Available online at website: https://jurnal.iaii.or.id/index.php/RESTI

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 10 No. 2 (2026) 465 - 475 e-ISSN: 2580-0760

A Hybrid Intersection Filtering and Recursive Feature Elimination Technique for Efficient Feature Reduction in High Dimensional Datasets

Akhmad Dahlan[1], Yoga Pristyanto[2], Anggit Ferdita Nugraha[3], Rifda Faticha Alfa Aziza[4], Ibnu Hadi Purwanto[5]
[1]Department of Informatics Management, Faculty of Computer Science, Universitas Amikom Yogyakarta, Yogyakarta, Indonesia
[2]Department of Information System, Faculty of Computer Science, Universitas Amikom Yogyakarta, Yogyakarta, Indonesia
[3]Department of Computer Engineering, Faculty of Computer Science, Universitas Amikom Yogyakarta, Yogyakarta, Indonesia
[4]Department of Informatics, Faculty of Computer Science, Universitas Amikom Yogyakarta, Yogyakarta, Indonesia
[5]Department of Information Technology, Faculty of Computer Science, Universitas Amikom Yogyakarta, Yogyakarta, Indonesia

*alland@amikom.ac.id

Abstract

High-dimensional datasets are commonly encountered in real-world machine learning applications and often degrade classification performance due to redundant and irrelevant features. In addition, the presence of excessive features increases computational complexity and processing time. Feature selection is therefore a crucial preprocessing step to improve model accuracy and efficiency. This study proposes a hybrid feature selection approach called Intersection Filtering based on Recursive Feature Elimination with Cross-Validation (IF-RFECV), which integrates wrapper-based and filter-based strategies to obtain a stable and optimal subset of features. The proposed method first applies Recursive Feature Elimination with Cross-Validation (RFECV) using multiple classification models to rank and select relevant features. Subsequently, an intersection filtering mechanism is employed to identify features that are consistently selected across different RFECV-based models, thereby reducing model-dependent bias and improving feature robustness. The effectiveness of IF-RFECV is evaluated using four benchmark datasets with varying dimensionality obtained from the KEEL and UCI repositories. Several classification algorithms, including Gradient Boosting, K-Nearest Neighbor, Naïve Bayes, Decision Tree, Random Forest, and Support Vector Machine, are used to assess model performance. Experimental results demonstrate that IF-RFECV produces a more compact feature subset compared to conventional RFECV while achieving superior performance in terms of accuracy, precision, recall, and F1-score on most datasets, particularly those with higher dimensionality. Although IF-RFECV requires slightly higher computational time due to its two-stage process, the performance gains and improved generalization justify this trade-off. These findings indicate that IF-RFECV is an effective and robust feature selection technique for high-dimensional classification problems.

Keywords: classification; feature selection; intersection filter; machine learning; RFECV

How to Cite: A. Dahlan, Y. Pristyanto, A. F. Nugraha, R. F. A. Aziza, and I. H. Purwanto, "A Hybrid Intersection Filtering and Recursive Feature Elimination Technique for Efficient Feature Reduction in High Dimensional Datasets", *J. RESTI (Rekayasa Sist. Teknol. Inf.)* , vol. 10, no. 2, pp. 465 - 475, Apr. 2026.
Permalink/DOI: https://doi.org/10.29207/resti.v10i2.7396

Received: December 18, 2025
Accepted: April 12, 2026
Available Online: April 26, 2026

This is an open-access article under the CC BY 4.0 License
Published by Ikatan Ahli Informatika Indonesia

1. Introduction

The existence of high-dimensional datasets is one of the problems that are often encountered in machine learning implementation processes [1-2]. A high-dimensional dataset refers to a condition in which the dataset used for training and testing contains a large number of features or attributes [3-5]. In general, such datasets frequently appear in real-world problems [6]. This condition can significantly affect the performance of classification models, as each feature plays an important role in shaping the quality of the resulting model [7]. When a dataset contains many features that are irrelevant or weakly related to the target class, the classification performance tends to degrade [6], [8-9]. In addition, the computational cost and processing time required for model training and evaluation increase substantially [10-12]. Therefore, feature selection is an

essential preprocessing step to reduce dimensionality and improve both model performance and efficiency.

Feature selection techniques are commonly categorized into filtering and wrapping approaches [13]. Filter-based methods evaluate features using statistical criteria independent of any learning algorithm, while wrapper-based methods assess feature subsets based on the performance of a specific classifier. Several filtering techniques, such as correlation criteria and mutual information, have been widely used due to their simplicity and low computational cost. On the other hand, Recursive Feature Elimination (RFE) is one of the most popular wrapper-based feature selection methods and has demonstrated promising results in various classification tasks [14].

Several previous studies have explored the use of RFE for feature selection. Research in [15] applied RFE as a wrapper technique and showed that feature elimination does not always lead to continuous performance improvement across different datasets and models [16]. A comparative study in [13] demonstrated that RFE outperformed Relief filtering, indicating that recursive elimination techniques are more effective in identifying relevant and discriminative features. Other studies reported that RFE can improve classification accuracy by removing less informative features [17]. However, further investigations revealed that RFE combined with cross-validation (RFECV) does not consistently enhance performance for all classifiers [18]. Comparisons between RFE and Principal Component Analysis (PCA) in [18-19] showed that RFE generally yields better results, as it preserves feature interpretability and relevance rather than transforming the feature space.

In addition to wrapper-based methods, several studies have examined filter-based feature selection techniques. Research in [20] compared Information Gain and Pearson Correlation Coefficient (PCC), concluding that Information Gain performs better because it can capture non-linear relationships between features and class labels. Similarly, [21] evaluated Correlation-Based Feature Selection, Information Gain, and Gini Ratio, showing that each method improves performance for certain classifiers but not universally. Further studies in [22] confirmed that filter-based methods can reduce computational load and enhance performance for specific models, although their effectiveness remains dataset- and model-dependent.

Based on these findings, it can be observed that no single feature selection technique consistently improves classification performance across different datasets and classifiers. Most existing approaches rely on a single filtering or wrapping strategy, or apply RFECV using a single learning model, which often leads to model-dependent and unstable feature subsets. To address this limitation, this study proposes a hybrid feature selection approach that combines wrapping and filtering techniques.

In this research, the feature selection process begins with Recursive Feature Elimination with Cross-Validation (RFECV) as the wrapping stage. The resulting feature subsets obtained from multiple classifiers are then refined using an intersection-based filtering mechanism, referred to as Intersection Filtering based on Recursive Feature Elimination with Cross-Validation (IF-RFECV). The main objective of this approach is to identify features that are consistently selected across different models, thereby reducing model bias and improving feature stability. The performance of the proposed IF-RFECV method is evaluated and compared with previous approaches [13], [19] to assess its effectiveness. This study makes three main contributions: (1) the proposed IF-RFECV method demonstrates better classification performance compared to conventional RFECV, (2) IF-RFECV provides an effective solution for selecting relevant features in high-dimensional datasets, and (3) the proposed approach can serve as a reference for future research on feature selection and high-dimensional data analysis.

