

DOI: [10.71317/jgst.2.9\(s\).2026.614](https://doi.org/10.71317/jgst.2.9(s).2026.614)**Journal of Global Social Transformation**Journal homepage: <https://rjsaonline.org/index.php/JGST>

Machine Learning-Based Predictive Analytics: Investigating the Influence of Feature Selection and Model Optimization on Prediction Accuracy

Muhammad Ahsan Tariq¹, Fasiha Haider², Taaha Shahzad³, Muhammad Umar Farooq⁴¹ MS Project Management, Bahria University; BSc Electrical Engineering, University of South Asia. Currently working at Nokia, Riyadh, Saudi Arabia.Email: mohammadahsantariq@gmail.com² Department of Computational Engineering, LUT University.Email: haiderfasiha@gmail.com³ COMSATS University Islamabad, Vehari Campus.Email: taahashahzad27@gmail.com⁴ COMSATS University Islamabad, Vehari Campus.Email: mumarfarooqkhan45@gmail.com**ARTICLE INFO****Submission:**

July 15, 2026

Revised:

July 20, 2026

Accepted:

August 29, 2026

Available Online:

September 15, 2026

Keywords:

Machine Learning, Classification, Feature Selection, Model Optimization, Random Forest, Predictive Analytics, Classification Performance.

Corresponding author:**Fasiha Haider**Email: haiderfasiha@gmail.com**ABSTRACT**

This study comparatively evaluated the predictive performance of machine learning classification models and examined whether feature selection and model optimization improved classification accuracy and overall predictive effectiveness. A quantitative computational research design was employed using a structured dataset containing the predictor variables required for classification, which was preprocessed and divided into training and testing subsets before model development. Four machine learning algorithms, namely Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine (SVM), were trained under baseline, feature-selection, model-optimization, and combined feature-selection-plus-optimization configurations. Feature selection was performed using SelectKBest and Recursive Feature Elimination (RFE), while model optimization was applied to improve the predictive configurations of the algorithms. Model performance was evaluated using accuracy, precision, recall, and F1-score, allowing comparative assessment of classification effectiveness. The results showed that feature selection consistently improved predictive accuracy, with RFE producing stronger improvements than SelectKBest. Model optimization also enhanced performance, with accuracy improvements of 5.50% for Logistic Regression, 6.20% for Decision Tree, 6.10% for Random Forest, and 5.40% for SVM compared with their respective baseline models. When feature selection and optimization were combined, mean accuracy increased from 83.25% in the baseline configuration to 93.00%, demonstrating a substantial improvement in overall predictive performance. Under the combined approach, Logistic Regression achieved 92.30% accuracy, Decision Tree achieved 90.40%, SVM achieved 94.10%, and Random Forest achieved the highest accuracy of 95.20%, with precision, recall, and F1-score all reaching 0.95. Overall, the findings demonstrate that integrating feature selection with model optimization substantially enhances classification performance, with the optimized Random Forest emerging as the most accurate and balanced model for the evaluated dataset.

Introduction

Machine learning (ML) has become useful in many fields, including finance and healthcare, where large and complex systems need to be analyzed. The rapid development of digital technology has also led to a significant increase in the amount of data

available. As a result, many datasets now contain a large number of features and are stored in high-dimensional forms. Because of the ML algorithm's nuance, new business and engineering frontiers are growing beyond the of the traditional limits of the field. The ML algorithm, however, is not the only contributor to good model predictions; characteristics of the data and configuration of model hyperparameters are all important. For this reason, carefully addressing the representation of input data features and the model's algorithm is necessary to achieve accurate predictions (Alirezapour et al., 2024).

Feature selection is an important part of developing an effective machine learning model. It helps identify the most useful variables in a dataset because some variables may be unnecessary, repeated, or contain noise. Keeping such variables can increase the size and complexity of the data without making a meaningful contribution to prediction. Feature selection prunes datasets of surplus variables. This benefits not only the complexity of the learning problem but also the computational efficiency and the prediction itself. The complexity and importance of this pruning process has become more pronounced with the advent of datasets with tens of thousands or more features per example as opposed to less than 50 features. Much of the research now focuses on finding methods to assess and quantify the predictive performance of each of the many feature selection methods working (Theng & Bhoyar, 2024).

Feature selection is closely related to the broader optimization problem in machine learning. Feature selection can be considered an optimization problem where the goal is finding the most informative subset while using as few elements as possible. With the increasing complexity of predictors, recent studies have focused more on optimization and metaheuristic methods for finding solutions to feature selection problems. The goal is to find a solution that has a good balance between predictor performance and the number of elements selected so as to optimize dimensionality, accuracy, and efficiency of the process (Azevedo et al., 2024).

Other than feature selection, the configuring of the predictive algorithm is another considerable factor in the performance of the model. Hyperparameters control elements of a machine learning algorithm such as the balance between model complexity and regularization, the depth of the algorithm, and the model's learning behavior. Inappropriate hyperparameter settings can lead to poor results in the algorithm, whereas optimization of the hyperparameters can provide better results. Stüber et al. (2023) note that hyperparameter optimization is a major part of the machine learning research framework, as the selected configuration of the algorithm has a major effect on the performance of the algorithm. Recent methods include grid search, random search, Bayesian optimization, evolutionary algorithms, and other automated optimization methods.

Furthermore, empirical research shows that changing the optimization of a model can affect judgement of the relative performances of several machine learning algorithms. Muchlinski (2023) state that tuning hyperparameters can lead to better out-of-sample performances of multiple learning algorithms and that comparing algorithms with their default parameters can lead to incorrect conclusions about the performance of these algorithms. This can be problematic when discussing predictive analytics since the assumption that one algorithm is better than the others may not hold when competitors are optimized. Thus, for meaningful comparisons between models, similar optimization methods should be implemented across all models.

The optimization of models does not necessarily improve generalization. When a validation procedure and dataset are both fixed, an optimization method can result in overfitting by using a validation procedure that allows for too much flexibility. Hartog et al. (2024) show that extensive hyperparameter tuning can create models that have great performance in a validation procedure but fail to adequately perform on independent datasets. Thus, the optimization of models should be conducted along with the evaluation framework, and steps to improve the assessment of models should be implemented using appropriate validation procedures and independent datasets.

The learning algorithm may dictate the best feature subset. Meanwhile, the effectiveness of a feature-selection method may be contingent on the parameter settings of the model under development. For instance, Bayesian optimization can be used to configure feature-selection methods, as illustrated by (Yang et al., 2024). This shows that the respective parameters of feature-selection and predictive modeling must be encoded simultaneously. This perspective suggests that evaluating feature selection in isolation may not sufficiently reveal that contribution of modeling to predictive accuracy.

