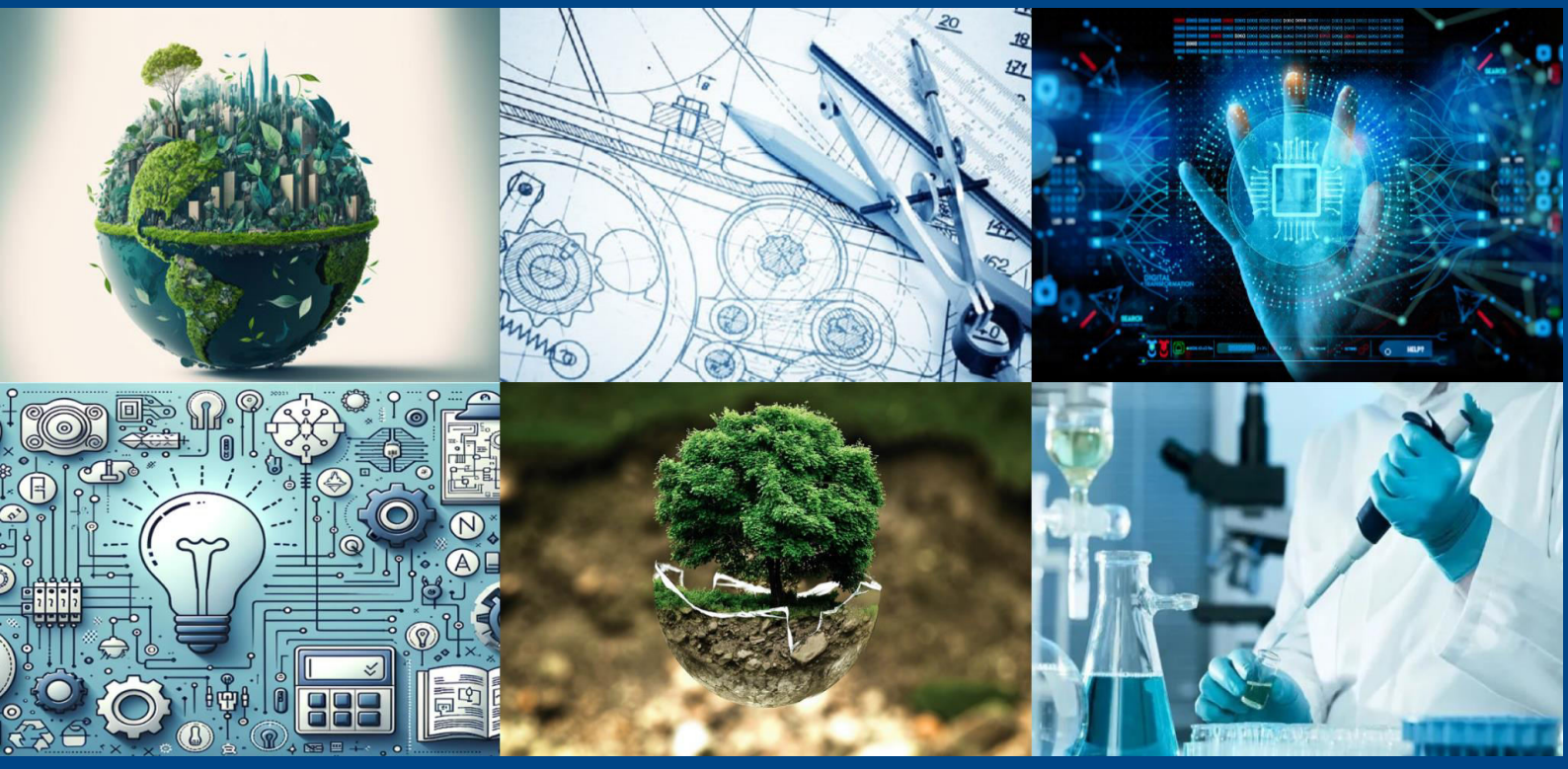
ISSN: 2582-7219

International Journal of Multidisciplinary Research in Science, Engineering and Technology

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 9.864

Volume 9, Issue 7, July 2026

© 2026 IJMRSET | Volume 9, Issue 7, July 2026| DOI:10.15680/IJMRSET.2026.0907019

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 9.864| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Comparative Performance Analysis of Machine Learning Models for Predictive Analytics

Arulkumar S, Rikshitha M, Paranithar S, and Mohammed Salaman A

Department of Computer Science with Data Analytics, Dr. N.G.P. Arts and Science College, Coimbatore, Tamil Nadu, India

ABSTRACT: The rapid growth of data-driven applications has increased the importance of predictive analytics techniques for extracting valuable insights and supporting intelligent decision-making. Machine learning algorithms provide effective solutions for classification and prediction problems by automatically identifying patterns from historical data. However, selecting an appropriate machine learning algorithm remains a challenging task because different models exhibit different learning capabilities and prediction performances.