2. Methods

The stages carried out in this paper can be seen in Figure 1. There are five main stages, including (1) Data Acquisition, (2) Preprocessing Data, (3) Features Selection, (4) Classification Model and (5) Evaluation.

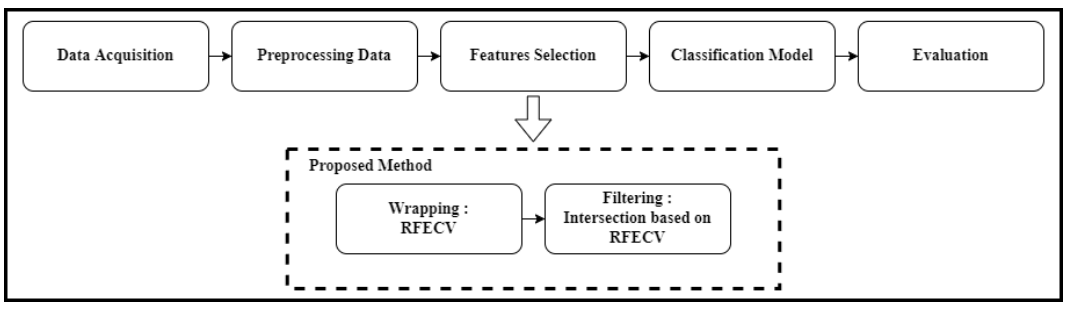


Figure 1. Research Workflow

2.1 Data Acquisition

The dataset used in this study is a public dataset obtained from KEEL Repository and UCI Machine Learning Datasets Repository. In this study, four types of datasets were used with varying features. Table 1 is information from the datasets used in this study.

Table 1. Performance Metrics

Datasets	Code	Num of Instances	Num of Features	Num of Classes
Mutagenesis-Bonds	D1	3995	17	2
Mutagenesis-Chains	D2	5349	25	2
WDBC	D3	569	30	2
Musk2	D4	6598	167	2

This study employs four benchmark datasets obtained from publicly available machine learning repositories that are widely used in feature selection and classification research [13], [18-19]. The selected datasets represent different domains and exhibit varying levels of dimensionality, instance size, and feature complexity. These characteristics make them suitable for evaluating the effectiveness of feature selection techniques in high-dimensional classification problems.

The first dataset, Mutagenesis-Bonds, is derived from the original Mutagenesis dataset and focuses on chemical bond-related attributes. It contains structural representations of chemical compounds and is widely utilized to evaluate machine learning models in predicting mutagenic properties. Due to the large number of bond-related features, this dataset is characterized by high dimensionality and the presence of redundant information.

The second dataset, Mutagenesis-Chains, is also sourced from the Mutagenesis dataset but emphasizes molecular chain descriptors. Similar to Mutagenesis-Bonds, it includes a substantial number of features that describe chemical structures, making it a challenging dataset for feature selection techniques that aim to identify the most informative attributes for classification tasks.

The third dataset, Wisconsin Diagnostic Breast Cancer (WDBC), is obtained from the UCI Machine Learning Repository. It comprises 569 instances with 30 numerical features derived from digitized images of fine needle aspirates of breast tissue. Although it has lower dimensionality compared to the Mutagenesis datasets, WDBC is commonly used as a benchmark dataset for evaluating classification models and assessing feature relevance in medical diagnosis applications.

The fourth dataset, Musk2, consists of molecular conformational data used to distinguish between musky and non-musky compounds. It contains a relatively high number of features describing molecular shape and spatial characteristics. Musk2 is frequently employed in feature selection research due to its high-dimensional nature and multiple-instance structure, which introduce additional complexity for classification models.

All datasets were preprocessed prior to the feature selection stage. This preprocessing includes handling missing values where applicable and applying feature scaling to ensure consistent evaluation across different classification algorithms. By utilizing datasets from different domains with diverse dimensional characteristics, this study provides a comprehensive evaluation of the proposed IF-RFECV method in various high-dimensional classification scenarios.

2.2 Preprocessing Data

In the data cleaning stage, several steps are carried out to ensure the quality and reliability of the dataset. The first step is *filling in blank data*, which involves identifying and handling missing values using appropriate techniques such as mean, median, or mode imputation, or in some cases, predictive modeling. This step ensures data completeness and prevents bias in the learning process. The second step is *eliminating data duplication*, where duplicate records are detected and removed to maintain data integrity and ensure that each entry represents unique information, avoiding potential distortion during model training.

The third step is *checking for data inconsistencies*, which focuses on verifying the consistency of data formats, units, and values. For instance, categorical variables are standardized (e.g., using consistent labels such as “Yes” and “No”), and numerical features are examined to ensure they fall within valid ranges. The fourth step is *correcting errors in data*, where mistakes such as typos, out-of-range values, or misclassified entries are identified and corrected using logical rules or domain knowledge.

After completing all these steps, it was found that the datasets used in this study contained no missing values or duplicate records, indicating that the data were already clean and ready for subsequent preprocessing stages.

2.3 Feature Selection

During the feature selection phase, a hybrid strategy integrating both wrapper and filter approaches was utilized to determine the most relevant and optimal subset of features. The process begins with the application of Recursive Feature Elimination with Cross-Validation (RFECV), a wrapper-based technique that iteratively evaluates and eliminates less important features based on model performance [23-26]. In this stage, a Decision Tree estimator is employed as the primary model within the feature elimination process. Meanwhile, multiple learning algorithms, including Gradient Boosting (GB), K-Nearest Neighbor (KNN), Naïve Bayes (NB), Decision Tree (DT), Random Forest (RF), and Support Vector Machine (SVM), are integrated with the RFECV framework to assess and rank feature importance.

The RFECV method works by fitting a model f to the training data (X, y), ranking features according to their importance scores $I(f, x_i)$, and recursively eliminating the least important features. The process is repeated until the optimal subset of features is obtained. The selection criterion is determined by cross-validation performance CV. The optimal feature subset is defined in Equation 1.

$$S^* = \arg max_{S \subseteq F} CV\ (fs) \quad (1)$$

In equation (1), $F = \{x_1, x_2, \ldots, x_n\}$ is the full set of features. $S \subseteq F$ is a candidate subset of features. fs is the model trained on subset S. $CV\ (fs)$ is the cross-validation score for subset S and S^* is the optimal subset that yields the best cross-validation score. During each iteration of the RFECV process, features are ranked according to their importance scores. The ranking function is defined in Equation 2.

$$R(x_i) = I(f, x_i) \quad (2)$$

In Equation 2, scores $I(f, x_i)$ represents the importance score of feature $x_i$with respect to the trained model f. Features with lower importance scores are considered less relevant to the predictive performance. The feature elimination step is performed by removing the feature with the minimum importance score. The updated feature subset at iteration $t + 1$ i s expressed in Equation 3.

