IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p-ISSN: 2278-8727, Volume 28, Issue 2, Ser. 1 (Mar. – Apr. 2026), PP 29-35
www.iosrjournals.org

Machine Learning-Based Cancer Risk Prediction Using Clinical And Lifestyle Factors

Thi Hong Loan Nguyen
(HaUI School Of Information & Communications Technology, Hanoi University Of Industry (Haui), Hanoi, Vietnam,

Abstract:
Background*: Cancer has become one of the leading causes of mortality worldwide, with increasing incidence and a noticeable shift toward younger age groups in many countries. Early identification of cancer risk is therefore essential for improving prevention strategies and supporting timely medical intervention. Recent advances in machine learning have enabled the analysis of clinical and lifestyle data to identify potential disease risks and support decision-making in healthcare. However, selecting an appropriate predictive model remains an important challenge in developing reliable prediction systems. Therefore, this study aims to develop a machine learning–based approach for cancer risk prediction using clinical and lifestyle factors and to determine the most effective predictive model through comparative analysis.*
Materials and Methods*: In this study, a dataset containing several clinical and lifestyle risk indicators was used, including age, gender, body mass index (BMI), smoking status, genetic risk, physical activity level, alcohol consumption, and cancer history. The dataset was divided into training and testing sets using stratified sampling to preserve class distribution. Several machine learning algorithms were implemented and compared, including Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), k-Nearest Neighbors (KNN), and AdaBoost. Hyperparameter optimization was performed using GridSearch combined with stratified cross-validation to improve model performance and ensure reliable evaluation.*
Results*: The experimental results show that the evaluated machine learning models are capable of predicting cancer risk using clinical and lifestyle features. Among the tested algorithms, the AdaBoost model achieved the best predictive performance with an accuracy of 0.96 and a ROC-AUC of approximately 0.97. Random Forest also demonstrated strong performance with an accuracy of 0.94 and a ROC-AUC of about 0.96. Other models such as KNN, SVM, Decision Tree, and Logistic Regression showed comparatively lower predictive performance.*
Conclusion: *The results indicate that the AdaBoost-based prediction approach can effectively identify potential cancer risk using clinical and lifestyle data. This method has potential application in supporting early cancer risk assessment and assisting the development of data-driven decision support systems in healthcare.*
Key Word: *Cancer risk prediction; Machine learning; AdaBoost; Clinical data; Lifestyle factors.*

Date of Submission: 09-03-2026 Date of Acceptance: 19-03-2026

I. Introduction

Cancer remains one of the leading causes of mortality worldwide and represents a significant public health challenge. According to global cancer statistics reported by the World Health Organization (WHO) and the GLOBOCAN database, approximately 19.3 million new cancer cases and nearly 10 million cancer-related deaths were reported worldwide in 2020. Among different cancer types, breast cancer is one of the most frequently diagnosed cancers globally. Early detection and accurate risk prediction are therefore essential for improving treatment outcomes and reducing mortality rates.

Traditional cancer diagnosis relies primarily on clinical examination, medical imaging, and pathological analysis. Although these approaches remain fundamental in clinical practice, they often depend heavily on expert interpretation and may not always detect cancer in its early stages. With the rapid growth of healthcare data and advances in computational technologies, machine learning has emerged as a promising approach for disease prediction and medical decision support. Machine learning algorithms can identify complex patterns in clinical datasets and assist healthcare professionals in detecting diseases more accurately and efficiently [1], [2].

Several studies have investigated cancer prediction using machine learning techniques. Kumari and Singh [3] developed a breast cancer prediction system using data mining techniques and demonstrated that machine learning algorithms can effectively classify malignant and benign tumors. Similarly, Naji et al. [4] evaluated multiple machine learning models, including Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), and k-Nearest Neighbor (KNN) and reported promising classification

performance. Austria et al. [5] conducted a comparative analysis using the Coimbra Breast Cancer Dataset and found that advanced models such as Gradient Boosting achieved strong predictive performance.

Other studies have explored different machine learning frameworks for improving prediction accuracy. Gupta et al. [6] proposed a machine learning-based model for breast cancer prediction and highlighted the importance of proper model selection and data preprocessing. Ensemble learning approaches have also been investigated to enhance classification performance. For example, Assegie et al. [7] introduced a breast cancer prediction model combining Decision Tree and AdaBoost, demonstrating that ensemble methods can improve predictive accuracy compared with individual classifiers. Similarly, ensemble techniques have been successfully applied in other medical prediction tasks. Hang et al. [8] proposed a stacking ensemble learning framework for heart disease risk prediction and showed that integrating multiple machine learning models can significantly improve predictive performance and robustness. In addition, Yarabarla et al. [9] applied machine learning techniques for breast cancer prediction and demonstrated their potential for supporting early disease detection.

