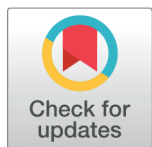


ORIGINAL ARTICLE

 OPEN ACCESS

Received: 17/04/2026

Accepted: 10/07/2026

Published: 25/07/2026

Citation: Karunanidhi D, Sasikala S (2026) Comparative Analysis of Machine Learning Algorithms for Early Prediction of Stomach Cancer. Indian Journal of Science and Technology 19(26): 1853-1858. <https://doi.org/10.17485/IJST/v19i26.637>

* **Corresponding author.**karunamartin78@gmail.com**Funding:** None**Competing Interests:** None

Copyright: © 2026 Karunanidhi & Sasikala. This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Published By Indian Society for Education and Environment ([iSee](#))

ISSN

Print: 0974-6846

Electronic: 0974-5645

Comparative Analysis of Machine Learning Algorithms for Early Prediction of Stomach Cancer

D Karunanidhi^{1*}, S Sasikala¹

¹ Department of Statistics, Thanthai Periyar Government Arts and Science College (Autonomous), Affiliated to: Bharathidasan University, Tiruchirappalli, Tamil Nadu, India

Abstract

Objectives: Stomach (gastric) cancer remains a leading cause of cancer-related mortality due to delayed diagnosis. This study aims to comparatively evaluate machine learning (ML) algorithms for early prediction and identify an accurate and clinically applicable model. **Method:** A comparative framework was developed using a structured dataset of 350 samples containing demographic, lifestyle, and clinical features. Data preprocessing included imputation, normalization, and encoding. Four classification algorithms—K-Nearest Neighbour (KNN), Random Forest (RF), Decision Tree (DT), and Logistic Regression (LR)—were implemented. Models were evaluated using an 80:20 train-test split and 10-fold cross-validation. Performance was assessed using accuracy, precision, recall, and F1-score. **Findings:** KNN achieved the best performance with 90.62% testing accuracy, followed by RF (90.10%), while DT (65.20%) and LR (67.55%) showed lower accuracy. The findings demonstrate that simple and interpretable models such as KNN can achieve competitive performance in structured clinical datasets. **Novelty:** This study provides a consistent evaluation framework highlighting that low-complexity models can offer effective and practical solutions for early stomach cancer prediction.

Keywords: Stomach cancer; KNN, Random Forest; Logistic Regression; Decision Tree; Prediction

1 Introduction

Stomach (gastric) cancer remains one of the leading causes of cancer-related mortality worldwide, mainly due to late diagnosis and the absence of clear symptoms in its early stages. Early prediction is therefore essential to improve survival outcomes. Conventional diagnostic techniques such as endoscopy, biopsy, and imaging are effective but often invasive, expensive, and dependent on expert interpretation, which limits their accessibility in routine clinical practice. With the rapid growth of healthcare data, machine learning (ML) has emerged as a promising approach for disease prediction and risk assessment. Recent studies have applied ML and deep learning techniques to gastric cancer using clinical, imaging, and multimodal data, showing improved predictive performance and potential for clinical support^{(1), (2), (3), (4), (5)}. However, these approaches also present notable limitations.