This research presents a comparative performance analysis of multiple machine learning models for predictive analytics. Five supervised learning algorithms, namely Logistic Regression, Support Vector Machine, K-Nearest Neighbor, Random Forest, and Gradient Boosting, are implemented and evaluated using the Breast Cancer Wisconsin Diagnostic dataset.

The experimental workflow includes exploratory data analysis, data preprocessing, feature scaling, model train-ing, and performance evaluation. The dataset is divided into training and testing subsets using an 80:20 ratio, and StandardScaler is applied to normalize feature values. The performance of the developed models is analyzed using accuracy, precision, recall, and F1-score metrics.

The experimental results demonstrate that Support Vector Machine achieves the highest predictive performance among all evaluated algorithms with an accuracy of 98.24% and an F1-score of 98.61%. The findings of this study highlight the importance of model selection and preprocessing techniques in developing reliable predictive analytics systems.

KEYWORDS: Machine Learning, Predictive Analytics, Classification, Support Vector Machine, Ensemble Learn-ing, Performance Evaluation

I. INTRODUCTION

In recent years, the rapid advancement of digital technologies has resulted in the generation of massive amounts of data across various domains. Extracting meaningful information from these large datasets has become essential for improving decision-making processes. Predictive analytics has emerged as an important approach that uses statistical techniques and machine learning algorithms to analyze historical information and predict future outcomes.

Machine learning is a branch of artificial intelligence that enables computer systems to learn patterns from data without being explicitly programmed. By identifying relationships between input variables and target outcomes, machine learning models can perform complex classification and prediction tasks efficiently.

The effectiveness of predictive analytics systems depends greatly on the selection of appropriate machine learning algorithms. Different algorithms use different mathematical approaches for learning patterns from data. Some algo-rithms provide better interpretability, while others achieve higher accuracy by capturing complex relationships among features.

Classification algorithms have been widely applied in various applications such as healthcare analysis, financial pre-diction, cybersecurity, and decision support systems. Among the available machine learning techniques, Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Random Forest, and Gradient Boosting are commonly used because of their reliability and effectiveness.

© 2026 IJMRSET | Volume 9, Issue 7, July 2026| DOI:10.15680/IJMRSET.2026.0907019

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 9.864| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Logistic Regression provides a simple and interpretable approach for binary classification problems. Support Vector Machine performs well in high-dimensional feature spaces by identifying optimal decision boundaries. K-Nearest Neighbor uses similarity-based learning, while Random Forest and Gradient Boosting improve prediction performance through ensemble learning techniques.

Although several machine learning algorithms are available, identifying the most suitable model for a specific dataset remains a significant challenge. A model that performs well on one dataset may not provide similar results on another dataset due to differences in feature distribution, dataset size, and complexity.

Therefore, comparative analysis of multiple machine learning models is important for understanding their strengths, limitations, and suitability for predictive analytics applications. Such evaluation provides valuable insights into algorithm selection and helps develop efficient data-driven solutions.

In this research, a comparative study is conducted using the Breast Cancer Wisconsin Diagnostic dataset. Five machine learning classification algorithms are implemented under the same experimental environment and evaluated using standard performance metrics.

The main contributions of this research are:

- A complete machine learning pipeline is developed including data exploration, preprocessing, model training, and evaluation.
- Five popular classification algorithms are compared using the same dataset and experimental conditions.
- Multiple evaluation metrics including accuracy, precision, recall, and F1-score are used to provide comprehensive performance analysis.
- The best-performing machine learning model is identified based on experimental results.

The remainder of this paper is organized as follows. Section II discusses related work and previous studies. Section III explains the proposed methodology, dataset characteristics, preprocessing techniques, and machine learning models. Section IV presents experimental results and comparative analysis. Section V provides discussion, conclusion, and future research directions.