Despite these advances, several limitations remain in previous studies. Many existing works primarily focus on image-based diagnosis or employ limited sets of clinical features, which may not fully capture the complex factors influencing cancer risk. However, cancer risk is often affected by a combination of clinical characteristics and lifestyle factors such as smoking habits, physical activity levels, and genetic predisposition. Furthermore, the predictive performance of machine learning models may vary depending on the dataset and experimental settings, making it necessary to conduct systematic comparative studies under consistent experimental conditions.

Motivated by these challenges, this study investigates the use of machine learning techniques for cancer risk prediction using clinical and lifestyle factors. Multiple machine learning algorithms are implemented and comparatively evaluated in order to identify the model that achieves the best predictive performance. The objective of this study is to improve the accuracy of cancer risk prediction and to contribute to the development of data-driven decision support systems for healthcare applications.

The remainder of this paper is organized as follows. Section II describes the materials and methods used in this study, including the dataset, feature description, and machine learning models applied for cancer risk prediction. Section III presents the experimental results obtained from the comparative evaluation of different machine learning algorithms. Section IV discusses the findings and analyzes the performance of the proposed approach. Finally, Section V concludes the paper and outlines potential directions for future research.

II. Materials And Methods

Dataset Description

In this study, a structured dataset containing demographic, clinical, and lifestyle-related variables was used to develop a machine learning model for cancer risk prediction. The dataset was obtained from the Kaggle data repository, specifically the "Cancer Prediction Dataset" available online [10]. The dataset includes eight explanatory variables: Age, Gender, Body Mass Index (BMI), Smoking status, Genetic Risk, Physical Activity level, Alcohol Intake, and Cancer History, along with a binary target variable, Diagnosis, which indicates whether the individual was diagnosed with cancer (1) or not (0).

Each record in the dataset represents a single individual. The variables capture different aspects that may influence cancer development. Age reflects the demographic factor associated with increased cancer incidence over time. Gender represents biological differences that may influence disease susceptibility. BMI indicates body composition and obesity-related risk. Smoking represents exposure to tobacco-related carcinogens, while Alcohol Intake reflects lifestyle behaviors associated with several cancer types. Genetic Risk represents hereditary predisposition, and Cancer History captures previous cancer-related medical history. Physical Activity is included as a behavioral factor that may influence overall health and disease risk.

To evaluate the predictive capability of the models, the dataset was divided into training and testing subsets using a stratified sampling strategy to maintain class balance. Specifically, 80% of the data was used for training and 20% for testing.

Data Preprocessing

Data preprocessing plays a crucial role in improving the reliability and predictive performance of machine learning models. In this study, a preprocessing pipeline was constructed to ensure consistent data transformation during model training and evaluation.

First, missing values were handled using median imputation, which replaces missing entries with the median value of the corresponding feature. Median imputation is widely used because it is robust to outliers and preserves the distribution of the data.

After handling missing values, feature scaling was applied using standardization. Specifically, the StandardScaler technique was used to transform numerical features to have zero mean and unit variance. This step is particularly important for algorithms that are sensitive to feature magnitude, such as Support Vector Machine (SVM) and k-Nearest Neighbor (KNN) [3].

All preprocessing steps were implemented using a pipeline structure provided by the scikit-learn framework. This approach ensures that identical transformations are applied consistently during both training and testing stages.

Machine Learning Models

In this study, six supervised machine learning models were implemented and comparatively evaluated for cancer risk prediction, including Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), k-Nearest Neighbor (KNN), and AdaBoost. These models were selected because they represent different learning principles, including linear classification, tree-based learning, distance-based learning, margin-based learning, and boosting-based ensemble learning. By combining these algorithms in a unified experimental setting, the study aimed to identify the model with the best predictive performance for clinical and lifestyle data. These algorithms have been widely used in previous studies on cancer prediction and medical decision support systems [1], [4], [6].