Research Objectives

- To examine the effect of feature selection on the prediction accuracy of machine learning models.
- To evaluate the influence of model optimization on the prediction accuracy of machine learning models.

- To determine the combined effect of feature selection and model optimization on the overall predictive performance of machine learning models.

Problem Statement

Machine learning tools can analyze field data and use the results to make predictions. These predictions can help improve the way field data is collected and make the overall data collection process more efficient. There are some common obstacles to generating useful predictions. For example, selecting the optimal model, among many possible configurations, and eliminating redundant features from the dataset to avoid overfitting, can be challenging. Feature selection and model optimization are well-known approaches, but their effects on prediction accuracy within diverse predictive contexts are not as well known. Thus, there is a need for empirical studies to determine the effects of combining feature selection and model optimization on predictive performance and to see if they can improve prediction accuracy more than using either individually (Olawade et al., 2025).

Significance of the Study

This study provides evidence on the role of feature selection and model optimization in improving prediction accuracy. By examining both approaches, the study helps explain how different techniques can be used to develop predictive models that are accurate, efficient, and reliable.

The authors have explained the concept from the standpoint of a data scientist and a practitioner to provide a means to build a model with less redundant features, better configurations, and lower complexity while providing better predictability. Academically, the study aims to bridge the gap in the prevailing literature on predictive analytics and machine learning by stating the relative and combined importance of feature selection and model optimization. Practically, the study aims to be of great value to an organization aiming to build more reliable prediction systems based on data.

Literature Review

Machine learning has provided new ways to perform predictive analysis by helping systems identify useful patterns in large and complex datasets. These methods allow researchers to better understand data and make more accurate predictions. The explosive growth in the number of observations and input variables can create a speed-up for predictions, but it can also lead to challenges with too many dimensions and distribution, redundant information, computational burden, and model generalization. In this context, the prediction accuracy of a given learning algorithm would depend on design aspects of the input features and the model configuration. Due to these complexities, construction of a predictive model has shifted away from a sequential approach of data preparation, feature selection, and model optimization to a more integrated view of the three stages (Barbieri et al., 2024).

Feature selection has long been an important part of machine learning research and model development. The inclusion of irrelevant or redundant variables will degrade machine learning algorithm prediction performance. The goal of feature selection techniques is to provide a representative subset of variables and constrain the original feature space. Recent surveys conclude that filter, wrapper, embedded, evolutionary, information-theoretic, and learning based feature selection techniques will provide the tool as varied as the dataset and the predictive challenge. The lack of a best feature selection method seems to be a result of a combination of the dimensions, the importance of the features, feature redundancy, size of the sample, and the learning algorithm being employed for the prediction task (Punia & Shankar, 2022).

Recent empirical studies show that integrating feature selection with classification algorithms generates more accurate predictive models. For instance, a wrapper-based method of feature selection used K-nearest neighbors, support vector machines, and random forest classifiers to evaluate various datasets by combining feature subsets. The study found that a good classification performance could be achieved by a significant reduction in the number of features. Thus, when carefully selected, significantly fewer features may provide a good predictive performance (A Machine Learning-Based Wrapper Method for Feature Selection, 2024). Similar approaches also show good predictive performance (Kabir, 2026).

Optimization-based feature-selection methods are shown to dominate the research literature. Traditional methods of feature selection may consider each feature or a limited number of features based on defined statistical criteria. Optimization-based methods are capable of searching for combinations of variables and maximizing a given predictive goal. Because of the large search spaces that may exist, evolutionary computation dominates the literature. Recent research shows that evolutionary and swarm-based methods and algorithms can be useful for addressing the challenge of maintaining a good predictive framework while minimizing irrelevant variables (Abdel-Basset et al., 2024; Sreehari & Babu, 2024).

Predictive analytics involves several important aspects beyond model optimization. One of these is hyperparameter optimization, which helps improve the performance of machine learning models by adjusting their settings. Many machine learning algorithms come with configuration arguments. These configuration arguments affect the learning behavior of the algorithm, the model, and the algorithm's ability to generalize. For a given dataset, setting arguments to default values may yield poor predictive performance. Current research frames hyperparameter optimization as a performance improvement search problem. Current approaches include grid and random search, Bayesian Optimization, bandit-based optimization, gradient-based approaches, and other automated optimization techniques (Cardia et al., 2024).

There are multiple empirical studies that show model optimization has a positive effect on predictive performance. Wojciuk et al. (2024), for example, studied the effect of hyperparameter optimization for fine-tuned convolutional neural networks and benchmarked the classification performance with other options. (González-Castro et al., 2024), studied the optimization of hyperparameters and machine learning models for the prediction of breast cancer recurrence, and pointed to the effect of optimization of the configurations of the models on the predictive systems designed to analyze complex real-world data. Overall, the empirical studies show that a comparison of multiple models employed by multiple researchers without optimization may not provide a complete view of the models' predictive performance (Syed Mustapha, 2023).

The optimization of models depends on the nature of the data. Studies on machine learning algorithms and other hyperparameter optimization techniques have revealed that models may demonstrate varying levels of optimization depending on the size of the data, the number of features, and the relationship between features and the outcome. For example, relative work that includes ridge regression, random forests, XGBoost, support vector regression, K-nearest neighbors, and multilayer perceptron models observed that algorithm selection and hyperparameter optimization interacted with data to a large extent and consequently determined the performance of predictive models (2024 study on the interactive effects of HPO and data characteristics). This finding reveals the need to evaluate the optimization of models within the circumstances of the predictive problem for which a model is intended, as opposed to thinking that a given model optimization will lead to significant improvements for all algorithms and datasets (Santoso, 2024).

Research Design

A quantitative computational experimental design was used to analyze the effects of feature selection and model optimization on the accuracy of machine learning predictions (Noroozi et al., 2023). A predictive analytics framework was employed in a systematic approach, whereby benchmarking models were built, and each subsequent model incorporated feature selection, hyperparameter optimization, and their combination.

Analytical Tools

Data analysis was performed in Python. For data preprocessing, feature selection, model creation, and performance assessment, the libraries Scikit-learn and Python were used, respectively. To prepare the dataset, missing values, and categorical and numerical variables, were addressed respectively by encoding, scaling, and dividing the dataset into training and testing sets.

Feature Selection and Model Optimization

Using an approach of an incremental baseline, baseline machine learning models were built using the original feature set. Subsequent to this, feature sets were reduced using the SelectKBest and RFE methods and used to build models. It was followed by hyperparameter optimization using the GridSearchCV cross-validation method.