$$F_{t+1} = F_t \left\{x_j | R(x_j)\ minimal\right\} \quad (3)$$

In Equation 3, F_t denotes the feature subset at iteration t, and the feature $x_j$with the lowest ranking score is removed. This recursive elimination process continues until a stopping criterion is achieved, such as reaching the highest cross-validation score or a predefined number of features. To enhance the robustness of feature selection, RFECV is applied using multiple base classifiers, resulting in several feature subsets. These subsets are denoted as: $S_{GB}, S_{KNN}, S_{NB}, S_{DT}, S_{RF}, S_{SVM}$. Each subset corresponds to features selected by a specific model, including Gradient Boosting (GB), K-Nearest Neighbor (KNN), Naïve Bayes (NB), Decision Tree (DT), Random Forest (RF), and Support Vector Machine (SVM). To identify features that are consistently selected across different models, an intersection operation is applied. The final feature subset is defined in Equation 4.

$$S_{IF-RFECV} = S_{GB} \cap S_{KNN} \cap S_{NB} \cap S_{DT} \cap S_{RF} \cap S_{SVM} \quad (4)$$

In Equation 4, $S_{\text{IF-RFECV}}$ represents the intersection of all feature subsets obtained from different models. This subset contains features that are consistently selected, indicating their stability and relevance across multiple learning algorithms. Finally, the features selected through IF-RFECV are used for classification model and performance evaluation. This hybrid approach ensures that the selected features are both model-driven (through RFECV) and statistically robust (through intersection filtering), which improves the accuracy, generalization, and computational efficiency of the final classification models.

In this study, Recursive Feature Elimination with Cross-Validation (RFECV) is employed as the core wrapper-based feature selection technique. RFECV iteratively ranks and removes the least important features based on the importance scores or coefficients provided by the underlying classifier, while model performance is evaluated using stratified five-fold cross-validation to determine the optimal number of selected features. This strategy ensures balanced class distribution across folds and provides a reliable trade-off between computational efficiency and performance estimation.

All classification algorithms used in this study adopt commonly used default parameter settings provided by the machine learning library to ensure a fair and reproducible evaluation. This design choice aims to assess the effectiveness of the proposed feature selection method rather than optimizing individual classifier performance through extensive hyperparameter tuning. Default configurations are applied for Gradient Boosting, Random Forest, Decision Tree, Support Vector Machine, K-Nearest Neighbor, and Naïve Bayes, allowing observed performance differences to be primarily attributed to the feature selection process. Consequently, this approach supports the main objective of evaluating the robustness and effectiveness of the proposed IF-RFECV method for high-dimensional feature selection.

2.4 Model Classification

This study employs multiple classification algorithms, including Gradient Boosting (GB), K-Nearest Neighbor (KNN), Naïve Bayes (NB), Decision Tree (DT), Random Forest (RF), and Support Vector Machine (SVM). The inclusion of diverse classifiers aims to ensure a comprehensive and unbiased evaluation of the proposed feature selection approach. Each algorithm reflects a different learning paradigm and possesses unique characteristics in terms of decision boundaries, sensitivity to features, and underlying learning mechanisms.

Gradient Boosting and Random Forest are categorized as ensemble learning methods capable of modeling complex non-linear relationships and feature interactions. Decision Tree is utilized as a base learner due to its interpretability and its sensitivity to irrelevant attributes, making it effective for assessing feature selection performance. Support Vector Machine is incorporated because of its strong theoretical basis and its effectiveness in handling high-dimensional data, particularly when the number of features is greater than the number of samples.

K-Nearest Neighbor is applied as an instance-based method that relies on distance calculations, making it highly sensitive to redundant and irrelevant features. In addition, Naïve Bayes is used as a probabilistic classifier that assumes feature independence. The

performance of both KNN and NB is strongly influenced by the quality of the selected features, thereby offering further insight into the robustness and effectiveness of the proposed feature selection method.

The utilization of multiple classification algorithms is a key contribution of this study to address model-dependent bias in feature selection, a limitation commonly observed in wrapper-based methods such as RFECV, where the selected feature subsets are highly influenced by the characteristics of a single classifier. Unlike conventional approaches that rely on one model, the proposed IF-RFECV method introduces an intersection filtering mechanism that retains only features consistently selected across diverse classifiers. This strategy ensures that the resulting feature subset is not only model-independent but also more stable and robust across different learning paradigms. Consequently, the proposed approach provides a more reliable framework for feature selection, significantly enhancing classification performance, feature stability, and generalization capability compared to traditional single-model RFECV methods.

2.5 Evaluation

The final stage of this study involves evaluating the performance of the developed model. In this phase, a comparative analysis is conducted between the proposed Intersection Filtering based on Recursive Feature Elimination with Cross-Validation (IF-RFECV) method and previously established approaches [13], [19] to assess its effectiveness. Several evaluation indicators are used in conducting testing and model comparisons, especially for the classification model. To measure and compare the performance of the classification models, several evaluation metrics are utilized, including Accuracy, Precision, Recall, F1 Score, and processing time (in seconds). The Accuracy metric is presented in Equation 5, Precision in Equation 6, Recall in Equation 7, and the F1 Score in Equation 8.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (5)$$

$$Precision = \frac{TP}{TP+FP} \quad (6)$$

$$Recall = \frac{TP}{TP+FN} \quad (7)$$

$$F1\ Score = \frac{2\ x\ Precision\ x\ recall}{precision\ +recall} \quad (8)$$

Based on Equations 5–8, the evaluation metrics are derived from the confusion matrix, which summarizes the classification results into four categories. True Positive (TP) refers to the number of instances that belong to the positive class and are correctly predicted as positive by the model. In contrast, False Positive (FP) indicates the number of instances that actually belong to the negative class but are incorrectly classified as positive.

False Negative (FN) represents the number of positive instances that are misclassified as negative, reflecting the model's inability to correctly identify certain positive cases. Meanwhile, True Negative (TN) denotes the number of negative instances that are correctly predicted as negative. These four components form the basis for calculating performance metrics such as Accuracy, Precision, Recall, and F1 Score.

3. Results and Discussions

The results and discussion section focuses on the feature selection process, testing the classification model and evaluating the comparison of the classification model.

3.1 Intersection Filtering based on RFE

The feature selection process was carried out in two sequential stages using a hybrid technique that integrates wrapper and filter approaches (IF-RFECV). Based on the procedures outlined in the methodology section, the results of the feature selection process for each dataset are presented. The first dataset, Mutagenesis-Bonds (D1), is used as the initial case study. The selected features derived from this dataset are summarized in Table 2 and illustrated in Figure 3.

