	97.35%	97.33%

Unlike Boruta, RFE selected only ten features to optimize model simplicity while maintaining predictive power. Table 5 shows that this reduction led to a slight drop in performance, particularly in simpler models such as DT and RFC. However, more robust models such as XGB, GBC, and LGBM still achieved over 97% accuracy. This demonstrates that a smaller subset of features can still perform well, although with slightly reduced accuracy compared to the full set or Boruta-selected features.

Table 6. Importance Feature: Each Disease with OvR

Class	Top 5 Features (Importance)
Anemia	Red Blood Cells (0.2575), White Blood Cells (0.2449), Hematocrit (0.1528), C-reactive Protein (0.0892), Insulin (0.0837)
Diabetes	Insulin (0.2451), BMI (0.2350), Cholesterol (0.2085), HbA1c (0.1018), Glucose (0.0995)
Healthy	Heart Rate (0.2831), Platelets (0.1695), Mean Corpuscular Volume (0.1324), HDL Cholesterol (0.1305), Mean Corpuscular Hemoglobin (0.1011)
Heart Disease	Systolic Blood Pressure (0.1478), Heart Rate (0.1337), Insulin (0.1255), Mean Corpuscular Volume (0.1233), Mean Corpuscular Hemoglobin Concentration (0.1066)
Thalassemia	Mean Corpuscular Hemoglobin (0.3406), White Blood Cells (0.1825), Red Blood Cells (0.1505), Diastolic Blood Pressure (0.0704), Mean Corpuscular Hemoglobin Concentration (0.0699)
Thrombocytopenia	Heart Rate (0.3390), Platelets (0.2523), LDL Cholesterol (0.1922), White Blood Cells (0.0850), HDL Cholesterol (0.0509)

To ensure the model's relevance to each disease class, a One-vs-Rest (OvR) analysis was conducted. Table 6 presents the top five most important features for each class. For instance, anemia classification was most influenced by red and white blood cell counts as well as hematocrit levels, while insulin, BMI, and cholesterol were highly predictive for diabetes. In the case of

thalassemia, the most critical features included mean corpuscular hemoglobin, white and red blood cells, and diastolic blood pressure.

When comparing these results with the features selected by Boruta, as shown in Figure 2 and RFE, as shown in Figure 3, it can be observed that several important features identified in the OvR analysis were retained. For example, Red Blood Cells, White Blood Cells, Hematocrit, and Insulin were selected by both Boruta and RFE, aligning with their high importance in anemia and other diseases. Similarly, BMI, Cholesterol, and HbA1c, which were relevant for diabetes, were also selected by Boruta. However, some notable features such as Heart Rate, which was important for multiple classes (Healthy, Heart Disease, and Thrombocytopenia), and Diastolic Blood Pressure, which was important for Thalassemia were not selected by RFE.

This comparison highlights that while both Boruta and RFE were able to preserve many clinically relevant features, certain disease-specific indicators were excluded, potentially due to their lower global ranking across all classes. Nevertheless, the overall overlap suggests that the feature selection methods retained key discriminative variables for most classes while effectively reducing dimensionality and potential noise.