Hyperparameter Optimization and Validation

To improve model performance, hyperparameter tuning was performed using GridSearch combined with stratified k-fold cross-validation. This strategy systematically evaluates multiple parameter combinations and selects the configuration that yields the best performance. Stratified cross-validation was adopted to preserve the class distribution in each fold and reduce evaluation bias. In this study, the optimal hyperparameters were selected based on the highest ROC-AUC score obtained during cross-validation, and the final models were further evaluated on the independent test set. Hyper-parameter optimization using cross-validation has been shown to significantly improve the predictive performance and generalization ability of machine learning models in medical applications [2], [8].

Evaluation Metrics

The predictive performance of the models was evaluated using Accuracy, Precision, Recall, F1-score, and ROC-AUC. Accuracy is defined as

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

Precision is defined as

$$Precision = \frac{TP}{TP + FP}$$

Recall is defined as

$$Recall = \frac{TP}{TP + FN}$$

and F1-score is calculated as

$$F_1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

where *TP, TN*, *FP*, and *FN* denote true positives, true negatives, false positives, and false negatives, respectively.

Algorithm Workflow

The overall workflow of the proposed cancer risk prediction framework is illustrated in Figure 1. The workflow consists of several sequential stages including data acquisition, preprocessing, model training, hyperparameter optimization, model evaluation, and feature importance analysis.

Initially, the dataset containing demographic and lifestyle variables is collected and organized in a structured format. The dataset includes variables such as Age, Gender, BMI, Smoking status, Genetic Risk, Physical Activity, Alcohol Intake, and Cancer History, while the target variable represents the cancer diagnosis outcome.

In the preprocessing stage, missing values are handled using median imputation, and numerical features are standardized using feature scaling to ensure consistent feature distributions. The processed dataset is then divided into training and testing subsets using stratified sampling in order to preserve class distribution.

Next, multiple machine learning models are trained using the training dataset. Hyperparameter tuning is performed using grid search combined with stratified cross-validation to identify the optimal model configuration. After training, the selected models are evaluated on the independent test dataset using several performance metrics.

Finally, the best-performing model is selected, and feature importance analysis is conducted to identify the most influential predictors associated with cancer risk.

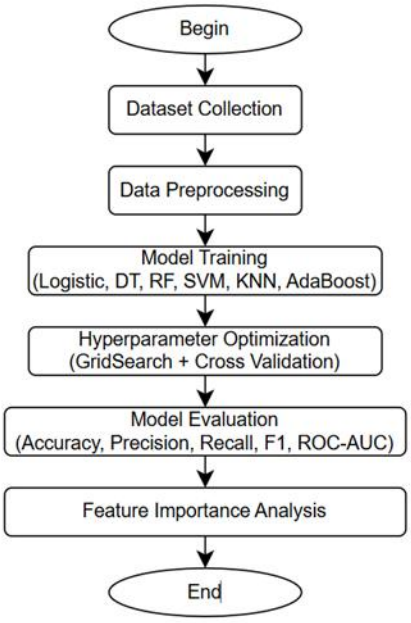

Figure 1. Workflow of the proposed machine learning framework for cancer risk prediction.

III. Results

Model Performance Comparison

Six machine learning algorithms were evaluated for cancer risk prediction, including Logistic Regression, Decision Tree, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Random Forest, and AdaBoost. Model performance was assessed using multiple evaluation metrics, including accuracy, precision, recall, F1-score, ROC-AUC, and cross-validation mean score.

Table 1 presents the performance comparison of all evaluated models.

Table 1: Performance comparison of machine learning models.

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC	CV Mean
AdaBoost	**0.96**	**0.99**	0.90	**0.94**	**0.97**	**0.96**
Random Forest	0.94	0.97	0.86	0.91	0.96	0.94
KNN	0.89	0.92	0.76	0.83	0.95	0.92
SVM	0.89	0.89	0.81	0.85	0.94	0.92
Decision Tree	0.90	0.88	0.83	0.86	0.94	0.89
Logistic Regression	0.84	0.80	0.77	0.79	0.92	0.92

Among the evaluated models, AdaBoost achieved the best overall performance, with an accuracy of 0.96, precision of 0.99, recall of 0.90, and an F1-score of 0.94. The model also obtained the highest ROC-AUC value (0.97) and the highest cross-validation score (0.96), indicating strong predictive capability and good generalization.

Random Forest achieved the second-best performance, with an accuracy of 0.94 and ROC-AUC of 0.962, demonstrating the effectiveness of ensemble-based learning approaches. In contrast, traditional classifiers such as Logistic Regression produced lower performance, with an accuracy of 0.84 and ROC-AUC of **0.92**. Overall, ensemble methods consistently outperformed single learners across most evaluation metrics.