Table 2. Mutagenesis-Bonds Dataset Selection Features Results

Method	Base Model	Selected Features	Features
RFECV	GB, KNN, DT, SVM	Charge1, Charge2, Quanta1, Quanta2	4
RFECV	NB	Charge1, Charge2, Elementtype1–9, Quanta1, Quanta2	13
RFECV	RF	Charge1, Charge2, Elementtype1–6, Quanta1, Quanta2	9
IF-RFECV	—	Charge1, Charge2, Quanta1, Quanta2	4

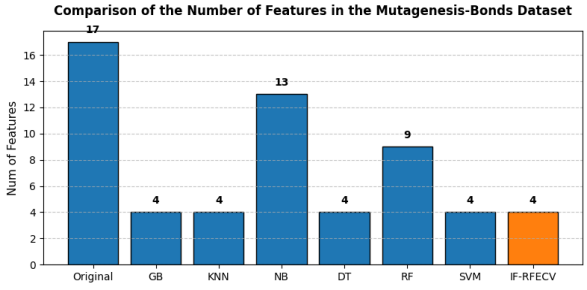


Figure 2. Comparison of the Number of Features in the Mutagenesis-Bonds Dataset

Table 2 and Figure 2 present a comparative analysis of feature selection results obtained from multiple RFECV-based models. The results show that Gradient Boosting (GB), K-Nearest Neighbor (KNN), Decision Tree (DT), and Support Vector Machine (SVM) consistently selected four features, while Naïve Bayes (NB) and Random Forest (RF) produced larger subsets with 13 and 9 features, respectively. This variation reflects differences in how each model evaluates feature relevance, with Naïve Bayes tending to retain more features and tree-based models performing implicit feature selection.

To enhance robustness, the IF-RFECV approach, as defined in equation (4), is applied to identify features consistently selected across all models. This process

yields a stable subset of four features—Charge1, Charge2, Quanta1, and Quanta2—for the Mutagenesis-Bonds (D1) dataset, demonstrating improved feature stability and generalization. The analysis is then extended to the Mutagenesis-Chains (D2) dataset, with the results summarized in Table 3 and illustrated in Figure 3.

Table 3 and Figure 3 present a comparative analysis of feature selection results obtained from RFECV-based models. The results indicate that Decision Tree (DT), Naïve Bayes (NB), and Random Forest (RF) each selected six features, while Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and Gradient Boosting (GB) retained 12, 14, and 16 features, respectively. This variation reflects differences in model sensitivity to feature importance, where tree-based and probabilistic models perform more aggressive feature elimination, while distance-based and ensemble models tend to retain a larger subset.

Table 3. Mutagenesis-Chains Dataset Selection Features Results

Method	Base Model	Selected Features	Features
RFECV	GB	Bond2, Q1, E3–n, E2–c, Q2, El1–h, El1–o, El1–cl, El1–c, El1–br, Charge3, Charge2, Charge1, El1–f, Q3	16
RFECV	KNN	Bond2, Q1, E3–n, E2–c, Q2, El1–h, El1–o, El1–cl, El1–c, Charge3, Charge2, Charge1, Q3	14
RFECV	NB, DT, RF	Q3, Charge1, Charge2, Charge3, Q1, El1–cl	6
RFECV	SVM	Bond2, Q1, E3–n, Q2, El1–h, El1–cl, El1–o, Charge3, Charge2, Charge1, E1–c, Q3	12
IF-RFECV	—	Q3, Charge1, Charge2, Charge3, Q1, El1–cl	6

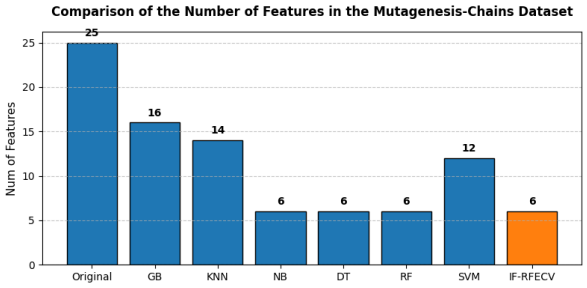


Figure 3. Comparison of the Number of Features in the Mutagenesis-Chains Dataset

To improve robustness, the IF-RFECV approach, as defined in equation (4), is applied to identify features consistently selected across all models. This process produces a stable subset of six features—Q3, Charge1, Charge2, Charge3, Q1, and E1=cl—for the Mutagenesis-Chains (D2) dataset, demonstrating improved feature stability and generalization. The analysis is then extended to the WDBC (D3) dataset, with the results summarized in Table 4 and illustrated in Figure 4.

Table 4 and Figure 4 present a comparative analysis of feature selection outcomes across multiple RFECV-based models. The results indicate that Gradient Boosting (GB), Decision Tree (DT), and Random Forest (RF) each retained five features, while Support Vector Machine (SVM) and Naïve Bayes (NB) selected six features, and K-Nearest Neighbor (KNN) retained eight features. This variation reflects differences in how each model captures feature relevance based on its inherent learning characteristics.

To identify the most robust features, the IF-RFECV approach, as defined in equation (4), is applied to retain only features consistently selected across all models. This process produces four core features—Concave_points3, Area3, Texture3, and Radius3—which represent the most discriminative attributes in the WDBC dataset. These features effectively reduce redundancy while preserving essential information, thereby enhancing feature stability and model generalization across classifiers.

Table 4. WDBC Dataset Selection Features Results

Method	Base Model	Selected Features	Features
RFECV	GB	Concave_points3, Area3, Texture3, Radius3, Concave_points1	5
RFECV	KNN	Area2, Compactness3, Area3, Perimeter3, Texture3, Radius3, Concavity3, Concave_points3	8
RFECV	NB	Area2, Area3, Texture3, Radius3, Concave_points3, Concavity3	6
RFECV	DT	Texture3, Concave_points3, Area3, Compactness2, Radius3	5
RFECV	RF	Texture3, Concave_points3, Area3, Radius3, Symmetry3	5
RFECV	SVM	Concavity3, Area3, Texture3, Radius3, Concave_points3, Compactness3	6
IF-RFECV	—	Concave_points3, Area3, Texture3, Radius3	4

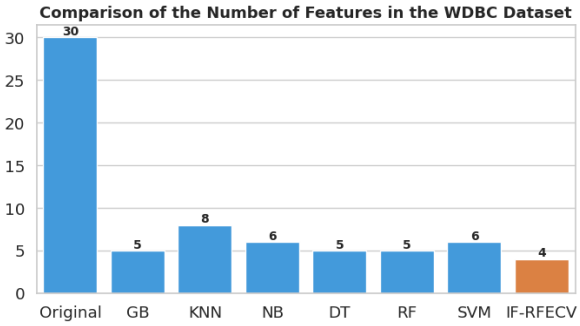


Figure 4. Comparison of the Number of Features in the WDBC Dataset

The analysis is subsequently extended to the Musk (D4) dataset, with the selected features summarized in Table 5 and illustrated in Figure 5.

