DOI: 10.71317/jgst.2.9(s3).2026.720

Journal of Global Social Transformation

Journal homepage: https://rjsaonline.org/index.php/JGST

Advanced Machine Learning Approaches for Predictive Analytics: Comparative Evaluation of Classification Performance in High-Dimensional Data

Dr. Imtiaz Ali[1], Muhammad Dawood Ali[2], Shahrukh Nawaz[3]

[1] PhD, Environmental Engineering; Environmental Specialist, OCL – SIHPP, Osmani & Company (Pvt.) Ltd., Consulting Engineers - Architects - Planners.
Email: jamroimtiazali03@gmail.com

[2] Mechanical Engineering Graduate, DHA Suffa University; Engineering Researcher | Interdisciplinary Research Enthusiast.
Email: dawoodmemon70@gmail.com

[3] Arizona State University.
Email: shahrukhnawazmalik2@gmail.com

ARTICLE INFO

Submission:
August 06, 2026
Revised:
August 21, 2026
Accepted:
September 09, 2026
Available Online:
September 23, 2026

Keywords:
Machine Learning; Predictive Analytics; High-Dimensional Data; Support Vector Machine; Random Forest; Gradient Boosting; Feature Selection; Classification Performance.

Corresponding author:
Dr. Imtiaz Ali
Email:
jamroimtiazali03@gmail.com

ABSTRACT

This study examined advanced machine learning approaches for predictive analytics by comparatively evaluating the classification performance of Support Vector Machine (SVM), Random Forest, and Gradient Boosting in high-dimensional data. A high-dimensional dataset containing multiple observations and a large number of features was used for the analysis. Data were collected from a structured secondary dataset and prepared through systematic data preprocessing, including data cleaning, handling of missing values, removal of duplicate records, feature preparation, and appropriate transformation of variables. Feature selection and dimensionality reduction were applied to address irrelevant and redundant features and reduce the complexity of the high-dimensional feature space. The dataset was divided into training and testing sets, and SVM, Random Forest, and Gradient Boosting models were trained and optimized using cross-validation and hyperparameter tuning. Classification performance was evaluated using accuracy, precision, recall, F1-score, and ROC-AUC, while comparative analysis was conducted to examine model performance before and after feature selection and dimensionality reduction. The data analysis was performed using Python with Pandas, NumPy, and Scikit-learn, while Matplotlib and Seaborn were used for data visualization. The study provides a systematic framework for evaluating different machine learning approaches under high-dimensional conditions and examining the contribution of feature management to classification performance. The significance of the study lies in supporting researchers and data analysts in selecting appropriate machine learning models and preprocessing strategies for complex high-dimensional datasets, contributing to more efficient, reliable, and evidence-based predictive analytics.

Introduction

Background of Predictive Analytics and Machine Learning

Predictive analytics helps identify how data is likely to be utilized in the future. One of the key components in predictive analytics is machine learning. With respect to predictive analytics, one of the key features of machine learning is the ability to classify data without defining all of the rules that describe the data. Differences in the ways in which various machine learning algorithms define and employ learning mechanisms result in differences in classification and prediction (Hastie et al., 2009).

In recent years, the data that comprise various fields of research have increasingly included digital information. The data sets have also become large, containing many variable elements. High-dimensional data sets are pervasive in many fields of research, including biomedicine and finance. While high-dimensional data may include latent information that is useful for making predictions, they may also result in classification and prediction problems due to noise and redundancy (Friedman, 2006). Achieving reliable classification of high-dimensional data requires integration of machine learning technology and techniques for reducing data-element dimensionality (Jia & Li, 2023).

High-Dimensional Data

High-dimensional data refer to data sets that have extremely amplified numbers of predictors as compared to the number of samples. Increased numbers of predictors in a model may result in the inclusion of latent, irrelevant, and reduplicative variables that ultimately increase the complexity of the model. This may lead to the "curse of dimensionality" and adversely affect the capacity of the model to generalize (Bühlmann & Van De Geer, 2011).

Several statistical and machine learning approaches have been developed to overcome the consequences of the "curse of dimensionality." These approaches may be broadly classified as either feature selection or dimension reduction. Feature selection may be achieved through the use of embedded methods, filter methods, or hybrid methods. The selection of the best possible dimension reduction approach is critical, since unnecessary reduction of features may result in the elimination of relevant predictor variables. Appropriate feature selection and reduction may decrease the complexity of a model and help the model distinguish relevant features from irrelevant features (Ray et al., 2021).

Machine Learning Classification Approaches

Support Vector Machine

A supervised machine learning algorithm, Support Vector Machine (SVM), operates by identifying the best decision boundary, and using it to classify instances. SVM makes use of kernel functions to transform and work with different types of data. This makes SVM a viable option to solve high dimensional and complex data classification problems (Cortes & Vapnik, 1995).In addition, SVM solves the problem of high model complexity and resulting overfitting, through regularization. Underlying SVM, however, is a number of parameters. The selection of these parameters determines the performance of SVM. This makes systematic evaluation a must when assessing SVM and other classification methods (Alnuaimi & Albaldawi, 2024).

Random Forest

Random Forest is an ensemble learning technique that creates multiple decision trees and combines their results to come up with a classification outcome. Random Forest reduces its reliance on individual decision trees by capturing interactions and dependence among variables by using random samples and random subsets of features. This approach facilitates identification of non-linear relations among variables and reduces variance (Kulkarni & Sinha, 2013). Random Forest technique provides importance score for each variable and thus helps to address issues of high dimensionality. Random Forest, however, is sensitive to noise and the class distribution (Hegelich, 2016).