II. RELATED WORK

Machine learning-based predictive analytics has gained significant attention due to its ability to analyze complex datasets and generate accurate predictions. Researchers have investigated various classification algorithms to improve prediction performance and develop reliable intelligent systems.

Cortes and Vapnik introduced Support Vector Machine, which became one of the most widely used classification algorithms. SVM constructs an optimal separating hyperplane by maximizing the margin between different classes. Due to its effectiveness in handling high-dimensional data, SVM has been applied successfully in various classification applications.

Breiman proposed the Random Forest algorithm, an ensemble learning approach that combines multiple decision trees to improve prediction accuracy. Random Forest reduces the possibility of overfitting by generating predictions based on multiple decision trees and has demonstrated strong performance in classification tasks.

Logistic Regression has been extensively used for binary classification problems because of its simplicity, computational efficiency, and interpretability [10]. Although it performs effectively on linearly separable datasets, its performance may decrease when handling complex nonlinear relationships.

K-Nearest Neighbor is another commonly used classification technique that predicts class labels based on similarity between data samples [6]. The algorithm calculates the distance between samples and assigns the class label based on similarity. Although KNN is simple and easy to understand, its performance depends on feature scaling and the selection of an appropriate value of K.

© 2026 IJMRSET | Volume 9, Issue 7, July 2026| DOI:10.15680/IJMRSET.2026.0907019

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 9.864| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Gradient Boosting algorithms have also received considerable attention due to their ability to improve prediction accuracy by combining multiple weak learners [7, 15]. These methods construct models sequentially, where each new model attempts to correct the errors produced by previous models. Scalable implementations such as XGBoost have further extended the practical applicability of boosting-based ensembles to large and high-dimensional datasets [8].

Several previous studies have focused on comparing machine learning algorithms for predictive analysis. However, the performance of individual algorithms varies depending on dataset properties, preprocessing strategies, and evaluation methods.

A comparative study provides a better understanding of algorithm behavior under similar experimental conditions. By evaluating multiple models using identical preprocessing and evaluation procedures, the strengths and limitations of different approaches can be identified [12, 13]. Reliable estimation of model performance also depends on the evaluation strategy adopted, such as train-test splitting or cross-validation [11].

This research performs a systematic comparison of five machine learning algorithms using the Breast Cancer Wisconsin Diagnostic dataset. The study focuses on evaluating model performance using multiple classification metrics and identifying the algorithm that provides the best predictive capability.

III. METHODOLOGY

The proposed methodology follows a structured machine learning workflow for developing and evaluating predictive classification models. The complete process consists of dataset collection, exploratory data analysis, preprocessing, model development, performance evaluation, and comparative analysis.

The workflow adopted in this research is designed to ensure consistent evaluation of all machine learning algorithms. Each model is trained and tested using the same dataset partition and preprocessing procedure.

A. Research Workflow

The major steps involved in the proposed workflow are described below:

1. **Dataset Collection:** The Breast Cancer Wisconsin Diagnostic dataset available in the Scikit-learn library is selected for experimental evaluation.
2. **Exploratory Data Analysis:** The dataset characteristics are analyzed by examining data structure, feature information, class distribution, and missing values.
3. **Data Preprocessing:** The dataset is prepared by separating input features and target labels. Feature scaling is applied to improve model learning performance.
4. **Model Training:** Five machine learning classification algorithms are trained using the processed training dataset.
5. **Performance Evaluation:** The trained models are evaluated using accuracy, precision, recall, and F1-score metrics.
6. **Comparative Analysis:** The performance of all models is compared to identify the most effective algorithm.

IV. EXPERIMENTAL SETUP

The experiments were implemented using Python programming language and Scikit-learn machine learning framework [3, 14]. The implementation environment includes commonly used data science libraries such as NumPy, Pandas, Matplotlib, Seaborn, and Scikit-learn.

The complete experimental process was performed in a Jupyter Notebook environment named MLModelComparison.ipynb.

© 2026 IJMRSET | Volume 9, Issue 7, July 2026| DOI:10.15680/IJMRSET.2026.0907019

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 9.864| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Table 1: Software Environment Used for Implementation

Component	Description
Programming Language	Python
Development Environment	Jupyter Notebook
Data Processing	Pandas, NumPy
Visualization	Matplotlib, Seaborn
Machine Learning	Scikit-learn
Dataset Source	Sklearn Built-in Dataset

A. Software Environment

The following tools and libraries were used for implementation (Table 1). The experimental setup provides a controlled environment for comparing different machine learning algorithms and analyzing their predictive performance.