Table 5 and Figure 5 present a comparative analysis of feature selection outcomes across multiple RFECV-based models applied to the Musk (D4) dataset. The results show that each model produces a distinct

number of selected features, with Gradient Boosting (GB) and Naïve Bayes (NB) retaining 55 features, Random Forest (RF) selecting 54 features, K-Nearest Neighbor (KNN) producing 56 features, Support Vector Machine (SVM) selecting 45 features, and Decision Tree (DT) resulting in 44 features. This variation reflects differences in how each model evaluates feature relevance based on its inherent learning mechanism.

To improve robustness, the IF-RFECV approach, as defined in equation (4), is applied to identify features consistently selected across models. This process reduces the feature set from an average of approximately 51 features to 42 intersecting features, resulting in a more compact yet representative subset. The selected features effectively capture the most relevant attributes of the Musk dataset, demonstrating that the integration of RFECV and intersection filtering enhances feature stability, reduces dimensionality, and preserves essential information for classification.

Table 5. WDBC Dataset Selection Features Results

Method	Base Model	Selected Features	Features
RFECV	GB	FE6, FE8, FE9, FE17, FE20, FE22, FE24, FE25, FE32, FE34, FE35, FE36, FE38, FE39, FE42, FE43, FE48, FE50, FE52, FE54, FE55, FE56, FE61, FE62, FE63, FE64, FE66, FE69, FE75, FE77, FE83, FE85, FE89, FE93, FE96, FE99, FE101, FE103, FE106, FE108, FE116, FE117, FE122, FE124, FE126, FE132, FE133, FE136, FE140, FE151, FE152, FE156, FE158, FE162, FE163	55
RFECV	KNN	FE3, FE4, FE6, FE8, FE9, FE17, FE20, FE22, FE24, FE25, FE29, FE32, FE34, FE35, FE36, FE38, FE39, FE42, FE43, FE48, FE50, FE52, FE54, FE55, FE56, FE62, FE63, FE64, FE66, FE69, FE75, FE77, FE83, FE89, FE93, FE96, FE99, FE101, FE103, FE106, FE108, FE116, FE117, FE122, FE124, FE126, FE132, FE133, FE136, FE140, FE151, FE152, FE156, FE158, FE162, FE163	56
RFECV	NB	FE4, FE6, FE8, FE9, FE10, FE17, FE20, FE22, FE24, FE25, FE32, FE34, FE35, FE36, FE38, FE39, FE42, FE43, FE48, FE50, FE52, FE54, FE55, FE56, FE62, FE63, FE64, FE66, FE69, FE72, FE75, FE77, FE83, FE89, FE93, FE96, FE99, FE103, FE106, FE108, FE116, FE117, FE122, FE124, FE126, FE132, FE133, FE140, FE151, FE152, FE156, FE158, FE161, FE162, FE163	55
RFECV	DT	FE9, FE17, FE20, FE22, FE25, FE32, FE34, FE35, FE36, FE38, FE39, FE42, FE43, FE50, FE52, FE54, FE55, FE56, FE62, FE63, FE64, FE66, FE75, FE77, FE83, FE89, FE93, FE96, FE99, FE101, FE103, FE106, FE108, FE122, FE124, FE126, FE132, FE133, FE140, FE151, FE152, FE156, FE158, FE163	44
RFECV	RF	FE6, FE8, FE9, FE17, FE20, FE22, FE24, FE25, FE29, FE32, FE34, FE35, FE36, FE38, FE39, FE42, FE43, FE48, FE50, FE52, FE54, FE55, FE56, FE61, FE62, FE63, FE64, FE66, FE69, FE75, FE77, FE83, FE89, FE93, FE96, FE99, FE101, FE103, FE106, FE108, FE116, FE117, FE122, FE124, FE126, FE132, FE133, FE140, FE151, FE152, FE156, FE158, FE162, FE163	54
RFECV	SVM	FE9, FE17, FE20, FE22, FE24, FE25, FE32, FE34, FE35, FE36, FE38, FE42, FE43, FE50, FE52, FE54, FE55, FE62, FE63, FE64, FE66, FE75, FE77, FE83, FE89, FE93, FE96, FE99, FE101, FE103, FE106, FE108, FE116, FE122, FE124, FE126, FE132, FE133, FE140, FE151, FE152, FE156, FE158, FE162, FE163	45
IF-RFECV	—	FE9, FE17, FE20, FE22, FE25, FE32, FE34, FE35, FE36, FE38, FE42, FE43, FE50, FE52, FE54, FE55, FE62, FE63, FE64, FE66, FE75, FE77, FE83, FE89, FE93, FE96, FE99, FE103, FE106, FE108, FE122, FE124, FE126, FE132, FE133, FE140, FE151, FE152, FE156, FE158, FE162, FE163	42

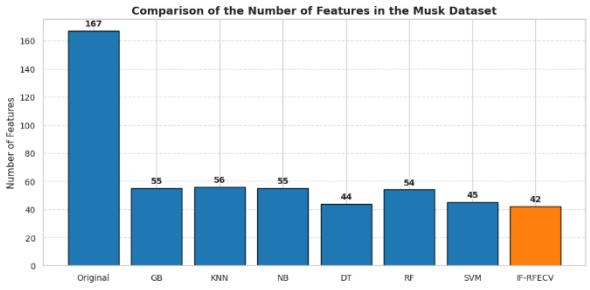


Figure 5. Comparison of the Number of Features in the WDBC Dataset

3.2 Model Evaluation and Comparison

Following the completion of the feature selection stage, the next step involves evaluating and comparing the performance of several classification models. This evaluation is conducted to examine how effectively each model performs when trained using the optimal feature subsets generated by the proposed method. The classification algorithms considered in this study include Gradient Boosting (GB), K-Nearest Neighbor (KNN), Naïve Bayes (NB), Decision Tree (DT), Random Forest (RF), and Support Vector Machine (SVM).

To provide a comprehensive assessment, five evaluation metrics are employed: Accuracy, Precision, Recall, F1-Score, and processing time (in seconds). Accuracy measures the overall correctness of the model, while Precision and Recall assess its ability to correctly identify positive instances and capture all relevant cases. The F1-Score, defined as the harmonic mean of Precision and Recall, offers a balanced evaluation of both metrics. In addition, processing time

is used to evaluate computational efficiency, which is particularly important in practical applications.

This section presents the performance evaluation and comparison of the classification models based on the metrics defined in Equations (5)–(8), including Accuracy, Precision, Recall, and F1-Score, along with processing time. These metrics are used to assess both predictive capability and computational efficiency. The results of the accuracy evaluation are summarized in Table 6.