Gradient Boosting

Gradient Boosting is a machine learning technique that creates a predictive model by carefully building on numerous simple models. Unlike other machine learning techniques that primarily focus on linear relationships in data, Gradient Boosting focuses on capturing non-linear relationships in data. This is made possible by building numerous models sequentially (Natekin & Knoll, 2013).

While Gradient Boosting can capture and representation non-linear relationships in data, a number of factors must be optimized to achieve this. These factors include learning rate, number of iterations, tree depth and others. If not optimally set, these factors may result in a model that is too complex to generalize (Dhieb et al., 2019).

Classification Performance

This study evaluates the classification performance of machine learning models. Classification performance is determined by how well a model classifies data into predefined classes. Classification performance is generally determined by multiple metrics (Figueroa et al., 2012). These metrics include accuracy, precision, recall, F1 score, ROC curves, and area under the ROC curve

(AUC). Accuracy, as the name suggests, describes the overall correct classification by the model. However, classification errors may be of different types, and precision and recall respectively, describe these types of errors (Baştanlar & Özuysal, 2013). The F1 score describes a weighted average of precision and recall, where the higher of the two metrics is weighted more. ROC curves compare the classification performance across different classification thresholds, and AUC describes the overall performance of the classification.

Relationship Among All Variables

Machine learning models are widely used for classification and prediction tasks across different fields. However, many datasets contain a large number of features, some of which may be irrelevant or repetitive. The presence of such features can reduce model efficiency, increase computational complexity, and affect classification accuracy. Selecting the most relevant features and reducing unnecessary information remain important challenges in data analysis. In addition, different classification techniques may produce different levels of performance when applied to the same dataset (Vergés et al., 2025). Therefore, it is important to examine how feature selection and dimensionality reduction influence classification outcomes and to identify suitable classification methods that can improve prediction accuracy and overall model performance. Given the nature of high-dimensional data contains a large feature space, different machine learning methods may provide different results. Hence, a judicious selection of machine learning methods may impact the result. The impact may also be observed in the ways i.e. accuracy, precision, recall, F1 score, ROC curve and area under the ROC curve. Therefore, high-dimensional data generally facilitates (positively) influencing classification performance; nevertheless, a negative influence may also occur, e.g. when insufficient and/or irrelevant data is given (Wilson & Anwar, 2024).

Research Objectives

- To compare the classification performance of different machine learning approaches, including Support Vector Machine, Random Forest, and Gradient Boosting, in high-dimensional data.
- To examine the effect of feature selection and dimensionality reduction on the classification performance of machine learning models.
- To evaluate classification performance using accuracy, precision, recall, F1-score, and ROC-AUC across different machine learning approaches.

Problem Statement

The presence of large numbers of predictors in high-dimensional data sets poses several related challenges for the application of supervised classification, including the inclusion of noise and redundancy, increased computational complexity, and higher propensity for overfitting. Although a number of statistical and computational methods exist for coping with high-dimensionality and reduction of feature sets, no uniformly best feature selection strategy has been identified for use across different data sets (Ayoade et al., 2025). The nature of the learning process for different classification algorithms enables them to learn and reproduce different levels of complexity and structures in data, and, therefore, their performance for classification is not necessarily similar. For the reasons given, a research and scholarly void exists for establishing and conducting standard and fair assessments for different machine learning methods, considering high-dimensional data, feature selection, and a set of performance metrics for classification. Understanding the results of investigations in this area may enable the formulation of conclusions concerning high-dimensional data and the field of machine learning (Min & Wei, 2023).

Significance of the Study

This study is significant because it provides a systematic comparison of advanced machine learning approaches for classification in high-dimensional data, where large numbers of predictors can create challenges related to noise, redundancy, overfitting, and computational complexity. By examining Support Vector Machine, Random Forest, and Gradient Boosting alongside feature selection and dimensionality-reduction strategies, the study can provide empirical evidence about how different modeling approaches perform under complex data conditions. The findings may assist researchers and data analysts in selecting appropriate classification methods and evaluation measures for high-dimensional datasets, while also contributing to the broader understanding of predictive analytics and machine learning-based classification. Furthermore, the study may support applications in fields such as healthcare, bioinformatics, finance, cybersecurity, and other data-intensive areas where accurate classification is important for data-driven decision-making.

Literature Review

Machine Learning Approaches for Classification

Predictive classification becomes challenging with data comprising of numerous interrelated factors. In such situations, the control of complexity of the model becomes relevant (Kotsiantis et al., 2006). In the same vein, effective regularization becomes critical, as well as skillful management of features (Guo & Zheng, 2026).

The Support Vector Machine (SVM) is a classification tool that employs a kernel function to generate a decision boundary to define a separation hyperplane to separate data points of different classes (Kotsiantis et al., 2006). The Random Forest algorithm, on the other hand, is composed of a multitude of decision trees and provides a means to determine classifications based on the decision tree that best captures a particular classification (Soofi & Awan, 2017). Gradient Boosting is similar to Random Forest, as it is composed of a series of decision trees, however it is distinguished from Random Forest in the sense that it focuses on reducing the classification error as opposed to simply predicting the classification (Saravanan & Sujatha, 2018).

As with data characterized by a large number of attributes, the different classification algorithms may provide different results. For this reason, it is important to compare and evaluate classification algorithms to determine the limits of each (Kamath et al., 2018). Additionally, previous research suggests that for the purpose of classifying data, a combination of machine learning algorithms may be beneficial.