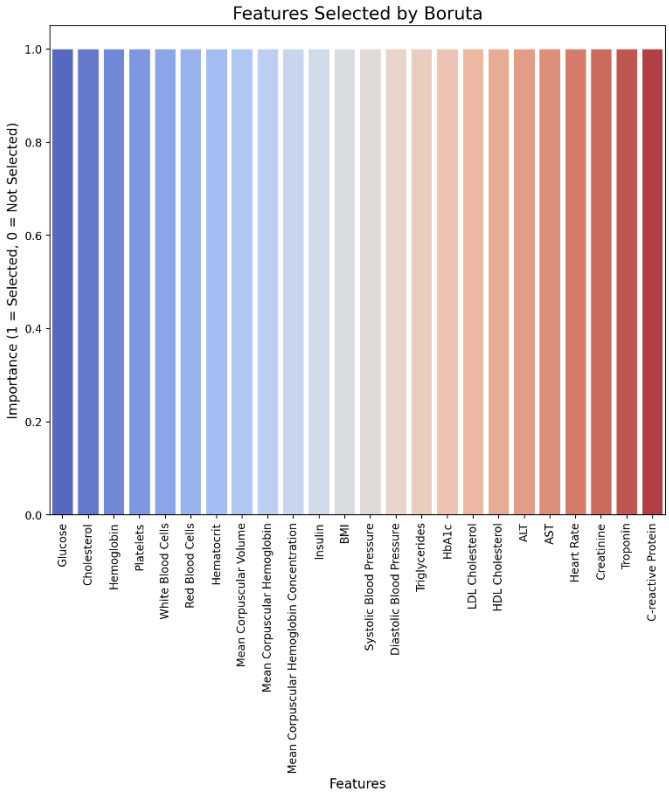

Figure 2. Feature Selection with the Boruta Method

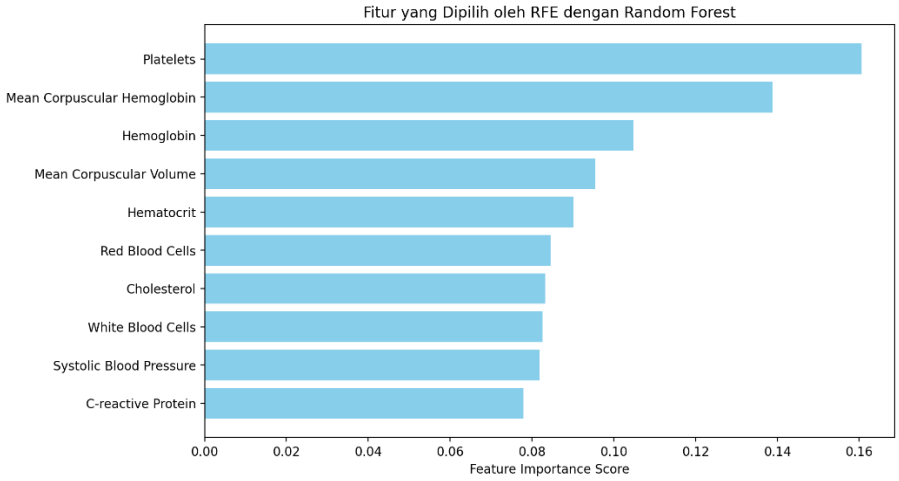

Figure 3. Feature Selection with RFE Method

4. Discussion

The experimental results indicate that feature selection techniques can noticeably impact classification performance, with different models exhibiting varying levels of sensitivity to feature set changes. The LGBM classifier consistently demonstrated the highest performance, regardless of feature selection method. However, the Decision Tree (DT) model showed a notable decline in accuracy and F1-score when Boruta and RFE were applied. This finding underscores that simpler models, like DT, are more sensitive to feature space changes. In contrast, more complex, ensemble-based models like LGBM, XGBoost, and GBC exhibit greater robustness in the presence of feature selection.

Interestingly, the Boruta method selected all 24 features as important, indicating that the full feature set holds valuable information for disease classification. In contrast, RFE identified only 10 features, concentrating on dimensionality reduction. Although this reduced feature set resulted in slightly lower performance scores, it still maintained competitive accuracy, particularly for LGBM, which achieved an accuracy above 97%. This underscores the trade-off between interpretability and performance in medical diagnostics—RFE simplifies complexity and can yield more interpretable models, but it may sacrifice some predictive power.

Further analysis using the One-vs-Rest (OvR) approach provided insights into the disease-specific features that drive classification accuracy. For example, anemia was strongly associated with red blood cells, white blood cells, and hematocrit, while diabetes was heavily influenced by features such as insulin, BMI, and cholesterol. Thrombocytopenia, on the other hand, showed a strong relationship with platelets and heart rate. These results confirm that the models capture disease-specific patterns and are not merely reliant on global trends, making them more interpretable and applicable to specific medical conditions.

One important implication of this study is the potential for more efficient disease detection by utilizing fewer features without sacrificing accuracy. Feature selection methods such as RFE can significantly reduce computational costs, time, and resources, making them especially valuable in time-sensitive medical applications where data acquisition may be costly or limited. Medical practitioners can achieve more efficient diagnostics by focusing on a more targeted set of features, ultimately accelerating clinical decision-making.

The strength of this study lies in its comprehensive evaluation of multiple machine learning models, accompanied by two feature selection techniques—Boruta and RFE. The findings suggest that high accuracy can be achieved even with a reduced feature set, making the models suitable for real-world medical datasets. Using both Boruta and RFE allows for a nuanced assessment of feature selection, offering insights into how each approach influences classification performance.

However, certain limitations must be taken into account. While Boruta's method of selecting all features as important may help capture a broader range of information, it could also lead to overfitting, especially in more complex datasets. Conversely, RFE, which aggressively reduces the feature set, may overlook potentially valuable features contributing to disease-specific classifications' accuracy. Additionally, this study only evaluates a limited number of machine learning models, and further research into additional algorithms could provide a more comprehensive understanding of the effects of feature selection on classification performance.