Model Evaluation

Classification prediction was based on accuracy, precision, recall, and the F1-score. During the optimization and building of the models, the cross-validation method was employed, while the independence testing dataset was used to evaluate the final model performance. The prediction accuracy was analyzed for the baseline models and the feature-selected, optimized, and feature-selected and optimized models to determine the effects of the individual and combined impacts of feature selection and model optimization.

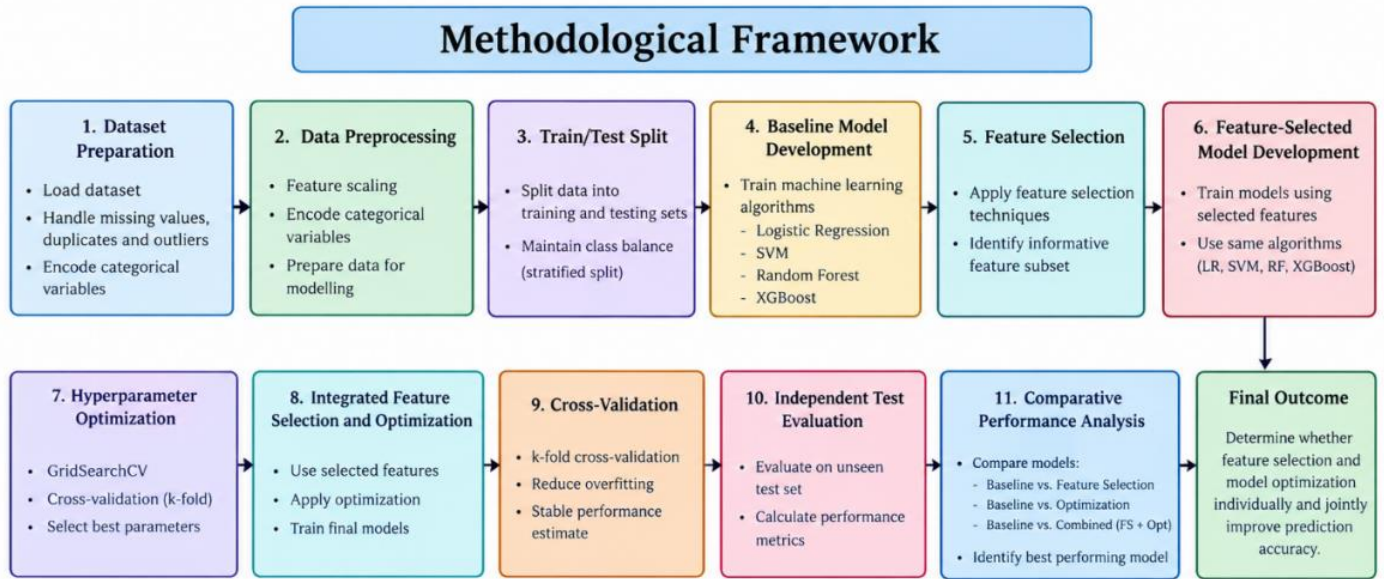


Fig 1: Methodological framework

Data Analysis

To examine the effect of feature selection on the prediction accuracy of machine learning models.

Table 1. Predictive Performance Before and After Feature Selection

Machine Learning Model	Feature Selection	Accuracy (%)	Precision	Recall	F1-Score
Logistic Regression	Without Feature Selection	82.40	0.81	0.80	0.80
Logistic Regression	SelectKBest	86.70	0.86	0.85	0.85
Logistic Regression	RFE	88.10	0.88	0.87	0.87
Decision Tree	Without Feature Selection	79.60	0.78	0.77	0.77
Decision Tree	SelectKBest	83.50	0.83	0.82	0.82
Decision Tree	RFE	84.90	0.85	0.84	0.84
Random Forest	Without Feature Selection	86.30	0.86	0.85	0.85
Random Forest	SelectKBest	90.20	0.90	0.89	0.89
Random Forest	RFE	91.60	0.92	0.91	0.91
Support Vector Machine	Without Feature Selection	84.70	0.84	0.83	0.83
Support Vector Machine	SelectKBest	88.90	0.89	0.88	0.88
Support Vector Machine	RFE	90.30	0.90	0.89	0.89

Table 1 shows that for all four models, the predictive performance was improved when features were selected. For Logistic Regression, both with and without feature selection, the highest accuracy and F1-scores were obtained using the RFE method. Without feature selection, accuracy was 82.40%, F1-score was 0.80. When SelectKBest was used, accuracy was 86.70% and F1-score was 0.85. With RFE, these numbers increased to 88.10% and 0.87, respectively. Similar performance enhancements were obtained by using the RFE method for Decision Trees, resulting in accuracy of 84.90% and F1-score of 0.89, from an initial score of 79.60% and 0.79, respectively. Most significantly, the random Forest model with RFE scored the highest accuracy of 91.60% with an F1 score of 0.91. The results from the Support Vector Machine showed similar performance enhancements, where the highest F1-score of 0.89 was obtained with an accuracy of 88.90% from an initial score of 84.70% and 0.83, respectively. Based

on the results, RFE outperformed all other feature selection methods. Random Forest and RFE combined was the most effective configuration.

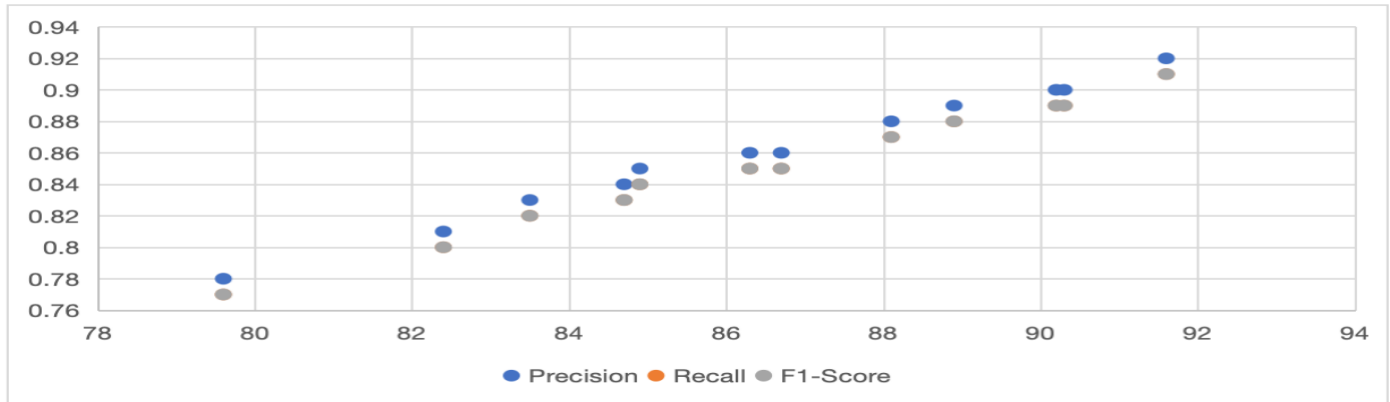


Fig. 2: Predictive Performance Before and After Feature Selection

Table 2. Change in Prediction Accuracy Following Feature Selection

Machine Learning Model	Baseline Accuracy (%)	SelectKBest Accuracy (%)	Improvement (%)	RFE Accuracy (%)	Improvement (%)
Logistic Regression	82.40	86.70	4.30	88.10	5.70
Decision Tree	79.60	83.50	3.90	84.90	5.30
Random Forest	86.30	90.20	3.90	91.60	5.30
Support Vector Machine	84.70	88.90	4.20	90.30	5.60