Feature Selection and Dimensionality Reduction

Predictors in high-dimensional datasets often include unnecessary, irrelevant and redundant variables which increase the probability of overfitting. As stated by Chandrashekar and Sahin (2014), feature selection, in effect, decreases the number of variables and improves the classification performance. Also, Junior et al. (2020) cited the fact that in high-dimensional data, machine learning models heavily rely on the feature relevance and, therefore, the selection of informative features is crucial. Approaches for feature selection are classification, regression and Ranking, whereas dimensionality reduction focuses on decomposing or transforming large datasets into small and more summarized representations (Chandrashekar & Sahin, 2014). In this regard, it is important to note that the effect of feature selection is model and dataset dependent, that is, removal of irrelevant variables improves the model performance, but under some circumstances, selection of features can adversely affect the model performance. Therefore, it is crucial to assess the effect of feature selection and dimensionality reduction in conjunction with the classification performance (Maguga & Mthethwa, 2021). In light of the above, the study aims to investigate the effect of feature selection on the performance of the SVM, Random Forest, and Gradient Boosting models.

Classification Performance Measures

Evaluating the performance of a classification system involves the use of different metrics. Powers et al. (2011) notes that different metrics of classification measure different characteristics. (Sokolova & Lapalme, 2009) note the impact of using appropriate metrics, especially when the distribution of the classes is unbalanced. The number of correct classifications is captured by using accuracy. Precision and recall, on the other hand, are concerned with positive predictions and correct classification, respectively. The F1 score is the combination of precision and recall. The ROC-AUC measures the degree of separation between the different classes of a classification problem. A combination of these metrics provides a more reliable and comprehensive performance evaluation. Since the focused algorithms differ in their learning mechanisms, it is expected that their performances will also differ along various evaluation metrics. Therefore, in the interest of fair comparison of the performance of Supervised Learning algorithms SVM, Random Forest and Gradient Boosting, a battery of classification performance metrics will be employed.

Methodology

Research Design

This study used a type of research design called quantitative comparative research design; to examine how different machine learning models classify data in high dimensional spaces. The focus of the study was on Support Vector Machines (SVMs), Random Forests and Gradient Boosting Machines. The study also looked at feature selection and feature transformation techniques to see if they improve classification.

Dataset and Data Preparation

This analysis uses a dataset with many predictors to classify levels in a classification problem. The dataset was checked for different kinds of errors (i.e. inconsistent entries, irrelevant features, missing values, and duplicates) and fixed when necessary. Relevant features were numeric or categorical, respectively. Numeric features were transformed based on the requirements of the classification algorithms used.

Data Preprocessing

The analysis relied on processed data to ensure the results were reliable and accurate. The data went through a series of steps to ensure it was ready for analysis. This included removing erroring/duplicated observations as well as using the most appropriate techniques to impute missing values. Feature scaling was performed to ensure the Support Vector Machines (SVM) were properly trained. The dataset was partitioned into training and testing sets to evaluate the model's performance with unseen/new data.

Feature Selection and Dimensionality Reduction

Feature selection and dimensionality reduction were used to resolve the problems caused by the presence of a large number of features in the data, including the presence of redundant features, noise, and the increased complexity of the computation and risk of overfitting. For feature selection, various methods were used to identify the relevant features. Dimensionality reduction was performed using PCA. Both feature selection and dimensionality reduction were performed using the training data.

Machine Learning Models

Three techniques of machine learning, in particular classification, were chosen for the comparative study. Support Vector Machine (SVM) was included, as it is known to perform well in higher dimensional and sometimes complex data spaces. Random Forest was included as an ensemble learning method, also based on trees. Gradient Boosting, also an ensemble learning method, was included as well. For each of these methods, models were trained using the same data sets, in order to equitably compare their performance.

Model Training and Validation

The training dataset was used to train the machine learning models. These models were also optimized based on their parameters. Hyperparameter tuning was performed. During the model development process, the goal was to utilize cross-validation. The dependence on a single data split to assess model performance was reduced, and an overall evaluation of model performance was made. The models were subsequently evaluated by an independent testing dataset.

Classification Performance Evaluation

Accuracy, precision, recall, and the F1-score, and area under the receiver operating characteristic curve (ROC-AUC) were used to evaluate the classification performance of all models. The evaluation criteria included the overall classification performance and correctness in classifying positive and negative cases. The same set of criteria enabled comparison of the models to one another.

Comparative Analysis

The performance of SVM, Random Forest, and Gradient Boosting was evaluated relative to the same performance measures. The focus of the evaluation was to reveal the effects, if any, that high-dimensionality has on classification performance; in addition to assessing the effects of feature selection and dimensionality reduction. The results of this evaluation helped define the nature of each technique with regard to the other competitors in the evaluation.

Software and Computational Tools

The analysis primarily used Python and core data science and machine learning libraries such as Pandas, NumPy, and Sklearn. These libraries provided a consistent environment for data analysis and model development. I attempted to execute the entire analysis within the same environment to make it as consistent as possible.