Table 6. Comparison of Accuracy Values

	Method	GB	KNN	NB	DT	RF	SVM
D1	RFECV	0.797	0.812	0.606	0.82	0.833	0.75
	IF-RFECV	0.796	0.799	0.713	0.811	0.819	0.75
D2	RFECV	0.828	0.82	0.66	0.829	0.844	0.784
	IF-RFECV	0.83	0.822	0.785	0.853	0.855	0.785
D3	RFECV	0.96	0.909	0.946	0.935	0.951	0.904
	IF-RFECV	0.953	0.9	0.974	0.942	0.964	0.911
D4	RFECV	0.801	0.737	0.771	0.736	0.759	0.848
	IF-RFECV	0.807	0.813	0.811	0.778	0.797	0.854

Table 6 demonstrates that the proposed feature selection method, Intersection Filtering based on RFECV (IF-RFECV), consistently achieves higher accuracy compared to the conventional RFECV approach, particularly in Datasets D2, D3, and D4, as well as across the majority of the evaluated classification models. This improvement highlights the effectiveness of integrating both wrapping and filtering strategies in refining the feature subset to enhance model performance.

The superior results achieved by IF-RFECV can be attributed to its ability to identify features that are not only relevant within individual models but also consistently significant across multiple learning algorithms. By focusing on intersecting features from different RFECV-based models, the IF-RFECV approach effectively minimizes redundancy and noise that may occur in model-specific feature spaces. Consequently, the resulting feature subset enhances the discriminative capacity of the models, leading to better generalization and higher predictive accuracy.

These findings suggest that the intersection-based filtering mechanism provides an additional layer of robustness to the traditional RFECV process, especially when applied to heterogeneous datasets with varying feature dependencies. Therefore, IF-RFECV can be considered a promising hybrid feature selection strategy for improving the reliability and efficiency of machine learning models in high-dimensional data scenarios.

The results of the precision evaluation are presented in Table 7. Table 7 reveals that the proposed feature selection method, Intersection Filtering based on RFECV (IF-RFECV), yields superior precision values compared to the conventional RFECV method across Datasets D2, D3, and D4, as well as in the majority of the evaluated classification models. This finding indicates that the IF-RFECV approach not only enhances model accuracy but also improves the ability of classifiers to correctly identify positive instances while minimizing false positives.

The improvement in precision demonstrates that the intersection-based filtering mechanism effectively refines the feature space by retaining only the most discriminative and stable features across multiple RFECV-based models. By eliminating redundant and irrelevant attributes, the IF-RFECV method enables classifiers to make more confident and reliable predictions. This leads to a more focused decision boundary, particularly beneficial in datasets where feature overlap or noise may otherwise reduce classification precision.

Table 7. Comparison of Precision Values

	Method	GB	KNN	NB	DT	RF	SVM
D1	RFECV	0.84	0.842	0.705	0.885	0.891	0.756
	IF-RFECV	0.838	0.833	0.75	0.876	0.877	0.755
D2	RFECV	0.867	0.87	0.808	0.911	0.913	0.793
	IF-RFECV	0.869	0.867	0.789	0.915	0.91	0.787
D3	RFECV	0.962	0.917	0.949	0.932	0.957	0.875
	IF-RFECV	0.962	0.91	0.971	0.959	0.968	0.918
D4	RFECV	0.554	0.409	0.463	0.448	0.517	0.8
	IF-RFECV	0.644	0.594	0.56	0.562	0.646	0.7

Overall, these results highlight the robustness of IF-RFECV in improving the reliability of classification outcomes. The method's ability to maintain high precision across diverse datasets and algorithms suggests that it effectively enhances the model's discriminative power, making it a valuable approach for feature selection in complex or high-dimensional learning environments.

The results of the recall evaluation are presented in Table 8. Table 8 illustrates that the proposed feature selection method, Intersection Filtering based on RFECV (IF-RFECV), achieves superior recall values compared to the conventional RFECV method across Datasets D2, D3, and D4, as well as in the majority of the evaluated classification models. This indicates that the IF-RFECV approach enhances the model's ability to correctly identify relevant positive instances, reducing the likelihood of false negatives and improving overall sensitivity.

Table 8. Comparison of Recall Values

	Method	GB	KNN	NB	DT	RF	SVM
D1	RFECV	0.902	0.919	0.874	0.87	0.885	0.981
	IF-RFECV	0.896	0.911	0.92	0.867	0.881	0.98
D2	RFECV	0.913	0.902	0.637	0.865	0.88	0.982
	IF-RFECV	0.918	0.908	0.983	0.892	0.899	0.989
D3	RFECV	0.972	0.942	0.97	0.939	0.975	0.956
	IF-RFECV	0.964	0.936	0.989	0.947	0.97	0.945
D4	RFECV	0.628	0.723	0.649	0.644	0.695	0.01
	IF-RFECV	0.67	0.696	0.621	0.754	0.749	0.05

The improvement in recall performance can be attributed to IF-RFECV's capability to retain only the most essential and generalizable features that contribute

meaningfully to distinguishing between classes. By combining the strengths of both wrapping and filtering strategies, the method preserves features that are consistently important across multiple RFECV-based models, ensuring that critical patterns in the data are not discarded during the reduction process. Consequently, the resulting models are better able to capture subtle but significant variations within the data, leading to a higher true positive rate.

Overall, these results indicate that IF-RFECV not only effectively reduces feature dimensionality but also improves model robustness by maintaining strong sensitivity across various classification algorithms and datasets. This highlights the potential of the proposed method for detection-oriented applications, particularly in scenarios where correctly identifying all relevant instances is more critical than minimizing false positives.

The evaluation results of the F1-Score are presented in Table 9. As shown in Table 9, the proposed IF-RFECV method consistently achieves higher F1-Score values compared to the conventional RFECV approach across datasets D1, D2, D3, and D4, as well as across most classification models. Since the F1-Score represents the harmonic mean of Precision and Recall, these results indicate that IF-RFECV provides a balanced improvement in both metrics, enabling more accurate identification of positive instances while reducing both false positives and false negatives.

Table 9. Comparison of F1 Score Values

	Method	GB	KNN	NB	DT	RF	SVM
D1	RFECV	0.869	0.878	0.758	0.879	0.884	0.854
	IF-RFECV	0.866	0.87	0.826	0.871	0.874	0.853
D2	RFECV	0.891	0.885	0.839	0.886	0.898	0.876
	IF-RFECV	0.892	0.887	0.876	0.902	0.905	0.876
D3	RFECV	0.964	0.923	0.956	0.95	0.971	0.942
	IF-RFECV	0.963	0.922	0.98	0.953	0.973	0.93
D4	RFECV	0.548	0.472	0.496	0.485	0.543	0.027
	IF-RFECV	0.59	0.596	0.542	0.573	0.63	0.091

The consistent improvement in F1-Score demonstrates that the IF-RFECV method effectively maintains equilibrium between Precision and Recall, ensuring that the enhancement in one metric does not come at the expense of the other. By integrating the intersection of features derived from multiple RFECV-based models, IF-RFECV is able to isolate the most stable and discriminative feature subset, thereby strengthening the classifier's overall predictive consistency.