ROC Curve Analysis

To further evaluate the discriminative ability of the models, Receiver Operating Characteristic (ROC) curves were generated for all classifiers. The figure 2 illustrates the trade-off between the true positive rate (TPR) and the false positive rate (FPR) across different classification thresholds.

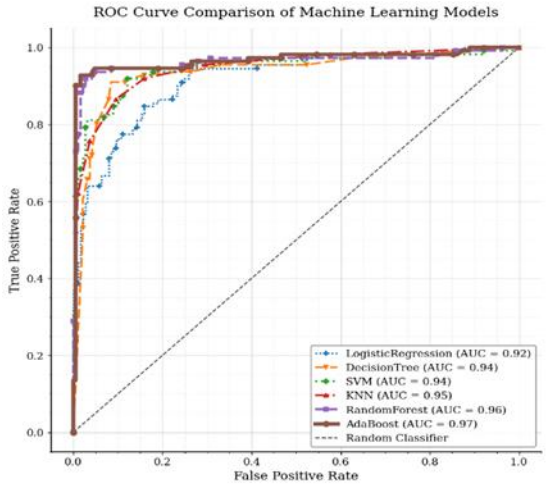

Figure 2. ROC curve comparison among the evaluated models.

The AdaBoost classifier achieved the highest area under the curve (AUC = 0.97), indicating the strongest ability to distinguish between high-risk and low-risk individuals. Random Forest also demonstrated strong discriminative capability with an AUC of 0.96, followed by KNN (0.95), SVM (0.94), Decision Tree (0.94), and Logistic Regression (0.92).

Figure 2 shows that ensemble-based models generally maintain higher true positive rates while keeping false positive rates relatively low, suggesting superior classification performance.

Feature Importance Analysis

To better understand the factors influencing cancer risk prediction, feature importance analysis was conducted using the Random Forest model. The importance score reflects the relative contribution of each feature to the prediction process.

Table 2 illustrates the ranking of the most important features.

Table 2: Feature importance ranking.

Feature	Importance
CancerHistory	**0.181**
AlcoholIntake	0.146
BMI	0.146
GeneticRisk	0.142
Age	0.126
PhysicalActivity	0.114
Gender	0.087
Smoking	0.057

Among all variables, CancerHistory emerged as the most influential feature, with an importance score of 0.18, indicating that historical cancer-related information plays a crucial role in predicting cancer risk.

Lifestyle-related factors such as AlcoholIntake, BMI, and PhysicalActivity also demonstrated notable importance in the model. Additionally, GeneticRisk and Age contributed significantly to the predictive performance, reflecting the combined influence of biological predisposition and demographic characteristics.

Other variables, including Gender and Smoking, showed comparatively lower importance scores within the current dataset.

IV. Discussion

Interpretation of model performance

The experimental results demonstrate that ensemble learning approaches provide superior performance for cancer risk prediction compared with traditional machine learning classifiers. Among the evaluated models, AdaBoost achieved the best predictive performance, with an accuracy of 0.96, an F1-score of 0.94, and an ROC-AUC of 0.97. These results indicate that the model has strong discriminative ability in distinguishing between high-risk and low-risk individuals.

The superior performance of AdaBoost can be attributed to its boosting mechanism, which sequentially combines multiple weak learners to reduce classification errors. By focusing on misclassified samples during the training process, AdaBoost effectively improves the overall predictive capability of the model. This mechanism

allows the model to capture complex and nonlinear relationships among risk factors that may not be adequately represented by simpler classifiers.

Random Forest also demonstrated strong performance, achieving the second-highest accuracy (0.94) and ROC-AUC (0.96). As another ensemble-based approach, Random Forest benefits from the aggregation of multiple decision trees, which improves model robustness and reduces variance. The strong performance of both AdaBoost and Random Forest suggests that ensemble learning techniques are particularly suitable for medical prediction tasks involving heterogeneous risk factors.

Interpretation of important risk factors

Feature importance analysis revealed several key variables that significantly contribute to cancer risk prediction. Among them, CancerHistory was identified as the most influential feature. This finding is consistent with established medical knowledge, as personal or family history of cancer is widely recognized as a strong predictor of future cancer risk.

Lifestyle-related variables, including AlcoholIntake and BMI, were also ranked among the most important predictors. Excessive alcohol consumption has been linked to an increased risk of several cancer types, including liver, colorectal, and breast cancers. Similarly, elevated BMI is associated with metabolic and hormonal changes that may promote tumor development.