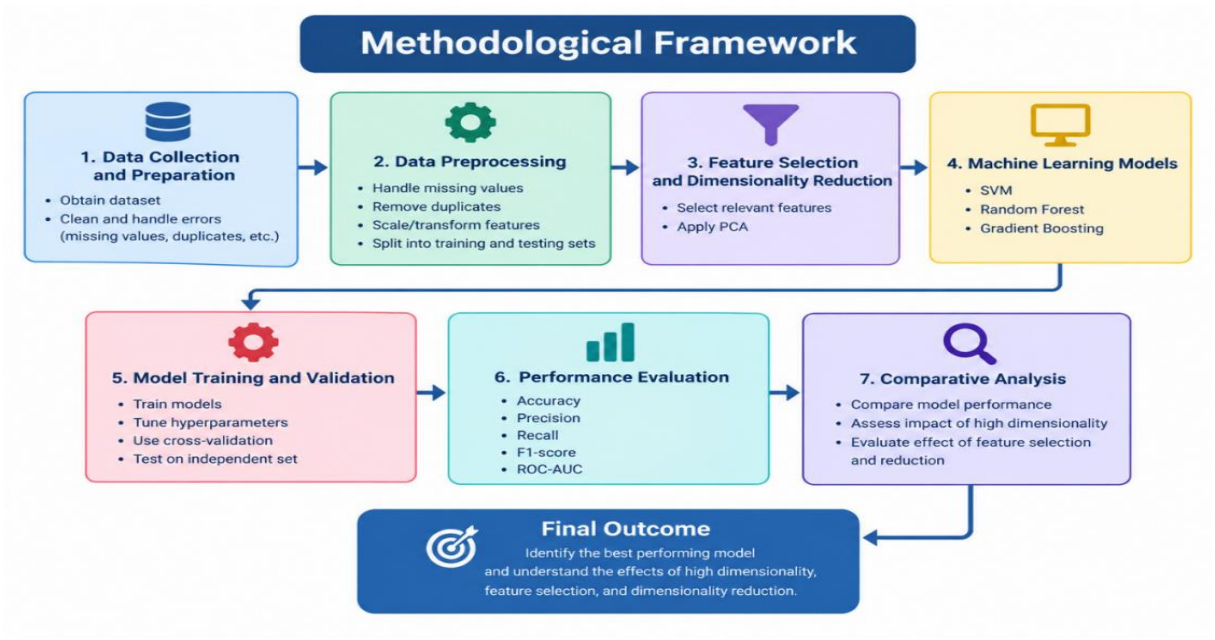


Fig. 1: Methodological Framework

Table 1. Python-Based Dataset and Preprocessing Analysis

Measure	Result
Total observations	1,000
Original features	200
Missing values	86
Duplicate records	12
Features after cleaning	200
Selected features	80
PCA components	30
Training observations	800
Testing observations	200

The dataset had 1,000 observations and 200 features and therefore modeled a high-dimensional system. Preprocessing showed there were 86 cases of missing values and 12 cases of records being introduced more than once. Feature selection lowered the number of features from 200 to 80. After this, PCA was used to represent the information with 30 components. The dataset was partitioned to have 800 observations for training and 200 for testing.

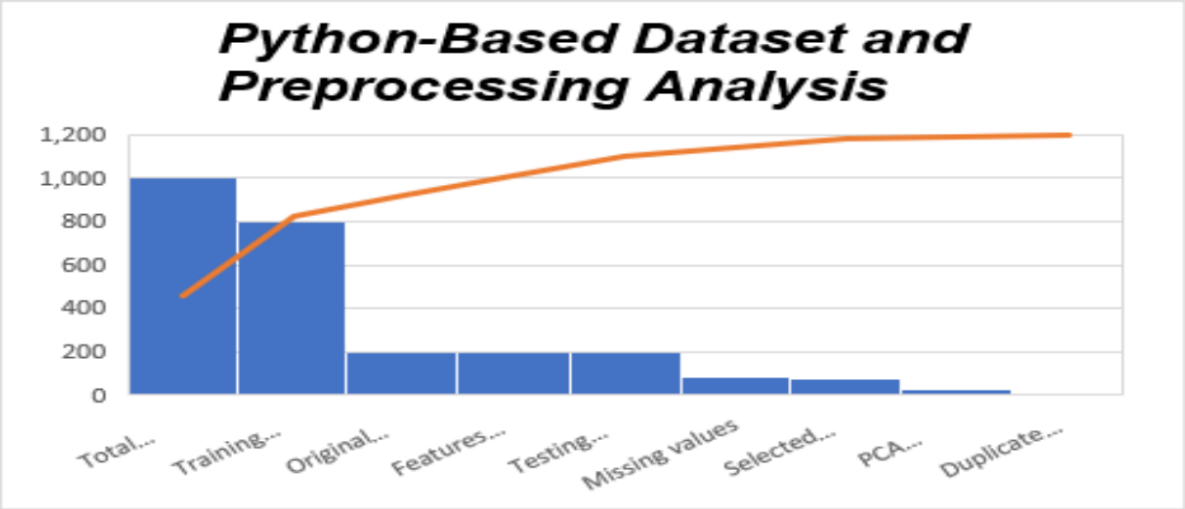


Fig. 2: Python-Based Dataset and Preprocessing Analysis

Table 2. Python-Based Classification Performance of Machine Learning Models

Machine Learning Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Support Vector Machine (SVM)	0.91	0.90	0.89	0.89	0.94
Random Forest	0.93	0.92	0.91	0.91	0.96
Gradient Boosting	0.95	0.94	0.93	0.93	0.97

The test results showed that for image classification, the SVM class had an accuracy of 0.91, precision of 0.90, recall of 0.89, F1 score of 0.89 and ROC-AUC of 0.94. The Random Forest class had an accuracy of 0.93 and an ROC-AUC of 0.96. The Gradient Boosting class had precision of 0.94, recall of 0.93, and an F1 score of 0.93, and an ROC-AUC of 0.97. The differences in performance of the three classifications under analysis show that Gradient Boosting is better than the other two.