These results further validate the robustness of the IF-RFECV approach as a hybrid feature selection technique that not only reduces feature dimensionality but also optimizes the trade-off between detection completeness and prediction accuracy. Consequently, IF-RFECV provides a more balanced and reliable performance across diverse datasets and classification algorithms, reinforcing its potential applicability in real-world machine learning scenarios.

The processing time evaluation results (in seconds) are presented in Table 10. Table 10 shows that the proposed feature selection method, Intersection Filtering based on RFECV (IF-RFECV), requires a slightly longer processing time compared to the conventional RFECV approach. This increase in computational time is primarily attributed to the two-phase mechanism of IF-RFECV, which first executes the RFECV process for each classification model and subsequently performs the intersection filtering to extract the most stable and relevant feature subset. As a result, the method inherently introduces an additional layer of computation after the RFECV stage.

Table 10. Comparison of the time processing (in seconds)

	Method	GB	KNN	NB	DT	RF	SVM
D1	RFECV	17.3	12.7	12.4	13.2	17.1	21.3
	IF-RFECV	19.3	13.1	12.6	13.4	20.2	27.5
D2	RFECV	45	42.23	41.68	41.93	45.78	54.22
	IF-RFECV	48	42.73	41.86	42.93	49.98	64.82
D3	RFECV	22.7	17.8	21	17.3	21.8	21.4
	IF-RFECV	23.9	18	21.2	17.5	23.9	21.8
D4	RFECV	282.5	256.4	254.9	253.9	271.1	251.5
	IF-RFECV	288.6	257.1	254.9	254.1	272.5	257.7

However, the slight increase in processing time is a reasonable trade-off considering the improved classification performance achieved in terms of Accuracy, Precision, Recall, and F1-Score across most datasets. This indicates that the additional computational cost incurred by IF-RFECV contributes positively to generating more robust and generalizable models. Therefore, while IF-RFECV may not be the fastest method, it demonstrates a favorable balance between computational efficiency and model performance, making it suitable for applications that prioritize accuracy and feature stability over minimal execution time.

The experimental results show that IF-RFECV consistently selects fewer features than standard RFECV across all evaluated datasets, with more pronounced reductions observed in high-dimensional datasets such as D1, D2, and D4. This improvement is achieved by applying an intersection filtering mechanism that retains only features consistently selected across multiple RFECV-based models, resulting in a more compact and stable feature subset.

In terms of classification performance, IF-RFECV achieves comparable or superior accuracy, precision, recall, and F1-score for most classifiers across D1, D2, D3, and D4, particularly for classifiers sensitive to irrelevant and redundant features such as K-Nearest Neighbor and Naïve Bayes. Although IF-RFECV introduces a moderate increase in computational time due to the additional filtering stage, this trade-off is considered acceptable given the improvements in feature reduction, robustness, and overall classification performance.

3.3 Discussion

The experimental results demonstrate that the proposed IF-RFECV method consistently improves feature reduction and classification performance across different datasets. The most significant improvements are observed in high-dimensional datasets D1, D2, and D4, where feature redundancy and model-dependent bias are more prominent. In these datasets, the intersection filtering mechanism effectively eliminates unstable and classifier-specific features, leading to a more compact and reliable feature subset.

For classifiers that are sensitive to irrelevant and redundant features, such as K-Nearest Neighbor and Naïve Bayes, the performance gains achieved by IF-RFECV are more evident across D1, D2, D3, and D4. This behavior indicates that reducing noisy features through intersection-based selection enhances distance-based and probabilistic learning processes. Meanwhile, ensemble-based classifiers, including Gradient Boosting and Random Forest, maintain competitive performance while benefiting from reduced feature dimensionality, suggesting that the selected features preserve sufficient discriminative information.

The additional computational cost introduced by IF-RFECV is primarily associated with the intersection filtering stage applied after RFECV. However, this overhead remains moderate and is outweighed by the advantages of improved feature stability and reduced dimensionality, particularly for high-dimensional datasets D1, D2, and D4. These findings confirm that IF-RFECV offers a robust and practical feature selection solution for high-dimensional classification tasks, where consistent and efficient feature reduction is required across multiple learning models.

4. Conclusions

This study proposed a hybrid feature selection approach, namely Intersection Filtering based on Recursive Feature Elimination with Cross-Validation (IF-RFECV), to address feature redundancy and model-dependent bias in high-dimensional datasets. Experimental results on four benchmark datasets (D1, D2, D3, and D4) demonstrate that the proposed method consistently selects a more compact feature subset compared to conventional RFECV while maintaining comparable or improved classification performance across multiple learning algorithms. The most significant improvements are observed in high-dimensional datasets, where IF-RFECV effectively enhances feature stability and reduces the influence of irrelevant and classifier-specific features.

From a practical perspective, IF-RFECV is particularly suitable for classification problems involving high-dimensional data and limited training samples, where robust and consistent feature selection is essential. Although the proposed method introduces a moderate increase in computational time due to the additional intersection filtering stage, this trade-off is justified by the improvements in feature reduction, robustness, and overall model generalization. Future work will focus on optimizing the computational efficiency of IF-RFECV, investigating alternative feature aggregation strategies, and extending the evaluation to larger-scale datasets and additional learning paradigms to further validate the applicability of the proposed approach.

Acknowledgements

Thanks to the Department of Research and Community Service, Universitas Amikom Yogyakarta for funding and supporting this research.

References

[1] N. Almusallam, Z. Tari, J. Chan, A. Fahad, A. Alabdulatif, and M. Al-Naeem, "Towards an Unsupervised Feature Selection Method for Effective Dynamic Features," *IEEE Access*, vol. 9, pp. 77149–77163, 2021, doi: 10.1109/ACCESS.2021.3082755.

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

[3] R. Islam, M. I. Sayed, S. Saha, M. J. Hossain, and M. A. Masud, "Android malware classification using optimum feature selection and ensemble machine learning," *Internet of Things and Cyber-Physical Systems*, vol. 3, no. January, pp. 100–111, 2023, doi: 10.1016/j.iotcps.2023.03.001.

[4] P. Morán, A. Robles-Gómez, A. Duque, L. Tobarra, and R. Pastor-Vargas, "Machine learning models and dimensionality reduction for improving the Android malware detection," *PeerJ Comput Sci*, vol. 10, pp. 1–31, 2024, doi: 10.7717/PEERJ-CS.2616.

[5] R. Islam, M. I. Sayed, S. Saha, M. J. Hossain, and M. A. Masud, "Android malware classification using optimum feature selection and ensemble machine learning," *Internet of Things and Cyber-Physical Systems*, vol. 3, no. October 2022, pp. 100–111, 2023, doi: 10.1016/j.iotcps.2023.03.001.

[6] A. Pathak, U. Barman, and T. S. Kumar, "Machine learning approach to detect android malware using feature-selection based on feature importance score," *Journal of Engineering Research (Kuwait)*, no. March, 2024, doi: 10.1016/j.jer.2024.04.008.