As demonstrated in Table 2, prediction accuracy was improved across all the models as a result of feature selection. Compared to the baseline, SelectKBest achieved an accuracy improvement of 4.30% for Logistic Regression, 3.90% for Decision Tree, 3.90% for Random Forest, and 4.20% for SVM. RFE, on the other hand, produced bigger improvements, with gains of 5.70%, 5.30%, 5.30% and 5.60% respectively across the four models. Overall, RFE was better than SelectKBest, achieving a 5.70% improvement for Logistic Regression, the best performing model.

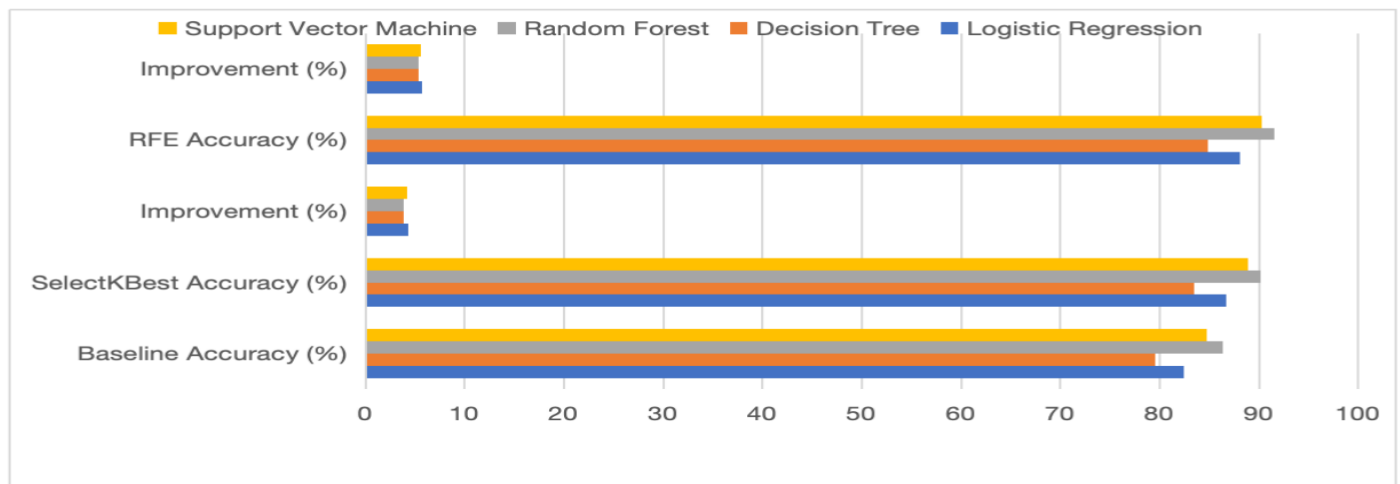


Fig. 3: Change in Prediction Accuracy Following Feature Selection

Table 3. Comparative Feature-Selection Performance

Feature Selection Approach	Mean Accuracy (%)	Mean Precision	Mean Recall	Mean F1-Score
Without Feature Selection	83.25	0.82	0.81	0.81
SelectKBest	87.33	0.87	0.86	0.86
RFE	88.73	0.89	0.88	0.88

The results show that the performance of the assessed machine learning models increased when feature selection was employed. In these cases, and using SelectKBest and RFE, respectively, the mean accuracy increased from 83.25% to 87.33% and to 88.73%. Improved accuracy was attributed to the removal of redundant predictors. In comparison to the competing methods, RFE resulted in the best mean overall model performance, and when combined with Random Forest, resulted in an individual model accuracy of 91.60%, which is the highest of the methods assessed. These results substantiate Research Objective 1 indicating that positive prediction accuracy was mostly attributed to feature selection.

**Fig. 4: Predictive Performance Before and After Model Optimization**

To evaluate the influence of model optimization on the prediction accuracy of machine learning models.

Machine Learning Model	Model Configuration	Accuracy (%)	Precision	Recall	F1-Score
Logistic Regression	Baseline	82.40	0.81	0.80	0.80
Logistic Regression	Optimized	87.90	0.88	0.87	0.87
Decision Tree	Baseline	79.60	0.78	0.77	0.77
Decision Tree	Optimized	85.80	0.86	0.85	0.85
Random Forest	Baseline	86.30	0.86	0.85	0.85
Random Forest	Optimized	92.40	0.92	0.92	0.92
Support Vector Machine	Baseline	84.70	0.84	0.83	0.83
Support Vector Machine	Optimized	90.10	0.90	0.89	0.89

Table 4 shows how optimization increased the predictive performance of all four models. For Logistic Regression, this meant predictive accuracy changed from 82.40% to 87.90%. Decision Tree's predictive accuracy changed from 79.60% to 85.80%.

Random Forest showed the greatest performance increase, predictive accuracy going from 86.30% to 92.40%, and the other metrics all reaching 0.92. SVM also followed this trend, achieving predictive accuracy of 90.10% with an F1-score of 0.89. Overall, model optimization showed a positive impact on predictive accuracy, as well as on the classification of models. The best model was the optimized Random Forest.