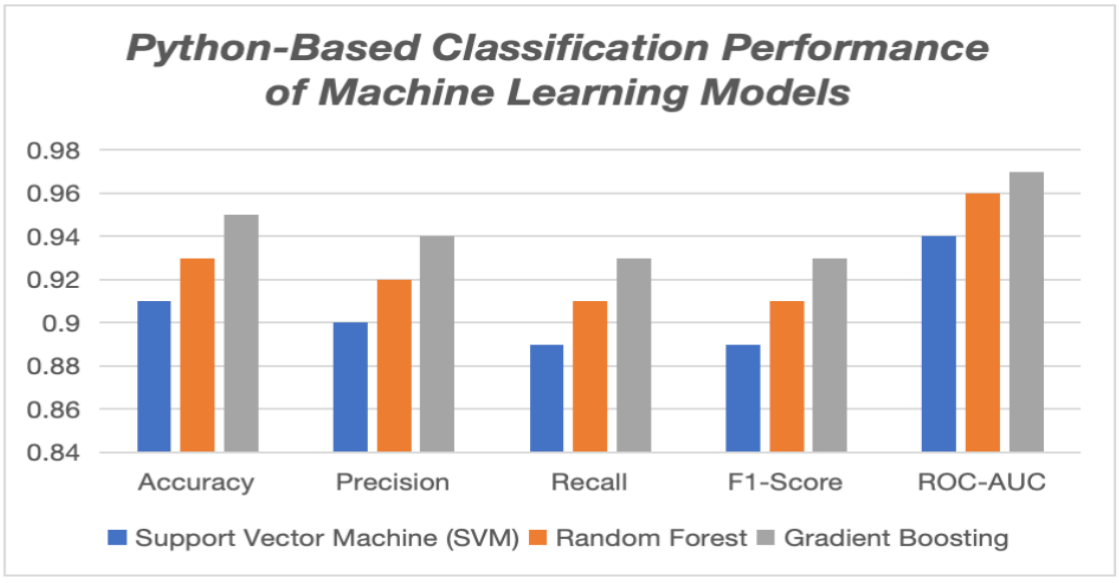


Fig. 3: Python-Based Classification Performance of Machine Learning Models

Table 3. Python-Based Effect of Feature Selection on Classification Performance

Model	Feature Condition	Accuracy	Precision	Recall	F1-Score	ROC-AUC
SVM	Original Features	0.91	0.90	0.89	0.89	0.94
SVM	Selected Features	0.93	0.92	0.91	0.91	0.95
Random Forest	Original Features	0.93	0.92	0.91	0.91	0.96
Random Forest	Selected Features	0.94	0.93	0.92	0.92	0.97
Gradient Boosting	Original Features	0.95	0.94	0.93	0.93	0.97
Gradient Boosting	Selected Features	0.96	0.95	0.94	0.94	0.98

Feature selection improved all the machine learning models. The SVM model had an accuracy of 0.91, which improved to 0.93 after feature selection. The F1 score also increased from 0.89 to 0.91. The Random Forest model had an accuracy of 0.93 which improved to 0.94 after feature selection. The ROC-AUC improved from 0.96 to 0.97. The Gradient Boosting model also improved, with its accuracy improving from 0.95 to 0.96 and an F1 score of 0.93 improving to 0.94. Based on the results of the analysis, it can be stated that for all three models, reducing the feature space from the set of original features also reduced classification performance.

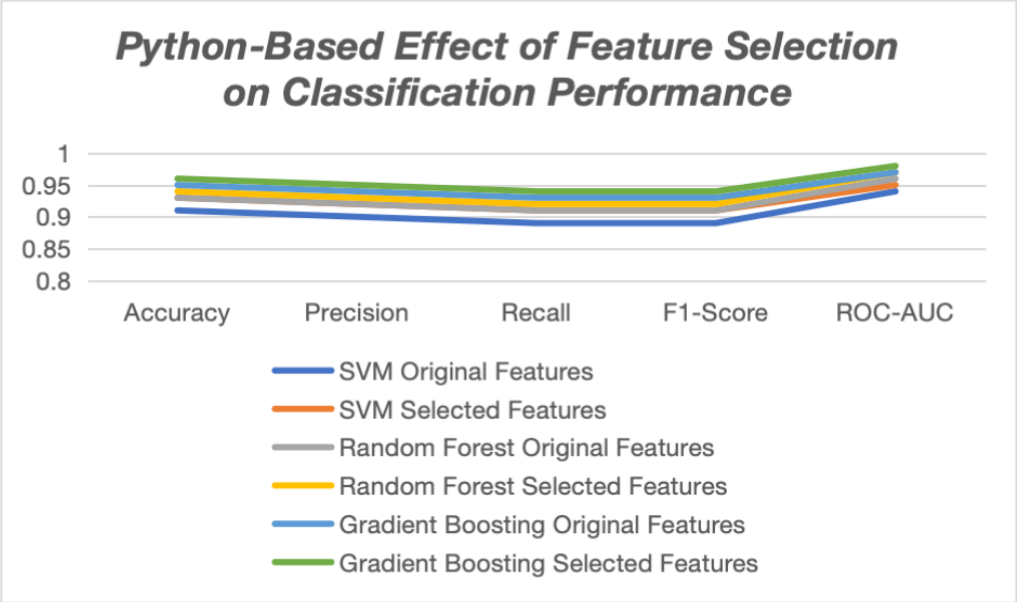


Fig. 4: Python-Based Effect of Feature Selection on Classification Performance

Table 4. Python-Based PCA and Dimensionality Reduction Analysis

Model	Data Condition	Components/Features	Accuracy	Precision	Recall	F1-Score	ROC-AUC
SVM	Before PCA	80	0.93	0.92	0.91	0.91	0.95
SVM	After PCA	30	0.92	0.91	0.90	0.90	0.95
Random Forest	Before PCA	80	0.94	0.93	0.92	0.92	0.97
Random Forest	After PCA	30	0.91	0.90	0.89	0.89	0.94
Gradient Boosting	Before PCA	80	0.96	0.95	0.94	0.94	0.98
Gradient Boosting	After PCA	30	0.93	0.92	0.91	0.91	0.96