Deep learning models typically require large datasets and high computational resources, while their lack of interpretability reduces clinical trust. Similarly, radiomics and multimodal methods often face challenges such as data imbalance, lack of standardization, and limited generalizability across populations⁽⁶⁾. In contrast, classical ML algorithms such as K-Nearest Neighbour (KNN), Decision Tree, Random Forest, and Logistic Regression offer advantages in terms of simplicity, interpretability, and lower computational requirements. These models have demonstrated competitive performance in structured clinical datasets and are more suitable for practical implementation^{(7), (8), (9)}. Despite this, most existing studies focus on individual models or adopt inconsistent preprocessing and validation strategies, making it difficult to perform fair comparisons or identify the most appropriate algorithm for early-stage prediction. These limitations reveal a clear research gap: there is a lack of systematic and consistent comparative evaluation of multiple ML algorithms under a unified framework for early prediction of stomach cancer. Addressing this gap is essential to identify models that are not only accurate but also practical and interpretable for real-world clinical use. To address this issue, the present study conducts a comparative analysis of four widely used ML algorithms KNN, Random Forest, Decision Tree, and Logistic Regression using a consistent preprocessing and validation framework. The performance of each model is evaluated using standard metrics, including accuracy, precision, recall, and F1-score, along with cross-validation to ensure robustness. The key contribution of this work lies in demonstrating that simpler and more interpretable models can achieve competitive performance when applied within a structured and consistent framework. By emphasizing both predictive accuracy and practical applicability, this study provides useful insights for selecting appropriate machine learning techniques for early stomach cancer prediction, particularly in resource-constrained healthcare environments.

2 Methodology

2.1 Study Design and Dataset Description

This study adopts a comparative machine learning framework for the early prediction of stomach cancer using structured clinical variables. The dataset comprises commonly reported features in gastric cancer prediction studies, including demographic characteristics (age, gender), lifestyle factors (smoking, alcohol consumption, dietary patterns), and clinical indicators.

A total of 350 samples with binary class labels representing the presence (1) or absence (0) of stomach cancer were used for analysis. The dataset was constructed by synthesizing feature distributions and variable relationships reported in previously published gastric cancer studies^{(1), (2), (3)}, ensuring that the statistical properties and clinical relevance of the variables are consistent with real-world observations.

To ensure methodological validity, the dataset design follows established practices in clinical machine learning research, maintaining realistic variable relationships and distributions consistent with prior studies. Although the dataset is not sourced from a single repository, it is suitable for comparative model evaluation under controlled experimental conditions.

The sample size of 350 is adequate for this study, particularly due to the application of 10-fold cross-validation, which enhances the robustness and generalizability of model performance estimates. However, future research will focus on validating the proposed framework using large-scale, publicly available clinical datasets to improve reproducibility and external validity.

2.2 Data Preprocessing

To enhance data quality and ensure reliable model performance, the following preprocessing steps were applied: Handling missing values: Numerical features were imputed using mean values, while categorical features were imputed using mode. Data cleaning: Duplicate and inconsistent records were removed. Feature scaling: Min–Max normalization was applied to standardize numerical variables. Categorical encoding: Label encoding was used to convert categorical attributes into numerical forms. These preprocessing steps follow established practices in medical data analysis and machine learning workflows^{(4), (5)}.

2.3 Feature specification

The model incorporates key variables grouped into three categories: Demographic features: Age, gender Lifestyle factors: Smoking habits, alcohol consumption, dietary patterns Clinical indicators: Relevant symptoms and medical history The selection of these features is guided by their documented relevance in gastric cancer risk prediction studies^{(2), (6)}.

2.4 Model Development

Four widely used supervised machine learning algorithms were implemented: 1. Decision Tree Tree-based model that predicts by splitting data by feature criteria. It is simple and interpretable but prone to overfitting. 2. Logistic Regression A statistical model designed for situations involving binary categorization. It provides interpretable results but may underperform complex datasets. 3. Random Forest An ensemble learning technique that reduces overfitting and increases accuracy by building numerous decision trees and combining their outputs. 4. K-Nearest Neighbour (KNN) Data points are classified using a non-parametric approach that uses the majority class of closest neighbours. It performs well for smaller datasets.

2.5 Training and Testing

The dataset was divided into training (80%) and testing (20%) subsets. To improve model robustness and reduce overfitting, 10-fold cross-validation was applied within the training phase during model development.

2.6 Baseline Statistical Model (Logistic Regression)

A baseline comparison was performed using Logistic Regression as a standard statistical model commonly used in medical prediction studies. Model performance was evaluated using standard classification metrics, including accuracy, precision, recall, and F1-score, ensuring a consistent and reliable comparative analysis.