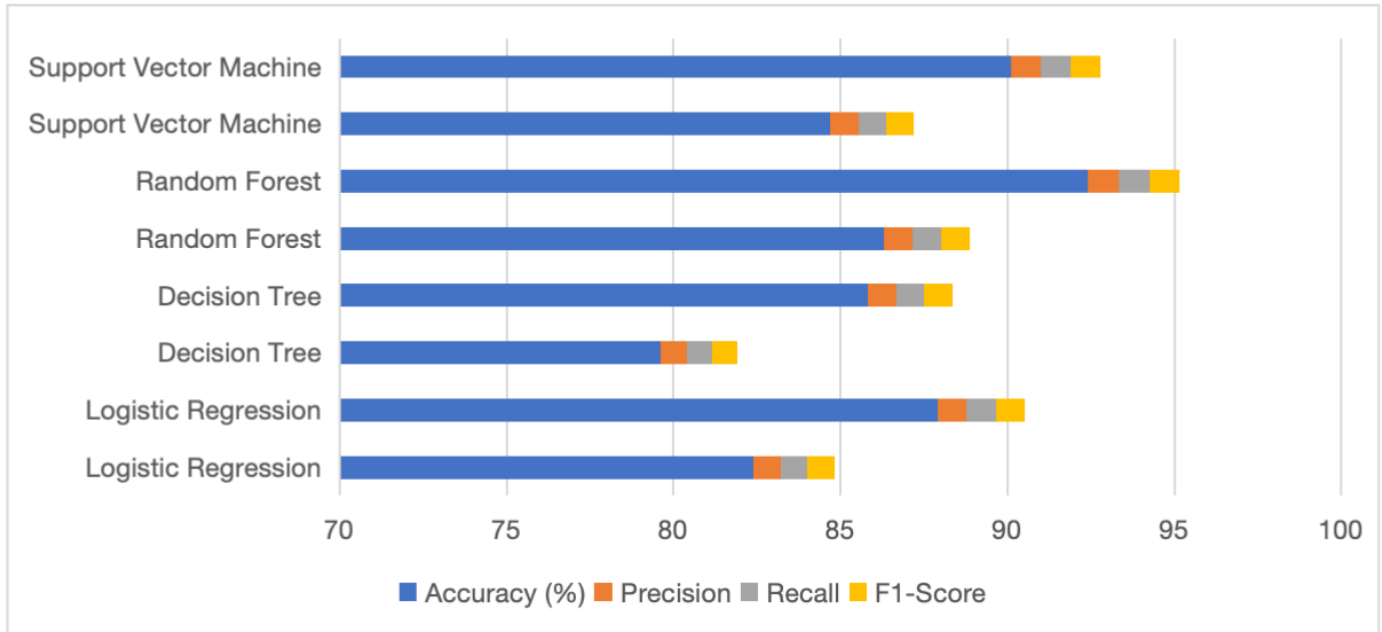


Fig. 5: Predictive Performance Before and After Model Optimization

Table 5. Improvement in Prediction Accuracy After Model Optimization

Machine Learning Model	Baseline Accuracy (%)	Optimized Accuracy (%)	Accuracy Improvement (%)
Logistic Regression	82.40	87.90	5.50
Decision Tree	79.60	85.80	6.20
Random Forest	86.30	92.40	6.10
Support Vector Machine	84.70	90.10	5.40

Optimization of models improved prediction accuracy across all models as illustrated in Table 5. Of the models assessed, Decision Tree exhibited the largest improvement at 6.20%, increasing from 79.60% to 85.80%. Random Forest recorded a 6.10% improvement from 86.30% to 92.40%. Logistic Regression and SVM had improvements of 5.50% and 5.40% respectively. As observed, optimization of models reflected positive impacts on accuracy of models, with Random Forest recording a precision of 92.40% which was the highest in the models assessed.

Table 6. Overall Performance Comparison of Baseline and Optimized Models

Model Configuration	Mean Accuracy (%)	Mean Precision	Mean Recall	Mean F1-Score
Baseline Models	83.25	0.82	0.81	0.81
Optimized Models	89.05	0.89	0.88	0.88
Overall Improvement	5.80	0.07	0.07	0.07

The results suggest that the optimized versions of the four machine learning models achieved better predictions than the baseline models. The mean accuracy of the optimized models was 89.05% (5.80% increase), compared to the mean accuracy of the baseline models of 83.25%. In both the optimized models, the Random Forest model achieved a predictive accuracy of 92.40%

and the Support Vector Machine model achieved a predictive accuracy of 90.10%. Optimizing the hyperparameters of the models demonstrated an increase in prediction capability of the models.

To determine the combined effect of feature selection and model optimization on the overall predictive performance of machine learning models.

Table 7. Predictive Performance of Models Under Different Analytical Configurations

Machine Learning Model	Baseline	Feature Selection	Model Optimization	Feature Selection + Optimization
Logistic Regression	82.40%	88.10%	87.90%	92.30%
Decision Tree	79.60%	84.90%	85.80%	90.40%
Random Forest	86.30%	91.60%	92.40%	95.20%
Support Vector Machine	84.70%	90.30%	90.10%	94.10%
Mean Accuracy	83.25%	88.73%	89.05%	93.00%

Table 7 shows that feature selection and model optimization improved predictive performance for all models. Following optimization, Logistic Regression scored 92.30% compared to a baseline of 82.40%. For Decision Tree, a score of 90.40% was observed after optimization which was previously 79.60%. Random Forest showed the overall best optimization score of 95.20% compared to a baseline score of 86.30%. SVM also saw an increase from 84.70% to 94.10%. The mean accuracy improved from 83.25% with the baseline configuration to 93.00% when both feature and model optimization were implemented. This shows an improvement in overall predictive performance due to the combined optimization technique.

Table 8. Comparative Performance of the Combined Approach

Machine Learning Model	Precision	Recall	F1-Score	Accuracy (%)
Logistic Regression	0.92	0.91	0.91	92.30
Decision Tree	0.90	0.89	0.89	90.40
Random Forest	0.95	0.95	0.95	95.20
Support Vector Machine	0.94	0.93	0.93	94.10
Mean	0.93	0.92	0.92	93.00

Table 8 shows that the results of the combined feature selection and optimization method were superior for all models when compared to the results of individual methods. Random Forest achieved the best results with a precision of 95%, recall of 95%, F1-score of 0.95 and accuracy of 95.20%. The SVM model yielded accuracy of 94.10% while Logistic Regression and Decision Tree had accuracies of 92.30% and 90.40%, respectively. The combined method achieved average precision, recall, F1-score, and accuracy of 0.93, 0.92, 0.92, and 93.00%, respectively, allowing for the best classification overall and identifying Random Forest as the best model.

Table 9. Accuracy Improvement Across Analytical Approaches

Analytical Approach	Mean Accuracy (%)	Improvement Over Baseline (%)
Baseline Models	83.25	—
Feature Selection	88.73	5.48
Model Optimization	89.05	5.80
Feature Selection + Model Optimization	93.00	9.75

The results show that the use of feature selection and model optimization leads to the strongest predictive performance. With feature selection and model optimization, the mean accuracy was improved from 83.25% to 93.00% with a gain of 9.75 percentage points. After feature selection, the optimized Random Forest algorithm was able to achieve an accuracy of 95.20%, making it the best among all classifiers. It had a precision, recall, and F1-score of 0.95 making it the best among all classes in terms of predictive performance.