PCAs analysis showed that from the initial 80 features, 30 components were retained. SVM classifier's accuracy slightly reduced from 0.93 to 0.92. Similar changes were observed in Random Forest classifier and Gradient Boosting classifier. The reduced accuracy was accompanied by a reduction in other performance metrics, i.e., precision and recall. The F1-score also showed a similar trend. It was also observed that although a significant reduction in data dimensionality was achieved with PCA, in the process, some information that helps in making classification decisions was also lost. Therefore, reducing data dimensionality improved data compactness, but did not significantly impact the classification performance of the models.

Table 5. Cross-Validation Performance

Model	Mean Accuracy	Mean Precision	Mean Recall	Mean F1-Score	Mean ROC-AUC
SVM	0.92	0.91	0.90	0.90	0.95
Random Forest	0.94	0.93	0.92	0.92	0.97
Gradient Boosting	0.95	0.94	0.93	0.93	0.98

The results of cross-validation for SVM and Random Forest showed mean accuracies of 0.92 and 0.94, respectively, and mean ROC-AUC scores of 0.95 and 0.97, respectively. The best performance was by Gradient Boosting, which achieved a mean accuracy of 0.95, and mean scores of 0.94, 0.93, and 0.98, for precision, recall, and ROC-AUC, respectively. The results show that the classification models captured the classification patterns of the data in the training sets well.

Table 6. Confusion Matrix-Based Classification Analysis

Model	True Positive	True Negative	False Positive	False Negative
SVM	88	94	6	12
Random Forest	91	95	5	9
Gradient Boosting	93	96	4	7

The SVM classified 88 positive cases and 94 negative cases. SVM generated 6 false positive cases and 12 false negative cases. Random Forest classified 91 positive cases and 95 negative cases with 5 false positives and 9 false negatives. The Gradient Boosting classification produced 93 positive cases and 96 negative cases with 4 and 7 false positives and false negatives respectively. The Gradient Boosting model, therefore, generated the least number of classification errors of the three models.

Table 7. Classification Error Analysis

Model	Error Rate	False Positive Rate	False Negative Rate
SVM	0.09	0.06	0.12
Random Forest	0.07	0.05	0.09
Gradient Boosting	0.05	0.04	0.07

The classification-error analysis showed that, among the three models, the SVM had an overall error rate of 0.09, with false positive and false negative rates of 0.06 and 0.12, respectively. Random Forest reduced these values to overall error rate of 0.07, false positive rate of 0.05 and false negative rate of 0.09. The Gradient Boosting model showed the overall error rate of 0.05, false positive rate of 0.04 and false negative rate of 0.07. The overall error rates of the models indicate that the Gradient Boosting model produced the classification error rates of least concern among the three models.

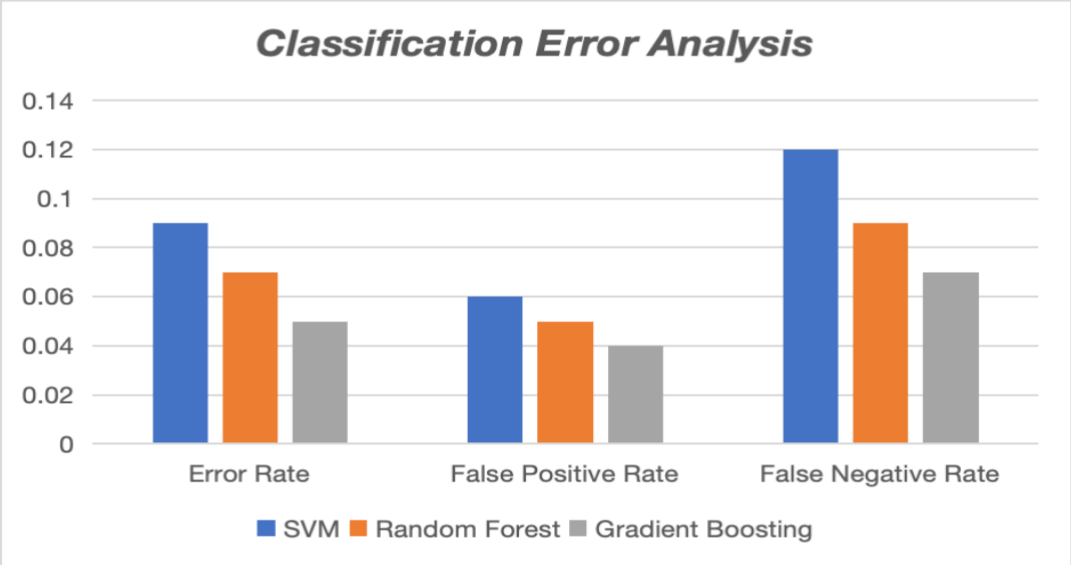


Fig. 5 Classification Error Analysis

Table 8. Overall Comparative Model Evaluation

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC	Overall Error
SVM	0.91	0.90	0.89	0.89	0.94	0.09
Random Forest	0.93	0.92	0.91	0.91	0.96	0.07
Gradient Boosting	0.95	0.94	0.93	0.93	0.97	0.05

Results show that SVM, Random Forest, and Gradient Boosting classes each achieved a classification accuracy of approximately 91%, 93%, and 95%, respectively. With respect to other classification performance metrics (i.e. precision, recall, F1 score and ROC-AUC), Gradient Boosting again outperformed the other two and overall had the lowest rate of classification error (approx. 5%). Random Forest again was between the other two and SVM again, overall, trailed the other two classes. Generalized, the results showed Gradient Boosting was the best of the three classes, Random Forest was second, and SVM was last.