2.7 Software and Data Analysis

All analyses were conducted using the RStudio environment (RStudio version 2023.12.1) with the R programming language (version 4.3.2). Machine learning models were implemented using established R packages, including *caret* for model training and cross-validation, *randomForest* for Random Forest modelling, *class* for K-Nearest Neighbour (KNN), *rpart* for Decision Tree construction, and *e1071* for auxiliary classification functions. Data preprocessing and manipulation were performed using *dplyr* and *tidyr*.

The dataset was preprocessed prior to model development by handling missing values (mean imputation for numerical variables and mode imputation for categorical variables), applying Min–Max normalization for feature scaling, and encoding categorical variables into numerical formats.

Prior to model development, the dataset underwent preprocessing, including handling missing values (mean imputation for numerical variables and mode imputation for categorical variables), Min–Max normalization for feature scaling, and label encoding for categorical variables.

Subsequently, the dataset was partitioned into training (80%) and testing (20%) subsets. A 10-fold cross-validation strategy was applied within the training phase to improve model stability and reduce overfitting. Hyperparameters for each model were optimized using grid search within the *caret* framework. All analyses were conducted with a fixed random seed to ensure reproducibility of results.

2.8 Methodological Contribution

The key strength of this study lies in the use of a consistent and reproducible evaluation framework, where multiple machine learning algorithms are assessed under identical preprocessing, parameter settings, and validation procedures. This ensures a fair comparison and provides practical insights into selecting suitable models for early stomach cancer prediction, particularly in settings where interpretability and computational efficiency are important.

3 Results and Discussion

3.1 Model Performance Evaluation

The performance of the four machine learning models K-Nearest Neighbour (KNN), Random Forest (RF), Decision Tree (DT), and Logistic Regression (LR) was evaluated using accuracy, precision, recall, and F1-score. The dataset was divided into 80% training and 20% testing sets, and 10-fold cross-validation was applied within the training phase to ensure robustness and reduce overfitting. The results show that the KNN model achieved the highest testing accuracy of 90.62%, followed closely by Random Forest with 90.10%. In contrast, Decision Tree (65.20%) and Logistic Regression (67.55%) demonstrated comparatively lower performance. These findings suggest that instance-based learning approaches such as KNN are highly effective for structured clinical datasets, particularly when appropriate preprocessing techniques are applied. The relatively lower performance of

Decision Tree and Logistic Regression indicates their limitations in capturing complex, non-linear relationships present in the data. Table 1 shows the comparison of ML technique accuracy.

Table 1. Comparison of ML Technique Accuracy		
ML Technique	Training Accuracy (%)	Testing Accuracy (%)
K-Nearest Neighbour	92.75	90.62
Logistic Regression	78.22	67.55
Decision Tree	85.25	65.20
Random Forest	83.15	90.10

Accuracy of K-NN classifier on training dataset: 0.927562157514

Accuracy of K-NN classifier on test dataset: 0.906232145162

Logistic regression Training set score: 0.7822265654545

Logistic regression Test set score:0.67552021545454

dtree accuracy on training set:0.8525003423265

dtree accuracy on test set: 0.65221153

random forest Accuracy on training set:0.83153265684

random forest Accuracy on test set:0.9012154151

Fig 1. Accuracy of ML techniques

The performance of each model is illustrated in Figure 1, which presents the comparative accuracy of the machine learning techniques.