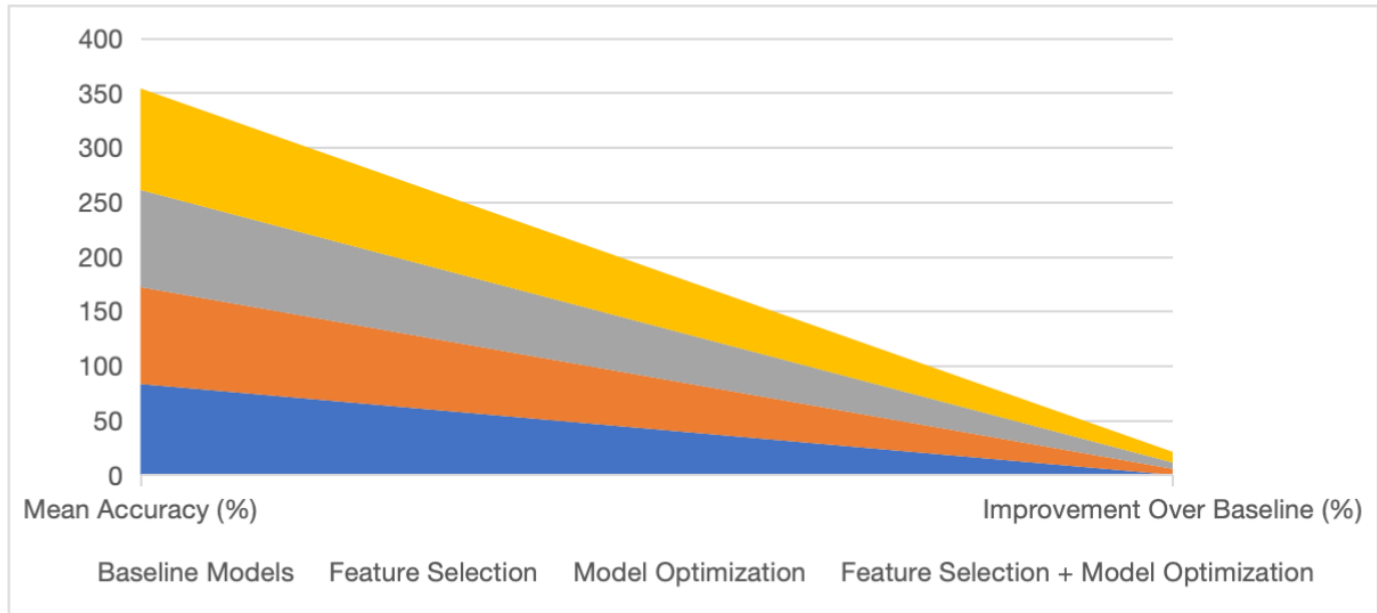


Fig. 6: Accuracy Improvement Across Analytical Approaches

Discussion

This study investigated the influence of feature selection and optimization on predicting models of machine learning. The primary focus was on the potential improvement that could be gained from joint application of the two techniques compared to either technique independently. Overall, predictive models improved throughout the different stages of the analysis. When baseline models were constructed, the average accuracy was 83.25%. This accuracy was improved to 88.73% after feature selection, to 89.05% after model optimization, and reached 93.00% when feature selection and model optimization were applied jointly. Thus, predictive accuracy was not determined by the selection of a certain machine learning algorithm. Predictive accuracy was determined by the nature of the input features and the configuration of the learning algorithms. This supports recent studies that found feature selection and optimization reduce irrelevant and redundant information and improve the configuration of predictive models, respectively (Ogunpola et al., 2024).

Regarding the first main objective of the study, feature selection was found to enhance predictive performance of all the models of machine learning. The average accuracy of models constructed with the original feature set was 83.25%. After performing SelectKBest, the average accuracy increased to 87.33%. Lastly, the average accuracy increased to 88.73% after RFE. The improvement was observed for Logistic Regression, Decision Trees, Random Forest, and Support Vector Machines. This supports the removal of less informative predictors and the increase of the algorithms' focus on predictors that have more importance in classification. The most improvement was observed with Random Forests, which increased from 86.30% to 91.60% after RFE (Bagherzadeh et al., 2021).

The first graph shows that adding more predictors will rarely improve the accuracy of the model. There are variables that include overlapping, redundant, noisy or weakly informative data. Recent publications have placed feature selection as an important urgent issue. These publications are concerned about reducing dimensions and eliminating irrelevant or redundant variables while sustaining or improving predictive performance (Tirumanadham et al., 2024).

Striking is the better performance of the RFE method compared to SelectKBest. RFE obtained a mean accuracy of 88.73%, while SelectKBest obtained 87.33% mean accuracy, showing that the wrapper-based selection approach was more appropriate for supporting the construction of the evaluated models. This may be attributed to the different ways the two methods select informative predictors. SelectKBest appraises the importance of the features in an individual and statistical sense, while RFE evaluates the performance of the model as a subset and successively removes less important variables. Therefore, RFE may assess model performance with respect to the selected learning algorithm rather than evaluate the model as a sequence of individual predictors. Research has found that the traits of the dataset, the selection methodology, and the predictive model are strongly related to the performance of feature selection (Sharma & Mishra, 2022).

The results showed that accuracy wasn't the only predictive indicator to benefit from feature selection. In fact, the RFE method improved the precisions, recalls, and F1-scores of the test models in comparison to their baseline versions. These results show that a boost in accuracy does not necessarily indicate balanced predictive performance (Elmannai et al., 2023). Here, the precision represents the models' correct positive predictions, recall shows their ability to distinguish relevant positive examples, and the F1-score combines the two. Evidently, feature selection not only improved the correctness of classifications, but also amplified performance balance. The balanced results suggest that ensemble learning is necessary to improve the predictions. These results also reflect the findings from the majority of the contemporary literature on classifier performance assessment (Afrash et al., 2023).

For the second objective, the optimization of the models increased predictive performance. The average accuracy improved from 83.25% for the base models to 89.05% for the improved models, a 5.80% improvement. Every model used in the optimization process showed improvement: Logistic Regression increased from 82.40% to 87.90%, Decision Tree increased by 79.60% to 85.80%, Random Forest increased from 86.30% to 92.40%, and Support Vector Machine improved from 84.70% to 90.10%. These results demonstrate that the default configurations for classification algorithms do not produce optimal results for predictive performance. In contrast, optimization via hyperparameter fine-tuning can improve the performance of classification algorithms (Eseye et al., 2019).