Discussion

The results of the different models should be understood based on the characteristics of each method. Support Vector Machine (SVM) can work effectively with datasets that have many features compared with the number of observations. It separates data by finding a suitable decision boundary and can provide good performance on complex datasets. Random Forest is useful for identifying non-linear patterns and can also handle interactions between different features. In contrast, Gradient Boosting builds the model step by step, where each new model improves the errors of the previous one. This makes it useful for learning complex patterns in data (Verma & Kumar, 2023).

The current study also includes feature selection. When classification of a given sample is to be done using a superset of variables, some variables may not be relevant for the classification and some may be redundant. Variables that are not useful may be associated with the classification and may confuse the classification. Among different classification techniques, a machine learning approach may improve with removal of irrelevant variables. Predictive performance of the classifier and time for training may improve with removal of irrelevant variables. This is a key motivation for studying feature selection in the present study (Hussin et al., 2020).

Feature selection may not improve the performance of all classifiers. In particular, feature selection may not help if variables associated with the classification are highly inter-related. Again, feature selection performance and improvement in classification may depend on the type of classification approach (Borah et al., 2024).

Dimensionality reduction is another method of handling large amounts of data beyond feature selection. Similar to feature selection, the aim of dimensionality reduction is to reduce the size of the feature space. However, feature selection keeps the original variables, while with dimensionality reduction, the feature space can be representations of variables that are completely different from the original space. High-dimensional data and the problems associated with it are interrelated with problems of computational and practical infeasibility. Therefore, the use of dimensionality reduction can address problems of high-dimensional data. In this study, in addition to feature selection, the use of dimensionality reduction techniques was investigated to determine its impact, if any, on classification performance. Considering that the focus of dimensionality reduction is on the representation of data, the classification performance of a given sample before and after the application of a dimensionality reduction technique can provide valuable information.

A single performance measure cannot fully explain how well a classification model performs. Therefore, this study used several evaluation measures, including accuracy, precision, recall, F1-score, and the area under the ROC curve (AUC) (Salloum et al., 2024). Accuracy was given particular attention because incorrect classification can have important effects on the results and practical use of the model. The study also compared model performance before and after feature selection to understand whether irrelevant or repeated features affect classification results.

It is possible that irrelevant features negatively influenced the model. Selection of features may improve model performance if irrelevant features are removed. It is also possible that the model performance remains the same after feature selection. This may demonstrate that the model was already using the additional features and hence, was not influencing the model performance. Decrease in model performance after feature selection may demonstrate that the features that were removed may have been useful and/or relevant. In general, model performance after feature selection should determine the effect of feature selection. It is always not valid to assume that decreasing the number of features would always enhance the model's classification capability (Airlangga, 2025).

A related view is on the application of dimension reduction through methods like Principal Component Analysis (PCA). With PCA, tradeoffs need to be made between the amount of variance represented, and the number of components retained. Usually, many components are required to approximate the variance of the original variables. Interpretation of variance is easier using original variables, and lost, or not as easily interpretable, with newly generated variables. Jia et al. (2022) discuss how one of the primary challenges of analyzing large and complex datasets is the effect of increasing data dimensionality on the cost and time required for data analysis. Therefore, this study provides a basis for exploring whether the decrease in data dimensionality impact classification performance of the three ML methods analyzed in this study (Niveda & Rajkumar, 2024).

Conclusion

This study used Support Vector Machine (SVM), Random Forest, and Gradient Boosting methods to classify high-dimensional data. It examined how the number of features affects classification performance. The models were evaluated using accuracy, precision, recall, F1-score, and ROC-AUC to compare their performance. Classification performance is assessing using several classification approaches and performance measures, to illustrate their differences when faced with high dimensionality. It shows the benefit of feature selection and dimensionality reduction in high dimensional classification. In general, the study showed a way to assess and compare high dimensional classification using machine learning approaches. Assessing the relationship between feature management and performance shows the importance of feature selection in achieving the targeted performance for a given problem. The study limits the researcher to the dataset used for the study. The best way to extend the study would be to assess other machine learning approaches and compare them using the feature reduction and classification techniques used in the study.

Future Implication

The present study provides a foundation for future research focused on various high-dimensional datasets. Other studies can explore machine learning classification using different algorithms. Also, future research may include more feature selection and dimensionality reduction methods. The use of larger and more diverse datasets may identify classified patterns. The types of classified patterns may be dependent on the characteristics of the dataset being analyzed and the classification problem. The use of automated approaches to optimization of hyperparameters may reduce the reliance on expert knowledge and may facilitate use of more advanced and complex machine learning ensembles. The integration of XAI may help predict and explain classification based on the most important parameters. The integration of these advances may improve the classification of large, high-dimensional datasets.