The variable GeneticRisk also demonstrated substantial importance, highlighting the role of inherited genetic predispositions in cancer susceptibility. In addition, Age emerged as an important predictor, which aligns with epidemiological studies showing that cancer incidence increases with age due to the accumulation of genetic mutations and long-term exposure to environmental risk factors.

Interestingly, variables such as Smoking and Gender exhibited relatively lower importance scores in the current dataset. While smoking is widely known to be a major risk factor for certain cancers, its lower importance in this model may be due to correlations with other lifestyle variables or limitations of the available dataset.

Practical implications

The findings of this study suggest that machine learning techniques can serve as effective tools for cancer risk assessment using readily available demographic and lifestyle information. The high predictive performance achieved by the AdaBoost model indicates that such approaches may be useful in supporting early risk screening and personalized preventive strategies.

In clinical practice, predictive models of this type could assist healthcare professionals in identifying individuals with elevated cancer risk, thereby enabling earlier monitoring, targeted screening programs, and lifestyle interventions. Additionally, the identification of key risk factors through feature importance analysis provides interpretable insights that may help guide preventive healthcare strategies.

V. Conclusion

This study demonstrates the feasibility of applying machine learning techniques to support cancer risk assessment based on readily available demographic and lifestyle information. The comparative evaluation of multiple classification algorithms highlights the effectiveness of ensemble learning approaches in handling heterogeneous health-related data and capturing complex relationships among risk factors. Beyond predictive performance, the analysis also provides insights into the relative contribution of different variables involved in cancer risk, which may help improve the interpretability of data-driven medical decision support systems. The proposed framework shows potential as a complementary tool for preventive healthcare and early risk screening. Future research may focus on integrating additional clinical or biological information and validating the framework across diverse populations to further enhance its reliability and practical applicability.

References

[1]. N. Fatima, L. Liu, H. Sha, And H. Ahmed, “Prediction Of Breast Cancer: Comparative Review Of Machine Learning Techniques And Their Analysis,” Ieee Access, Vol. 8, Pp. 150360–150376, 2020.

[2]. J. A. M. Sidey-Gibbons And C. J. Sidey-Gibbons, “Machine Learning In Medicine: A Practical Introduction,” Bmc Medical Research Methodology, Vol. 19, No. 64, 2019.

[3]. M. Kumari And V. Singh, “Breast Cancer Prediction System,” Procedia Computer Science, Vol. 132, Pp. 371–376, 2018.

[4]. M. A. Naji, S. El Filali, K. Aarika, E. H. Benlahmar, R. Ait Abdelouhahide, And O. Debauche, “Machine Learning Algorithms For Breast Cancer Prediction And Diagnosis,” Procedia Computer Science, Vol. 191, Pp. 487–492, 2021.

[5]. Austria, Y. D., Goh, M. L., Sta. Maria, L. B., Lalata, J. P., Goh, J. E., & Vicente, H. N. (2019). Comparison Of Machine Learning Algorithms In Breast Cancer Prediction Using The Coimbra Dataset. International Journal Of Simulation: Systems, Science And Technology. Https://Doi.Org/10.5013/Ijssst.A.20.S2.23.

[6]. A. Gupta, M. Garg, D. Kaushik, And A. Verma, “Machine Learning Model For Breast Cancer Prediction,” In Proc. Ieee Int. Conf. On I-Smac (Iot In Social, Mobile, Analytics And Cloud), 2020, Pp. 472–476.

[7]. T. A. Assegie, R. L. Tulasi, And N. Komal Kumar, “Breast Cancer Prediction Model With Decision Tree And Adaptive Boosting,” Iaes International Journal Of Artificial Intelligence, Vol. 10, No. 1, Pp. 184–190, 2021.

[8]. D. T. Hang, N. V. Bao, N. D. Hung, And D. V. Sang, "Enhancing Heart Disease Risk Prediction Using Stacking Ensemble Learning," In Proc. Fifth International Conference On Intelligent Systems And Networks (Icisn 2025), Lecture Notes In Networks And Systems, Vol. 1596, Springer, Singapore, 2026.
[9]. Yarabarla, M.S., Ravi, L., & Sivasangari, A. (2019). Breast Cancer Prediction Via Machine Learning. 2019 3rd International Conference On Trends In Electronics And Informatics (Icoei), 121-124.
[10]. Cancer Prediction Dataset. Kaggle. Available: Https://Www.Kaggle.Com/. Accessed: 2026.