[7] A. Newaz and S. Muhtadi, "Performance Improvement of Heart Disease Prediction by Identifying Optimal Feature Sets Using Feature Selection Technique," in *2021 International Conference on Information Technology (ICIT)*, 2021, pp. 446–450. doi: 10.1109/ICIT52682.2021.9491739.

[8] R. Islam, M. I. Sayed, S. Saha, M. J. Hossain, and M. A. Masud, "Android malware classification using optimum feature selection and ensemble machine learning," *Internet of Things and Cyber-Physical Systems*, vol. 3, no. January, pp. 100–111, 2023, doi: 10.1016/j.iotcps.2023.03.001.

[9] N. Almusallam, Z. Tari, J. Chan, A. Fahad, A. Alabdulatif, and M. Al-Naeem, "Towards an Unsupervised Feature Selection Method for Effective Dynamic Features," *IEEE Access*, vol. 9, pp. 77149–77163, 2021, doi: 10.1109/ACCESS.2021.3082755.

[10] E. Strelcenia and S. Prakoonwit, "Effective Feature Engineering and Classification of Breast Cancer Diagnosis: A Comparative Study," *BioMedInformatics*, vol. 3, no. 3, pp. 616–631, 2023, doi: 10.3390/biomedinformatics3030042.

[11] D. Ö. Şahin, O. E. Kural, S. Akleylek, and E. Kılıç, "Permission-based Android malware analysis by using dimension reduction with PCA and LDA," *Journal of Information Security and Applications*, vol. 63, no. October, p. 102995, 2021, doi: 10.1016/j.jisa.2021.102995.

[12] C. Marquez-Vera, A. Cano, C. Romero, and S. Ventura, "Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data," *Applied Intelligence*, vol. 38, no. 3, pp. 315–330, 2013, doi: 10.1007/s10489-012-0374-8.

[13] Khadijah, A. Adorada, P. W. Wirawan, and K. Kurniawan, "The Comparison of Feature Selection Methods in Software Defect Prediction," *ICICoS 2020 - Proceeding: 4th International*

Conference on Informatics and Computational Sciences, pp. 0–5, 2020, doi: 10.1109/ICICoS51170.2020.9299022.

[14] S. A. Roseline and S. Geetha, "Android Malware Detection and Classification using LOFO Feature Selection and Tree-based Models," *J Phys Conf Ser*, vol. 1911, no. 1, 2021, doi: 10.1088/1742-6596/1911/1/012031.

[15] M. Riajuliislam, K. Z. Rahim, and A. Mahmud, "Prediction of Thyroid Disease(Hypothyroid) in Early Stage Using Feature Selection and Classification Techniques," *2021 International Conference on Information and Communication Technology for Sustainable Development, ICICT4SD 2021 - Proceedings*, pp. 60–64, 2021, doi: 10.1109/ICICT4SD50815.2021.9397052.

[16] S. Murugesan, R. S. Bhuvaneswaran, H. Khanna Nehemiah, S. Keerthana Sankari, and Y. Nancy Jane, "Feature Selection and Classification of Clinical Datasets Using Bioinspired Algorithms and Super Learner," *Comput Math Methods Med*, vol. 2021, 2021, doi: 10.1155/2021/6662420.

[17] Y. Yin *et al.*, "IGRF-RFE: a hybrid feature selection method for MLP-based network intrusion detection on UNSW-NB15 dataset," *J Big Data*, vol. 10, no. 1, 2023, doi: 10.1186/s40537-023-00694-8.

[18] A. Gupta, A. Bhatt, V. P. Singh, A. Mathur, A. V Waghmare, and S. Joglekar, "Disturbance Detection in Power Quality Data using Principle Component Analysis, Recursive Feature Elimination and Logistic Regression," in *2025 IEEE 5th International Conference on Sustainable Energy and Future Electric Transportation (SEFET)*, Jul. 2025, pp. 1–4. doi: 10.1109/SEFET65155.2025.11255423.

[19] M. Riajuliislam, K. Z. Rahim, and A. Mahmud, "Prediction of Thyroid Disease(Hypothyroid) in Early Stage Using Feature Selection and Classification Techniques," *2021 International Conference on Information and Communication Technology for Sustainable Development, ICICT4SD 2021 - Proceedings*, pp. 60–64, 2021, doi: 10.1109/ICICT4SD50815.2021.9397052.

[20] Y. Deng, Y. Wei, and Y. Li, "Credit Risk Evaluation Based on Data Mining and Integrated Feature Selection," *ICSPCC 2020 - IEEE International Conference on Signal Processing, Communications and Computing, Proceedings*, pp. 8–11, 2020, doi: 10.1109/ICSPCC50002.2020.9259483.

[21] Y. Deng, Y. Wei, and Y. Li, "Credit Risk Evaluation Based on Data Mining and Integrated Feature Selection," *ICSPCC 2020 - IEEE International Conference on Signal Processing, Communications and Computing, Proceedings*, pp. 8–11, 2020, doi: 10.1109/ICSPCC50002.2020.9259483.

[22] M. Chemmakha, O. Habibi, and M. Lazaar, "Improving Machine Learning Models for Malware Detection Using Embedded Feature Selection Method," *IFAC-PapersOnLine*, vol. 55, no. 12, pp. 771–776, 2022, doi: 10.1016/j.ifacol.2022.07.406.

[23] K. M. Sujon, I. U. Ahmed, M. M. I. Molla, N. Jahan, and S. K. Utsho, "The Impact of Recursive Feature Elimination and Information Gain on Machine Learning Models for Student Performance Prediction," in *2025 International Conference on Quantum Photonics, Artificial Intelligence, and Networking (QPAIN)*, Jul. 2025, pp. 1–6. doi: 10.1109/QPAIN66474.2025.11172062.

[24] S. K and J. A. Arul Jothi, "Disease Prediction and Biomarker Identification using Recursive Feature Elimination and Deep Forest classifier from Human Gut Metagenomic Data," in *2025 IEEE International Symposium on Technology and Society (ISTAS)*, 2025, pp. 1–7. doi: 10.1109/ISTAS65609.2025.11269618.

[25] C. Yi, M. Guo, Y. Lei, and T. Qiu, "Research on the Application of Ensemble Learning Model Based on Recursive Feature Elimination in Engine Efficiency Prediction," in *2025 11th International Conference on Energy Materials and Electrical Engineering (ICEMEE)*, Aug. 2025, pp. 425–429. doi: 10.1109/ICEMEE66610.2025.11276539.

[26] Arini, Hidayat, and A. Fiade, "Performance Evaluation of RFE-CV, Class Weights, and ADASYN for Network Attack Classification on UNSW-NB15 Using Ensemble Learning," in *2025 13th International Conference on Cyber and IT Service Management (CITSM)*, 2025, pp. 1–7. doi: 10.1109/CITSM67730.2025.11291428.