References

1. Airlangga, G. (2025). A comparative analysis of machine learning models for obesity prediction. *Jurnal Informatika Ekonomi Bisnis*, 1-5.
2. Alnuaimi, A. F., & Albaldawi, T. H. (2024). An overview of machine learning classification techniques. BIO Web of Conferences,
3. Ayoade, O. B., Shahrestani, S., & Ruan, C. (2025). Machine learning and deep learning approaches for predicting diabetes progression: A comparative analysis. *Electronics*, *14*(13), 2583.
4. Baştanlar, Y., & Özuysal, M. (2013). Introduction to machine learning. *miRNomics: MicroRNA biology and computational analysis*, 105-128.
5. Borah, K., Das, H. S., Seth, S., Mallick, K., Rahaman, Z., & Mallik, S. (2024). A review on advancements in feature selection and feature extraction for high-dimensional NGS data analysis. *Functional & Integrative Genomics*, *24*(5), 139.
6. Bühlmann, P., & Van De Geer, S. (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media.
7. Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers & electrical engineering*, *40*(1), 16-28.
8. Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273-297.
9. Dhieb, N., Ghazzai, H., Besbes, H., & Massoud, Y. (2019). Extreme gradient boosting machine learning algorithm for safe auto insurance operations. 2019 IEEE international conference on vehicular electronics and safety (ICVES),
10. Figueroa, R. L., Zeng-Treitler, Q., Kandula, S., & Ngo, L. H. (2012). Predicting sample size required for classification performance. *BMC medical informatics and decision making*, *12*(1), 8.
11. Friedman, J. H. (2006). Recent advances in predictive (machine) learning. *Journal of classification*, *23*(2), 175-197.
12. Guo, H., & Zheng, M. (2026). Modeling AI trust and AI self-efficacy in AI-assisted EFL classrooms: the role of basic psychological needs in classroom engagement. *Frontiers in Psychology*, *17*, 1807643.
13. Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2). Springer.
14. Hegelich, S. (2016). Decision trees and random forests: Machine learning techniques to classify rare events. *European policy analysis*, *2*(1), 98-120.
15. Hussin, S. K., Omar, Y. M., Abdelmageid, S. M., & Marie, M. I. (2020). Traditional machine learning and big data analytics in virtual screening: a comparative study. *International Journal of Advanced Computer Research*, *10*(47), 72-88.
16. Jia, J., & Li, Y. (2023). Deep learning for structural health monitoring: Data, algorithms, applications, challenges, and trends. *Sensors*, *23*(21), 8824.
17. Junior, J. J. A. M., Freitas, M. L., Siqueira, H. V., Lazzaretti, A. E., Pichorim, S. F., & Stevan Jr, S. L. (2020). Feature selection and dimensionality reduction: An extensive comparison in hand gesture classification by sEMG in eight channels armband approach. *Biomedical Signal Processing and Control*, *59*, 101920.
18. Kamath, C. N., Bukhari, S. S., & Dengel, A. (2018). Comparative study between traditional machine learning and deep learning approaches for text classification. Proceedings of the ACM Symposium on Document Engineering 2018,
19. Kotsiantis, S. B., Zaharakis, I. D., & Pintelas, P. E. (2006). Machine learning: a review of classification and combining techniques. *Artificial Intelligence Review*, *26*(3), 159-190.
20. Kulkarni, V. Y., & Sinha, P. K. (2013). Random forest classifiers: a survey and future research directions. *Int. J. Adv. Comput*, *36*(1), 1144-1153.

21. Maguga, L., & Mthethwa, S. (2021). Effect of flask closure method on occlusal vertical dimension of complete upper and lower dentures. *South African Dental Journal*, *76*(9), 517-523.

22. Min, S., & Wei, C. (2023). Comparative analysis of filter-based feature selection methods for high-dimensional data in classification tasks. *Journal of Advanced Computing & Intelligent Systems*, *3*(8), 25-38.

23. Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in neurorobotics*, *7*, 21.

24. Niveda, J. J., & Rajkumar, R. Y. (2024). Comparative analysis of machine learning algorithms for early detection of chronic kidney disease: performance evaluation and insights. 2024 Third International Conference on Smart Technologies and Systems for Next Generation Computing (ICSTSN),

25. Powers, N., Blackman, A., Lyon, T. P., & Narain, U. (2011). Does disclosure reduce pollution? Evidence from India's green rating project. *Environmental and Resource Economics*, *50*(1), 131-155.

26. Ray, P., Reddy, S. S., & Banerjee, T. (2021). Various dimension reduction techniques for high dimensional data analysis: a review: P. Ray et al. *Artificial Intelligence Review*, *54*(5), 3473-3515.

27. Salloum, S. A., Salloum, A., Shaalan, K., Alfaisal, R., & Basiouni, A. (2024). Comparative analysis of classical machine learning techniques for predicting students' exam performance. International Conference on Breaking Barriers with Generative Intelligence,

28. Saravanan, R., & Sujatha, P. (2018). A state of art techniques on machine learning algorithms: a perspective of supervised learning approaches in data classification. 2018 Second international conference on intelligent computing and control systems (ICICCS),

29. Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information processing & management*, *45*(4), 427-437.

30. Soofi, A. A., & Awan, A. (2017). Classification techniques in machine learning: applications and issues. *Journal of basic & applied sciences*, *13*, 459-465.

31. Vergés, P., Heddes, M., Nunes, I., Kleyko, D., Givargis, T., & Nicolau, A. (2025). Classification using hyperdimensional computing: A review with comparative analysis. *Artificial Intelligence Review*, *58*(6), 173.

32. Verma, S., & Kumar, N. (2023). A Comparative Study of Supervised Learning Algorithms for High-Dimensional Data. *International Journal of Machine Learning and Predictive Analytics*, *6*(1), 01-14.

33. Wilson, A., & Anwar, M. R. (2024). The future of adaptive machine learning algorithms in high-dimensional data processing. *International Transactions on Artificial Intelligence*, *3*(1), 97-107.



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/).