5. Conclusions

This study demonstrates the significant role of feature selection techniques in enhancing multi-disease classification performance using machine learning models. Among the evaluated classifiers, the Light Gradient Boosting Machine (LGBM) consistently achieved the highest accuracy, showing strong resilience to changes in the feature set. The comparison between Boruta and Recursive Feature Elimination (RFE) revealed a trade-off between model interpretability and predictive performance, with RFE offering a more compact and interpretable feature subset while maintaining high accuracy.

Furthermore, class-wise analysis through the One-vs-Rest (OvR) approach highlighted the disease-specific relevance of features, confirming the models' ability to capture meaningful medical patterns rather than relying solely on global trends. These findings suggest that efficient and accurate medical diagnosis is possible even with reduced feature sets, which is particularly valuable in clinical environments with limited data resources.

Overall, this work supports the integration of feature selection in medical machine learning pipelines to improve both performance and interpretability. Future research could apply other feature selection techniques, such as Mutual Information or Genetic Algorithms, to compare their effectiveness in multi-disease classification tasks. Exploring alternative ensemble models, like Stacking or Voting classifiers, could also enhance performance. Additionally, evaluating these methods on diverse datasets, including those with missing or imbalanced data, would help assess the models' robustness and generalizability. Finally, integrating domain-specific medical knowledge with feature selection methods could further improve both the models' interpretability and performance, offering a more personalized approach to disease classification.

References

[1] C.-H. Hsu *et al.*, "Effective multiple cancer disease diagnosis frameworks for improved healthcare using machine learning," *Measurement*, vol. 175, p. 109145, 2021, doi: https://doi.org/10.1016/j.measurement.2021.109145.

[2] M. M. Ahsan and Z. Siddique, "Machine learning-based heart disease diagnosis: A systematic literature review," *Artificial Intelligent in Medicine*, vol. 128, no. C, Jun. 2022, doi: 10.1016/j.artmed.2022.102289.

[3] V. V. Vijayan and C. Anjali, "Prediction and diagnosis of diabetes mellitus — A machine learning approach," in *2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS)*, 2015, pp. 122–127. doi: 10.1109/RAICS.2015.7488400.

[4] P. Zhang and M. N. Kamel Boulos, "Generative AI in Medicine and Healthcare: Promises, Opportunities and Challenges," *Future Internet*, vol. 15, no. 9, 2023, doi: 10.3390/fi15090286.

[5] A. Rajkomar *et al.*, "Scalable and accurate deep learning with electronic health records," *NPJ Digit Med*, vol. 1, no. 1, p. 18, 2018, doi: 10.1038/s41746-018-0029-1.

[6] F. K. Alsheref and W. H. Gomaa, "Blood Diseases Detection using Classical Machine Learning Algorithms," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 7, 2019, doi: 10.14569/IJACSA.2019.0100712.

[7] K. Arumugam, M. Naved, P. P. Shinde, O. Leiva-Chauca, A. Huaman-Osorio, and T. Gonzales-Yanac, "Multiple disease prediction using Machine learning algorithms," *Materials Today: Proceedings*, vol. 80, pp. 3682–3685, 2023, doi: https://doi.org/10.1016/j.matpr.2021.07.361.

[8] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[9] M. Kursa, A. Jankowski, and W. Rudnicki, "Boruta - A System for Feature Selection," *Fundamenta Informaticae*, vol. 101, pp. 271–285, Jan. 2010, doi: 10.3233/FI-2010-288.

[10] A. Natekin and A. Knoll, "Gradient boosting machines, a tutorial," *Front Neurorobot*, vol. Volume 7-2013, 2013, [Online]. Available: https://www.frontiersin.org/journals/neurorobotics/articles/10.3389/fnbot.2013.00021

[11] S. Zhou, S. Wang, Q. Wu, R. Azim, and W. Li, "Predicting potential miRNA-disease associations by combining gradient boosting decision tree with logistic regression," *Computational Biology and Chemistry*, vol. 85, p. 107200, 2020, doi: https://doi.org/10.1016/j.compbiolchem.2020.107200.

[12] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, in KDD '16. ACM, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[13] K. Budholiya, S. K. Shrivastava, and V. Sharma, "An optimized XGBoost based diagnostic system for effective prediction of heart disease," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 7, pp. 4514–4523, 2022, doi: https://doi.org/10.1016/j.jksuci.2020.10.013.