B. Dataset Description

The Breast Cancer Wisconsin Diagnostic dataset is used in this research for evaluating the performance of machine learning classification algorithms. The dataset is publicly available through the Scikit-learn library and is widely used as a benchmark dataset for binary classification problems.

The dataset consists of 569 data samples with 30 numerical input features. These features are computed from digitized images of breast mass samples and represent different characteristics of cell nuclei [9].

The objective of the classification task is to predict whether a sample belongs to a malignant or benign class. The target variable contains two class labels representing the nature of the sample.
The dataset characteristics are summarized in Table 2.

Table 2: Characteristics of Breast Cancer Wisconsin Dataset

Property	Description
Dataset Name	Breast Cancer Wisconsin Diagnostic
Total Instances	569
Number of Features	30
Feature Type	Numerical
Target Classes	2
Missing Values	0
Source	Scikit-learn Library

The dataset was selected because it provides a clean and structured environment for evaluating classification algorithms. Since the dataset does not contain missing values, additional data cleaning procedures were not required.

C. Exploratory Data Analysis

Exploratory Data Analysis (EDA) was performed before model development to understand the characteristics and distribution of the dataset.
The initial analysis included examining the dataset structure, identifying feature information, checking missing values, and analyzing class distribution.
The dataset was loaded using the Scikit-learn function:
loadbreastcancer()
The feature matrix and target variable were separated as follows:

$$X = \text{Features}, \qquad y = \text{Target} \tag{1}$$

© 2026 IJMRSET | Volume 9, Issue 7, July 2026| DOI:10.15680/IJMRSET.2026.0907019

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 9.864| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

The dataset information was analyzed to verify the number of samples, features, and data types. Missing value analysis was performed to ensure data quality. The analysis confirmed that the dataset contained zero missing values.

The class distribution was also visualized to understand the balance between malignant and benign samples. Balanced class distribution helps improve the reliability of machine learning model evaluation.

D. Data Preprocessing

Data preprocessing is an essential step in machine learning because the quality of input data directly affects model performance.

The following preprocessing operations were performed:

- Separation of feature variables and target labels.
- Splitting the dataset into training and testing subsets.
- Feature scaling using StandardScaler.
- Preparing the processed data for model training.

The dataset was divided into training and testing sets using an 80:20 ratio. The training dataset contained 455 ×30 samples, while the testing dataset contained 114 ×30 samples.

E. Feature Scaling

Feature scaling was applied using StandardScaler from the Scikit-learn library. Scaling is necessary because machine learning algorithms may perform differently when input features have different numerical ranges.

Standardization transforms the feature values using the following equation:

$$z = \frac{x - \mu}{\sigma} \quad (2)$$

where:

- x represents the original feature value.
- μ represents the mean of the feature.
- σ represents the standard deviation.
- z represents the scaled feature value.

Feature scaling improves the performance of distance-based and margin-based algorithms, particularly Support Vector Machine and K-Nearest Neighbor.

V. MACHINE LEARNING MODELS

Five supervised machine learning classification algorithms were selected for comparative evaluation. These algorithms represent different learning approaches, including linear classification, distance-based learning, kernel-based learning, and ensemble learning.

A. Logistic Regression

Logistic Regression is a statistical classification algorithm used for predicting categorical outcomes. It estimates the probability of class membership using a logistic function. The algorithm is widely used because of its simplicity, fast execution, and ability to provide interpretable results.

In this study, Logistic Regression is used as a baseline classification model for comparison with other advanced algorithms.

B. Support Vector Machine

Support Vector Machine (SVM) is a powerful supervised learning algorithm designed for classification problems.

The main objective of SVM is to identify an optimal hyperplane that separates different classes with maximum margin. SVM performs effectively on high-dimensional datasets because it focuses on maximizing the distance between class boundaries.

Due to its strong generalization ability, SVM is widely applied in medical classification, image analysis, and pattern recognition applications.

© 2026 IJMRSET | Volume 9, Issue 7, July 2026| DOI:10.15680/IJMRSET.2026.0907019

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 9.864| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

C. K-Nearest Neighbor

K-Nearest Neighbor (KNN) is a non-parametric classification algorithm that predicts the class of a sample based on the majority class among its nearest neighboring samples.