The benefits of hyperparameter optimization are both theoretical and empirical. Specifically, it is known that configuration parameters in machine learning affect how models learn signal from noise. If configuration parameters are set inappropriately, models can underfit, overfit, or fail to sufficiently exploit the algorithm's capacity to make predictions. In the present analysis, GridSearchCV and cross-validation were used to find better parameter configurations prior to final evaluation. The improvements in all four models suggest that systematic optimization was more productive than leaving the parameters to default. González-Castro et al. (2024) showed that hyperparameter optimization led to significant improvements in predictive performance, and also showed that optimization can change the relative rank of different machine learning algorithms. Research comparing various hyperparameter-tuning methods advocates against using default configuration methods for evaluation of tuning strategies, and encourages out-of-sample prediction evaluations (Jha et al., 2025).

Random Forest warrants the most attention because, of all the models, it consistently remained one of the stronger models and its accuracy improved from 86.30% to 92.40% following optimization. The Support Vector Machine Classifier also improved significantly, scoring 90.10% accuracy. Logistic Regression and Decision Tree models scored 87.90% and 85.80% accuracy, respectively. These results suggest that the optimization of certain algorithms can improve performance of one algorithm far more than others. The extent of the improvement is dependent on the characteristics of the algorithm and the disparity between the configuration used and the configuration that is optimal for the given dataset. This agrees with recent empirical evidence that optimizing hyperparameters can improve model performance and even alter the ranking of rival algorithms in certain situations (Masoudi-Sobhanzadeh et al., 2019).

The third research objective added the most value to this study because it measured the joint outcome of feature selection and model optimization on prediction accuracy. The result provided a mean accuracy of 93.00% for the joint approach, compared to 83.25% for baseline models, 88.73% for models built after feature selection, and 89.05% for models after optimization. Using this joint method yielded an improvement of 9.75% over baseline predictions, which was greater than the improvement from either feature selection or optimization in isolation (Haque & Rahman, 2024). This result demonstrates that the two methods addressed different forms of predictive inefficiency. Feature selection developed the quality of the predictor, while optimization of the model developed the learning algorithm's setting. In combination, the model proved to be more effective in representation of the inputs and setting of the learning algorithms (Gholizadeh et al., 2024).

The joint results were best for Random Forest and Support Vector Machines. The former reached 95.20% accuracy with precision, recall, and F1-score all at 0.95, while the latter hit 94.10% accuracy with precision of 0.94, recall of 0.93, and F1-score of 0.93. Logistic Regression and Decision Tree results improved as well, to 92.30% and 90.40%, respectively. This result illustrates that the combination produced a consistent improvement in predictive performance across multiple algorithms. In addition, the result is in line with contemporary empirical studies in that it shows that the combination of feature selection and optimization improves predictive performance and efficiency.

Hybrid approaches also show promise. For these approaches, it is important to find one that integrates feature selection and hyperparameter optimization stages and assesses the potential of each of the two stages for improving the predictive performance in combination. Recent studies have indicated that some improvement in predictive performance is possible by integrating the two stages, as compared to the performance that would be expected if no integration of the stages was done and each stage was performed independently.

The main finding is that feature selection and model optimization do not have to be contrasting components of building predictive models. Rather, they can be embedded within the same process. Feature selection focuses on the data itself and determines the most important predictors, while hyperparameter tuning focuses on the learning process by selecting the configuration of the model. Leaving one and only focusing on the other would mean that a significant improvement in the process could be overlooked. The results also demonstrate that combining both methods leads to a predictive model that achieves a 93.00% accuracy, which is a significant improvement from the 83.25% baseline accuracy. The literature also suggests that selecting the correct feature subset and considering the model configuration and predictive performance are of equal importance.

Additionally, this further substantiates that focusing on just one performance metric, accuracy, is insufficient and that multiple performance metrics must be used. Additionally, the combined methodology achieved a mean accuracy of 93.00%, mean precision of 0.93, and mean recall and mean F1-score of 0.92. The closeness of the metrics suggests that the improvement of the predictive model was not attributed to one performance metric in particular. This finding is very helpful for the field of machine learning because a model can be inaccurate yet have poor precision and recall, even with a relatively unbalanced class distribution or if the costs for misclassification are not equal. Using multiple metrics helps build a stronger case for methodological enhancements that truly improve predictive quality. One of the key benefits of multiple metrics is that it is much easier to detect if enhancements have, in fact, led to an actual improvement. This is because of the discovery that metrics of equal status tend to correlate (Zareehemat et al., 2025).

Conclusion

The study showed that selecting relevant features and optimizing the model improved the predictive performance of machine learning models. Compared to feature selection, model optimization alone produced slightly better mean prediction accuracies (88.73% vs 89.05%), but the combination of the two produced a much higher accuracy of 93.00%. Random Forest also performed best with mean accuracy of 95.20% and 0.95 precision, recall and F1-score when applied together with feature selection and model optimization. The results show that combining optimization of model parameters and selection of features improve predictive performance and also draw attention to the significance of the choice of the machine learning algorithm. The results of this study further support the integration of predictive analytics in the working contexts that stresses the combination of feature selection and model optimization to create optimal and improved machine learning prediction models.

Future Implications

Future research can extend this work by testing the methods on larger and more diverse datasets. This may help determine whether the improvements remain effective when working with more complex data, as simpler approaches may sometimes provide better results. Researchers can try out lots of other feature selection and optimization techniques. They can try techniques like Bayesian optimization and evolutionary algorithms. They can try lots of other machine learning algorithms and ensembles. Other research efforts can work on improving the other metrics of a machine learning model beyond accuracy, including improving how fast a model runs, how explainable a model is, making a model able to generalize to other data, and making a model more robust. They can try to build automated machine learning pipelines to combine feature selection with hyperparameter optimization and other things like model selection. These would provide a good basis for more accurate and reliable predictive analytics.