[14] Md. J. Raihan, Md. A.-M. Khan, S.-H. Kee, and A.-A. Nahid, "Detection of the chronic kidney disease using XGBoost classifier and explaining the influence of the attributes on the model using SHAP," *Scientific Reports*, vol. 13, no. 1, p. 6263, 2023, doi: 10.1038/s41598-023-33525-0.

[15] Ibrahim Karabayir *et al.*, "Predicting Parkinson's Disease and Its Pathology via Simple Clinical Variables," *Journal of Parkinsons Disease*, vol. 12, no. 1, pp. 341–351, Sep. 2021, doi: 10.3233/JPD-212876.

[16] G. Ke *et al.*, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," in *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf

[17] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *The Annals of Statistics*, pp. 1189–1232, 2001.

[18] C. Dewi and R.-C. Chen, "Human Activity Recognition Based on Evolution of Features Selection and Random Forest," in *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*, 2019, pp. 2496–2501. doi: 10.1109/SMC.2019.8913868.

[19] B. K. Swain, S. Mohapatra, M. Mishra, and R. Sharma, "A unified approach for Parkinson's disease recognition: imbalance mitigation and grid search optimized boosting with LightGBM," *Medical & Biological Engineering & Computing*, vol. 62, no. 11, pp. 3471–3491, 2024, doi: 10.1007/s11517-024-03139-3.

[20] A. Sharma and B. Singh, "AE-LGBM: Sequence-based novel approach to detect interacting protein pairs via ensemble of autoencoder and LightGBM," *Computers in Biology and Medicine*, vol. 125, p. 103964, 2020, doi: https://doi.org/10.1016/j.compbiomed.2020.103964.

[21] R.-C. Chen, C. Dewi, S.-W. Huang, and R. E. Caraka, "Selecting critical features for data classification based on machine learning methods," *Journal of Big Data*, vol. 7, no. 1, p. 52, 2020, doi: 10.1186/s40537-020-00327-4.

[22] G. Manikandan, B. Pragadeesh, V. Manojkumar, A. L. Karthikeyan, R. Manikandan, and A. H. Gandomi, "Classification models combined with Boruta feature selection for heart disease prediction," *Informatics in Medicine Unlocked*, vol. 44, p. 101442, 2024, doi: https://doi.org/10.1016/j.imu.2023.101442.

[23] B. F. Darst, K. C. Malecki, and C. D. Engelman, "Using recursive feature elimination in random forest to account for correlated variables in high dimensional data," *BMC Genetics*, vol. 19, no. 1, p. 65, 2018, doi: 10.1186/s12863-018-0633-8.

[24] E. J. Michaud, Z. Liu, and M. Tegmark, "Precision Machine Learning," *Entropy*, vol. 25, no. 1, 2023, doi: 10.3390/e25010175.

[25] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene Selection for Cancer Classification using Support Vector Machines," *Machine Learning*, vol. 46, no. 1, pp. 389–422, 2002, doi: 10.1023/A:1012487302797.

[26] R. Y. Krisnabayu, A. Ridok, and A. Setia Budi, "Hepatitis Detection using Random Forest based on SVM-RFE (Recursive Feature Elimination) Feature Selection and SMOTE," in *Proceedings of the 6th International Conference on Sustainable Information Engineering and Technology*, in SIET '21. New York, NY, USA: Association for Computing Machinery, 2021, pp. 151–156. doi: 10.1145/3479645.3479668.

[27] E. Aboelnaga, "Multiple disease prediction Dataset," *Kaggle,* 2013, [Online]. Available: https://www.kaggle.com/datasets/ehababoelnaga/multiple-disease-prediction/data.

[28] F. Y. Osisanwo, J. E. T. Akinsola, O. Awodele, J. O. Hinmikaiye, O. Olakanmi, and J. Akinjobi, "Supervised Machine Learning Algorithms: Classification and Comparison," *International Journal of Computer Trends and Technology (IJCTT)*, vol. 48, no. 3, pp. 128–138, Jun. 2017, [Online]. Available: http://www.ijcttjournal.org

[29] M. T R, V. K. V, D. K. V, O. Geman, M. Margala, and M. Guduri, "The stratified K-folds cross-validation and class-balancing methods with high-performance ensemble classifiers for breast cancer classification," *Healthcare Analytics*, vol. 4, p. 100247, 2023, doi: https://doi.org/10.1016/j.health.2023.100247.

[30] J. Xu, "An extended one-versus-rest support vector machine for multi-label classification," *Neurocomputing*, vol. 74, no. 17, pp. 3114–3124, 2011, doi: https://doi.org/10.1016/j.neucom.2011.04.024.