The algorithm calculates the distance between samples and assigns the class label based on similarity.

Although KNN is simple and easy to understand, its performance depends on feature scaling and the selection of an appropriate value of K.

D. Random Forest

Random Forest is an ensemble learning algorithm that combines multiple decision trees to improve classification performance.

Each decision tree generates an individual prediction, and the final output is determined through majority voting.

Random Forest reduces overfitting and provides stable performance by combining multiple weak learners into a stronger predictive model.

E. Gradient Boosting

Gradient Boosting is another ensemble learning technique that builds models sequentially.

Each new model attempts to reduce the errors produced by previous models, resulting in improved prediction accuracy.

Gradient Boosting is effective for handling complex relationships within datasets and has been widely used in predictive analytics applications.

F. Evaluation Metrics

The performance of all machine learning models was evaluated using four standard classification metrics.

Accuracy

Accuracy measures the overall percentage of correctly classified samples.

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

Precision

Precision represents the proportion of correctly predicted positive samples among all predicted positive samples.

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

Recall

Recall measures the ability of a model to correctly identify actual positive samples.

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

F1-Score

F1-score is the harmonic mean of precision and recall and provides a balanced evaluation of classification performance.

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{6}$$

© 2026 IJMRSET | Volume 9, Issue 7, July 2026| DOI:10.15680/IJMRSET.2026.0907019

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 9.864| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

VI. EXPERIMENTAL RESULTS

The experimental evaluation was performed to compare the predictive performance of five machine learning classification models. All models were trained using the same training dataset and evaluated using the same testing dataset to ensure a fair comparison.

The performance analysis was conducted using accuracy, precision, recall, and F1-score metrics. These evaluation measures provide a comprehensive understanding of model prediction capability by considering both correctly classified and incorrectly classified samples.

The obtained results from all machine learning models are presented in Table 3.

Table 3: Performance Comparison of Machine Learning Models

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression	97.36%	97.22%	98.59%	97.90%
Support Vector Machine	98.24%	97.26%	100%	98.61%
K-Nearest Neighbor	94.73%	95.77%	95.77%	95.77%
Random Forest	96.49%	95.89%	98.59%	97.22%
Gradient Boosting	95.61%	95.83%	97.18%	96.50%

The experimental results indicate that Support Vector Machine achieved the highest performance among all evaluated algorithms.

The SVM model obtained an accuracy of 98.24%, which is higher than Logistic Regression, Random Forest, Gradient Boosting, and K-Nearest Neighbor.

The F1-score of 98.61% demonstrates that SVM maintained a strong balance between precision and recall. The recall value of 100% indicates that the model successfully identified all positive class samples in the testing dataset.

A. Performance Visualization

To provide a clear comparison between different machine learning algorithms, graphical analysis was performed.

The class distribution visualization helps to understand the distribution of target classes within the dataset. It provides an overview of the number of samples belonging to malignant and benign categories.

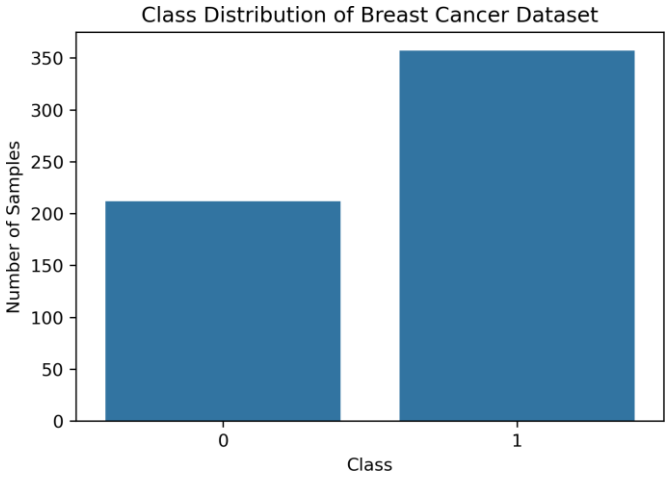


Figure 1: Class distribution of Breast Cancer Wisconsin dataset.

The accuracy comparison graph presents the performance differences among all five machine learning models.

As shown in Figure 2, Support Vector Machine achieved the highest accuracy compared with other classification algorithms.