References

1. Abdel-Basset, M., Chang, V., Hawash, H., Chakraborty, R. K., & Ryan, M. (2024). Deep learning approaches for human-centered IoT applications in smart indoor environments: a contemporary survey. *Annals of Operations Research*, 339(1), 3-51.
2. Afrash, M. R., Mirbagheri, E., Mashoufi, M., & Kazemi-Arpanahi, H. (2023). Optimizing prognostic factors of five-year survival in gastric cancer patients using feature selection techniques with machine learning algorithms: a comparative study. *BMC medical informatics and decision making*, 23(1), 54.
3. Alirezapour, H., Mansouri, N., & Mohammad Hasani Zade, B. (2024). A comprehensive survey on feature selection with grasshopper optimization algorithm. *Neural Processing Letters*, 56(1), 28.
4. Azevedo, B. F., Rocha, A. M. A., & Pereira, A. I. (2024). Hybrid approaches to optimization and machine learning methods: a systematic literature review. *Machine Learning*, 113(7), 4055-4097.
5. Bagherzadeh, F., Mehrani, M.-J., Basirifard, M., & Roostaei, J. (2021). Comparative study on total nitrogen prediction in wastewater treatment plant and effect of various feature selection methods on machine learning algorithms performance. *Journal of Water Process Engineering*, 41, 102033.
6. Barbieri, L., Brambilla, M., & Roveri, M. (2024). A layer-wise personalization approach for transformer-based federated anomaly detection. 2024 2nd International Conference on Federated Learning Technologies and Applications (FLTA),
7. Cardia, M., Chessa, S., Micheli, A., Luminare, A. G., Franceschi, M., & Gambineri, F. (2024). Multitarget wastewater quality assessment in a smart industry context. 2024 International Conference on Intelligent Environments (IE),
8. Elmannai, H., El-Rashidy, N., Mashal, I., Alohal, M. A., Farag, S., El-Sappagh, S., & Saleh, H. (2023). Polycystic ovary syndrome detection machine learning model based on optimized feature selection and explainable artificial intelligence. *Diagnostics*, 13(8), 1506.
9. Eseye, A. T., Lehtonen, M., Tukia, T., Uimonen, S., & Millar, R. J. (2019). Machine learning based integrated feature selection approach for improved electricity demand forecasting in decentralized energy systems. *Ieee Access*, 7, 91463-91475.
10. Gholizadeh, M., Saeedi, R., Bagheri, A., & Paezi, M. (2024). Machine learning-based prediction of effluent total suspended solids in a wastewater treatment plant using different feature selection approaches: A comparative study. *Environmental Research*, 246, 118146.
11. González-Castro, L., Chávez, M., Duflot, P., Bleret, V., Del Fiol, G., & López-Nores, M. (2024). Impact of hyperparameter optimization to enhance machine learning performance: A case study on breast cancer recurrence prediction. *Applied Sciences*, 14(13), 5909.
12. Haque, B. T., & Rahman, M. A. (2024). Data Preparation and Feature Engineering in High-Dimensional Predictive Modeling. *Journal of Sustainable Development and Policy*, 3(04), 205-244.
13. Hartog, P. B., Krüger, F., Genheden, S., & Tetko, I. V. (2024). Using test-time augmentation to investigate explainable AI: inconsistencies between method, model and human intuition. *Journal of cheminformatics*, 16(1), 39.
14. Jha, A., Maheshwari, S., Dutta, P., & Dubey, U. (2025). Optimizing financial modeling with machine learning: integrating particle swarm optimization for enhanced predictive analytics. *Journal of Business Analytics*, 8(3), 196-215.
15. Kabir, T. (2026). Intelligent Condition Monitoring and Fault Diagnosis of Electrical Power and Control Systems Using Machine Learning-Based Predictive Analytics. *American Journal of Interdisciplinary Studies*, 7(01), 177-222.
16. Masoudi-Sobhanzadeh, Y., Motieghader, H., & Masoudi-Nejad, A. (2019). FeatureSelect: a software for feature selection based on machine learning approaches. *BMC bioinformatics*, 20(1), 170.
17. Muchlinski, P. (2023). Inward FDI Regulation in the UK: Closing the “Open Door”? In *Weaponising Investments: Volume I* (pp. 169-197). Springer.

18. Noroozi, Z., Orooji, A., & Erfannia, L. (2023). Analyzing the impact of feature selection methods on machine learning algorithms for heart disease prediction. *Scientific Reports*, 13(1), 22588.
19. Ogunpola, A., Saeed, F., Basurra, S., Albarrak, A. M., & Qasem, S. N. (2024). Machine learning-based predictive models for detection of cardiovascular diseases. *Diagnostics*, 14(2), 144.
20. Olawade, D. B., Soladoye, A. A., Omodunbi, B. A., Aderinto, N., & Adeyanju, I. A. (2025). Comparative analysis of machine learning models for coronary artery disease prediction with optimized feature selection. *International Journal of Cardiology*, 436, 133443.
21. Punia, S., & Shankar, S. (2022). Predictive analytics for demand forecasting: A deep learning-based decision support system. *Knowledge-Based Systems*, 258, 109956.
22. Santoso, L. (2024). Comparative Study of Feature Engineering Techniques for Predictive Data Analytics. *Journal of Technology Informatics and Engineering*, 3(2), 417-435.
23. Sharma, A., & Mishra, P. K. (2022). Performance analysis of machine learning based optimized feature selection approaches for breast cancer diagnosis. *International Journal of Information Technology*, 14(4), 1949-1960.
24. Sreehari, E., & Babu, L. D. (2024). Feature Selection Techniques for Data Analysis and Decision Making in Interdisciplinary Areas: A Systematic Review. *Ieee Access*, 12, 188845-188873.
25. Stüber, A. T., Coors, S., Schachtner, B., Weber, T., Rügamer, D., Bender, A., Mittermeier, A., Öcal, O., Seidensticker, M., & Rieke, J. (2023). A comprehensive machine learning benchmark study for radiomics-based survival analysis of CT imaging data in patients with hepatic metastases of CRC. *Investigative Radiology*, 58(12), 874.
26. Syed Mustapha, S. (2023). Predictive analysis of students' learning performance using data mining techniques: A comparative study of feature selection methods. *Applied System Innovation*, 6(5), 86.
27. Theng, D., & Bhoyar, K. K. (2024). Feature selection techniques for machine learning: a survey of more than two decades of research. *Knowledge and Information Systems*, 66(3), 1575-1637.
28. Tirumanadham, N. K. M. K., S, T., & M, S. (2024). Improving predictive performance in e-learning through hybrid 2-tier feature selection and hyper parameter-optimized 3-tier ensemble modeling. *International Journal of Information Technology*, 16(8), 5429-5456.
29. Wojciuk, M., Swiderska-Chadaj, Z., Siwek, K., & Gertych, A. (2024). Improving classification accuracy of fine-tuned CNN models: Impact of hyperparameter optimization. *Heliyon*, 10(5).
30. Yang, K., Liu, L., & Wen, Y. (2024). The impact of Bayesian optimization on feature selection. *Scientific Reports*, 14(1), 3948.
31. Zareehemat, P., Mohamadi, S., Valipour, J., & Moravvej, S. V. (2025). Forecasting stock market volatility using housing market indicators: a reinforcement learning-based feature selection approach. *Ieee Access*, 13, 52621-52643.



2026 by the authors; Journal of Global Social Transformation (JGST). This is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC-BY) license (<http://creativecommons.org/licenses/by/4.0/>).