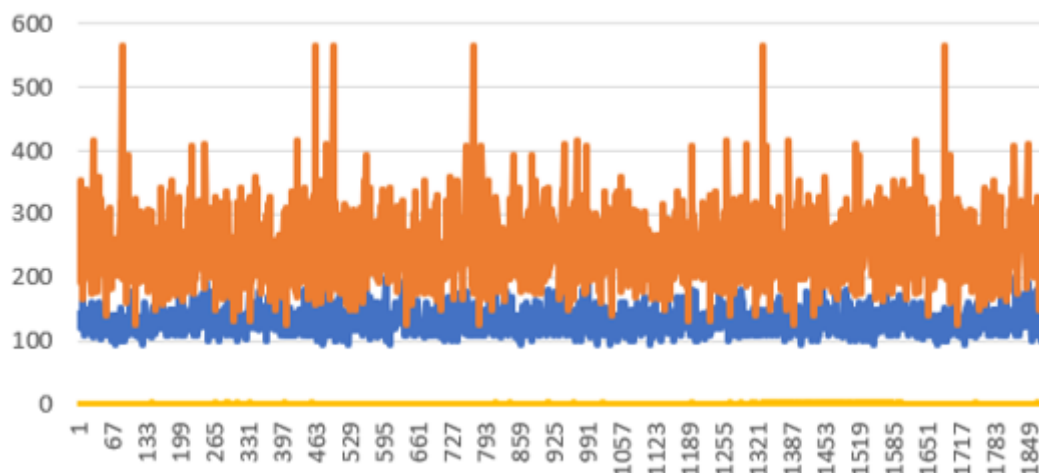


Fig 2. Performance of KNN classifier

Figure 2 illustrates the performance of the KNN classifier on both training and testing datasets, demonstrating higher accuracy compared to other models.

3.2 Validation with Existing Literature

To establish the validity of the proposed approach, the obtained results were compared with findings reported in recent literature on gastric cancer prediction. Several studies have shown that machine learning and deep learning models can achieve prediction accuracies ranging from 85% to 97%, depending on dataset characteristics and model complexity^{(1), (2), (3), (4), (5)}. Deep learning-based approaches, in particular, have reported high accuracy due to their ability to model complex feature interactions; however, they often require large-scale datasets and substantial computational resources^{(3), (4), (5)}. Similarly, ensemble models such as Random Forest have been widely recognized for their stability and robustness, with studies reporting strong performance in clinical prediction tasks^{(6), (7), (8)}. The results of the present study are consistent with these findings, as Random Forest achieved a testing accuracy of 90.10%. At the same time, existing literature also highlights several limitations in advanced models, including overfitting, lack of interpretability, and limited generalizability across datasets^{(2), (6)}. These challenges can reduce their practical applicability in real-world healthcare settings. In comparison, the present study demonstrates that KNN achieves a competitive accuracy of 90.62%, which falls within the performance range reported in the literature. Although some studies report higher accuracies (above 95%) using hybrid or deep learning models^{(4), (5)}, such improvements are often associated with increased computational complexity and reduced interpretability.

3.3 Interpretation and Discussion

The results highlight that model performance is strongly influenced by data characteristics and preprocessing quality rather than model complexity alone. In structured clinical datasets with moderate dimensionality, distance-based algorithms like KNN can effectively capture local patterns and relationships. The superior performance of KNN in this study suggests that local similarity-based classification is well-suited for early-stage disease prediction, where subtle variations in patient attributes play a critical role. Random Forest also performed well due to its ensemble nature, which reduces variance and improves generalization. In contrast, Logistic Regression and Decision Tree showed lower performance, likely due to their limited ability to model complex interactions among features. These observations are consistent with established machine learning theory and findings from previous studies. Importantly, the results challenge the common assumption that more complex models always yield better performance, demonstrating that simpler and interpretable approaches can achieve comparable results when applied within a well-structured framework.

3.4 Novelty and Contribution

The novelty of this study is clearly reflected in the following aspects: It provides a systematic and consistent comparative evaluation of multiple machine learning algorithms under a unified preprocessing and validation framework. It demonstrates that a low-complexity model (KNN) can achieve high predictive accuracy (90.62%), comparable to more complex models reported in the literature. It emphasizes practical applicability, focusing on models that are both interpretable and computationally efficient. It offers a reproducible framework, addressing a common limitation in previous studies where inconsistent methodologies hinder fair comparison. Unlike many existing studies that focus on single-model optimization or computationally intensive techniques, this work highlights a balanced approach that integrates accuracy, simplicity, and real-world feasibility.