© 2026 IJMRSET | Volume 9, Issue 7, July 2026| DOI:10.15680/IJMRSET.2026.0907019

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 9.864| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

The visualization demonstrates that kernel-based learning methods can achieve better classification performance when dealing with high-dimensional numerical datasets.

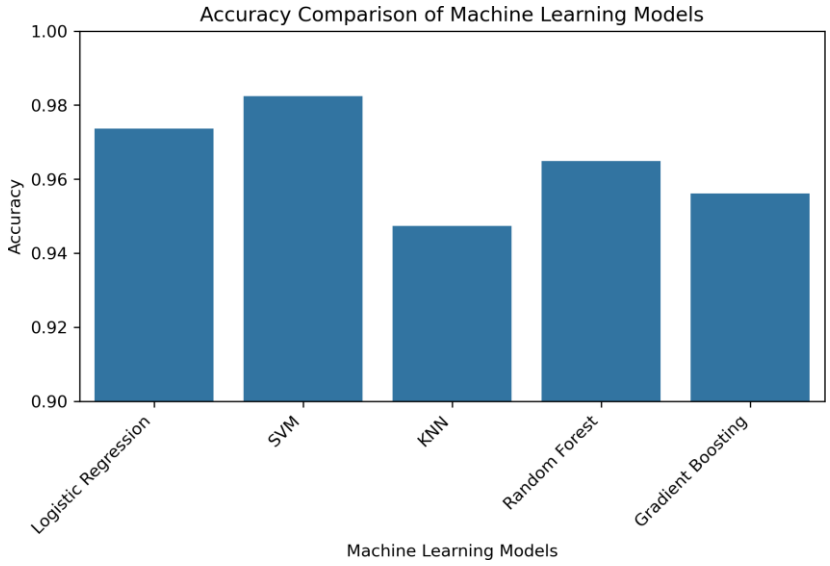


Figure 2: Accuracy comparison of machine learning models.

B. Confusion Matrix Analysis

A confusion matrix was generated for the best-performing Support Vector Machine model to analyze classification outcomes in detail.

The confusion matrix provides information about true positive, true negative, false positive, and false negative predictions.

The generated confusion matrix for the SVM model is shown in Figure 3.

The confusion matrix indicates that the SVM model produced highly accurate predictions with very few classification errors.

The lower number of incorrect predictions demonstrates the effectiveness of SVM in separating different classes within the dataset.

VII. DISCUSSION

The comparative analysis demonstrates that machine learning algorithms provide effective solutions for predictive analytics classification tasks.

Among all evaluated models, Support Vector Machine achieved the best overall performance. This superior performance can be attributed to its ability to identify an optimal separating boundary between classes.

The Breast Cancer Wisconsin dataset contains multiple numerical features with different characteristics. Since SVM uses a margin-based learning approach, it can effectively handle complex relationships among high-dimensional features.

Feature scaling also contributed significantly to improving model performance. StandardScaler transformed all feature

© 2026 IJMRSET | Volume 9, Issue 7, July 2026| DOI:10.15680/IJMRSET.2026.0907019

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 9.864| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

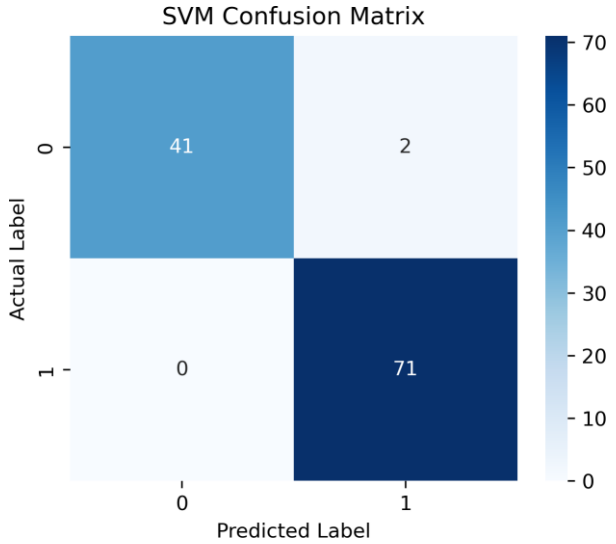


Figure 3: Confusion matrix of Support Vector Machine model.

values into a common range, allowing algorithms such as SVM and KNN to perform more effectively.

Logistic Regression achieved strong performance with an accuracy of 97.36%. The model provides simple interpretation and requires less computational complexity, making it useful for applications where explainability is important.