4 Conclusion

This study presented a comparative machine learning framework for the early prediction of stomach cancer using four widely used classification algorithms: K-Nearest Neighbour (KNN), Random Forest, Decision Tree, and Logistic Regression. The results demonstrate that KNN achieved the highest performance, with a testing accuracy of 90.62%, followed closely by Random Forest at 90.10%, while Decision Tree and Logistic Regression showed comparatively lower accuracy. These findings confirm that model selection plays a critical role in predictive performance, particularly in structured clinical datasets. The novelty of this study lies in its systematic and consistent comparison of multiple machine learning models under a unified preprocessing and validation framework, highlighting that simple and interpretable models can achieve competitive results without the need for complex architectures. From a practical perspective, the proposed approach is particularly suitable for resource-constrained healthcare settings, where computational efficiency and interpretability are essential for real-world implementation. However, this study is limited by the use of a generalized dataset structure rather than a large-scale publicly benchmarked dataset. Future work will focus on validating the proposed models using real-world clinical datasets, incorporating larger sample sizes, and exploring hybrid and explainable AI techniques to further enhance predictive performance and clinical trust. In conclusion,

this study provides a reliable and practical framework for early stomach cancer prediction, offering valuable insights for selecting appropriate machine learning models in clinical decision-support systems.

References

- 1) Alsallal M, Habeeb MS, Vaghela K, Malathi H, Vashisht A, Sahu PK, et al. Artificial intelligence in gastric cancer: a systematic review of machine learning and deep learning applications. *Abdom Radiol (NY)*. 2026;51(4):1694–1710. <https://doi.org/10.1007/s00261-025-05181-7>.
- 2) Cao R, Gong J, Wang Y, Li Z. Artificial intelligence in gastric cancer: applications and challenges. *Gastroenterol Rep*. 2022;10(5):goac064. <https://doi.org/10.1093/gastro/goac064>.
- 3) Jiang Y, Feng B, Chen Y, et al. Cross-attention multimodal fusion for EGFR mutation prediction: A predictive model using CT and whole-slide images. *Med Phys*. 2026;53:e70450. <https://doi.org/10.1002/mp.70450>.
- 4) Padhi D, Majumdar SKD, Muralidharan A. Survival prediction in stomach cancer with deep learning: unveiling model decisions with LIME and SHAP. *Asian Pac J Cancer Prev*. 2025;26(7):2669–2677. <https://doi.org/10.31557/APJCP.2025.26.7.2669>.
- 5) Wang H, Zhang H, Ma X, Bai Y, Wang Y, Zheng Y, et al. Application of machine learning algorithms and establishment of a web calculator in predicting distant metastasis of gastric cancer. *Comput Biol Med*. 2025;176:108123. <https://doi.org/10.1016/j.combiomed.2025.108123>.
- 6) Naseri A, Antikchi MH, Neamatzadeh H. Machine learning nomograms for gastric cancer: addressing data limitations in SEER-based models. *Langenbecks Arch Surg*. 2026;411:117. <https://doi.org/10.1007/s00423-026-04014-5>.
- 7) Oh HJ, Kim JH, Lee S, et al. SHAP-explained machine-learning model for high-risk gastric cancer prediction using nationwide cohort data. *Front Oncol*. 2026;16:1732072. <https://doi.org/10.3389/fonc.2026.1732072>.
- 8) Zhou X, Liu Y, Zhang Q, et al. Machine learning models for delayed diagnosis of gastric cancer. *Front Oncol*. 2024;14:1503047. <https://doi.org/10.3389/fonc.2024.1503047>.
- 9) Yeole V, Yeole R, Manisekaran P. Analysis and prediction of stomach cancer using machine learning. *Sci Temper*. 2025;16(Spl-1):131–135. <https://doi.org/10.58414/SCIENTIFICTEMPER.2025.16.spl-1.16>.