Random Forest achieved an accuracy of 96.49% due to its ensemble-based learning mechanism. The combination of multiple decision trees improves stability and reduces overfitting.

Gradient Boosting achieved competitive performance by sequentially improving weak learners. However, its accuracy was slightly lower compared with SVM because of differences in learning strategy and dataset characteristics.

K-Nearest Neighbor obtained the lowest accuracy among the evaluated algorithms. Since KNN relies on distance calculations, its performance can be affected by the distribution of feature values and selection of neighboring samples. Overall, the experimental findings show that no single algorithm is universally optimal for all classification problems. The effectiveness of a model depends on dataset properties, preprocessing techniques, and algorithm characteristics. For the selected dataset, Support Vector Machine provided the most reliable predictive performance and can be considered the most suitable model among the evaluated approaches.

VIII. CONCLUSION

This research presented a comparative performance analysis of machine learning models for predictive analytics. The main objective of this study was to evaluate and compare different supervised learning algorithms under the same experimental environment.

Five machine learning classification algorithms, namely Logistic Regression, Support Vector Machine, K-Nearest Neighbor, Random Forest, and Gradient Boosting, were implemented using the Breast Cancer Wisconsin Diagnos-tic dataset.

The complete machine learning workflow included dataset loading, exploratory data analysis, missing value analysis, train-test splitting, feature scaling, model training, and performance evaluation.

The performance of the developed models was evaluated using four important classification metrics: accuracy, precision, recall, and F1-score. These metrics provided a detailed understanding of model prediction capability and reliabil-ity.

The experimental results showed that Support Vector Machine achieved the best performance among all evaluated algorithms. The model obtained an accuracy of 98.24% and an F1-score of 98.61%, demonstrating its effectiveness for classification of high-dimensional numerical data.

The study also revealed that preprocessing techniques such as feature scaling play an important role in improving machine learning performance. Proper algorithm selection combined with suitable preprocessing methods can significantly enhance predictive analytics systems.

The comparative evaluation presented in this research provides useful insights into the behavior of different machine learning algorithms and helps in selecting suitable models for classification-based predictive applications.

IX. FUTURE SCOPE

Although the proposed approach achieved promising results, several improvements can be explored in future research. Future work may focus on applying advanced optimization techniques such as Grid Search, Random Search, and Bayesian Optimization to improve model performance by identifying optimal hyperparameters.

The study can also be extended by incorporating cross-validation techniques to obtain more reliable performance estimates and reduce dependency on a single train-test split.

Another possible extension is the comparison of traditional machine learning approaches with advanced deep learning

techniques such as Artificial Neural Networks, Convolutional Neural Networks, and Transformer-based models. Explainable Artificial Intelligence (XAI) techniques can also be integrated to improve model transparency and provide better understanding of prediction decisions.

Furthermore, evaluating the proposed methodology on larger and more diverse real-world datasets can improve the generalizability of the research findings.

The development of automated machine learning pipelines can also be considered to simplify model selection, training, and deployment processes for predictive analytics applications.

X. ACKNOWLEDGMENT

The authors would like to thank the Department of Computer Science with Data Analytics, Dr. N.G.P. Arts and Science College, for providing the necessary academic support and resources for completing this research work.

REFERENCES

[1] C. Cortes and V. Vapnik, “Support-vector networks,” *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.
[2] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, pp. 5–32, 2001.
[3] F. Pedregosa et al., “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
[4] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, 2009.
[5] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.
[6] T. Cover and P. Hart, “Nearest neighbor pattern classification,” *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967.
[7] J. H. Friedman, “Greedy function approximation: A gradient boosting machine,” *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.
[8] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in *Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.
[9] W. N. Street, W. H. Wolberg, and O. L. Mangasarian, “Nuclear feature extraction for breast tumor diagnosis,” in *Proc. IS&T/SPIE Int. Symp. on Electronic Imaging*, vol. 1905, 1993, pp. 861–870.

[10] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*, 3rd ed., Wiley, 2013.
[11] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proc. 14th Int. Joint Conf. on Artificial Intelligence (IJCAI)*, 1995, pp. 1137–1143.
[12] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd ed., Morgan Kaufmann, 2011.
[13] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed., Morgan Kaufmann, 2011.
[14] A. Ge´ron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*, 2nd ed., O'Reilly Media, 2019.
[15] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119–139, 1997.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com