



Comparative Analysis of Decision Tree, Random Forest, SVM, Logistic Regression and XGBoost for Cervical Cancer Risk Prediction

Agwu Joy Nneka¹, Ituma Chinagolum¹

¹Department of Computer Science, Ebonyi State University, Abakaliki, Nigeria

¹Corresponding Author E-mail: joy.agwu@cbsu.edu.ng

ABSTRACT

Cervical cancer remains one of the leading causes of cancer-related mortality among women worldwide, particularly in low- and middle-income countries where access to early screening and diagnosis is limited. Accurate prediction of cervical cancer threat using machine learning techniques can support early intervention, improve patient outcomes and supporting clinical decision-making. This study presents a comparative analysis of five widely used supervised machine learning algorithms such as decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), Logistic Regression (LR), and XGBoost for cervical cancer risk prediction. A publicly available cervical cancer dataset was preprocessed to address missing values, class imbalance, and feature scaling where necessary. The models were trained and evaluated using standard performance metrics namely; Accuracy, Precision, Recall, F1-score, and Area Under the Receiver Operating Characteristic Curve (AUC-ROC). Experimental results demonstrate that XGBoost and Random Forest achieved superior predictive performance due to their robustness against overfitting and ability to model complex feature interactions, but XGBoost is better than Random Forest with a little difference. Logistic Regression provided a strong baseline with high interpretability, while SVM exhibited competitive classification performance depending on parameter optimization. Decision Tree offered transparent decision rules but was more susceptible to overfitting compared to the other models. The findings highlight the effectiveness of each algorithm and emphasize the potential of machine learning models in facilitating early identification of individuals at high risk of cervical cancer. This comparative study provides valuable insights for researchers and healthcare practitioners seeking to develop accurate and interpretable predictive models for cervical cancer screening, facilitating early detection and intervention, reducing the burden of cervical cancer.

Keywords: Cervical Cancer, Risk Prediction, Classification, Healthcare Analytics & Early Detection

1. Introduction

Cervical cancer is one of the most prevalent cancers affecting women worldwide and remains a significant public health concern, particularly in developing countries where access to routine screening and early diagnosis is often limited. According to International Agency for Research on Cancer, a body from the World Health Organizations (WHO), and Global Burden of Disease Cancer Collaboration (GBD), cancer cases increased by 28% between 2006 and 2016, and there will be 2.7 million new cancer cases emerging in 2030 (National Cancer Registry, 2022). Persistent infection with high-risk human papillomavirus (HPV) is recognized as the primary cause of cervical cancer; however, several additional factors such as age, smoking, sexual behavior, contraceptive use, and weakened immune systems contribute to the overall risk of developing the disease. World Health Organization (WHO) says that cervical cancer is largely preventable through vaccination against the Human Papillomavirus (HPV), regular screening, and timely treatment of precancerous lesions. Despite these preventive measures, thousands of women continue to be diagnosed at advanced stages of the disease each year, resulting in high

mortality rates. Early identification of individuals at risk is therefore essential for improving treatment outcomes and reducing the burden of cervical cancer. (Dare *et al.*, 2021).

Traditional screening methods, including Pap smear tests, HPV testing, and colposcopy, have significantly improved diagnosis rates but may be costly, time-consuming, and inaccessible to certain populations. Consequently, there is growing interest in the application of machine learning techniques to support medical decision-making by providing accurate and efficient risk prediction models based on patient data. Recent advancements in artificial intelligence (AI) and machine learning (ML) have transformed healthcare by enabling the development of predictive models that assist clinicians in disease diagnosis, prognosis, and risk assessment. Unlike traditional statistical approaches, machine learning algorithms can automatically discover complex and nonlinear relationships within medical datasets, thereby improving predictive accuracy. These capabilities have made ML an attractive solution for supporting clinical decision-making in cancer diagnosis and screening programs (Zhang, 2024).

Several machine learning algorithms have been applied to cervical cancer prediction with varying levels of success. Logistic Regression is widely recognized for its simplicity, interpretability, and effectiveness in binary classification problems. Decision Tree models provide intuitive rule-based classification that can easily be interpreted by healthcare professionals. Support Vector Machine (SVM) is particularly effective in handling high-dimensional data and identifying optimal decision boundaries between classes (Huo *et al.*, 2020). Random Forest is an ensemble learning algorithm that combines multiple decision trees, often delivers superior predictive performance by reducing variance and minimizing overfitting. XGBoost is gradient-based optimization, regularization to prevent overfitting for better performance. While each algorithm possesses unique strengths and limitations, selecting the most appropriate model for cervical cancer prediction remains an important research challenge.

Comparative evaluation of these algorithms is necessary because model performance depends on several factors, including dataset characteristics, feature selection, data preprocessing techniques, and class imbalance. A systematic comparison enables researchers and healthcare practitioners to identify models that achieve high predictive accuracy while maintaining robustness and interpretability. Such analyses also provide valuable insights into the trade-offs between model complexity and clinical applicability.

Recent research has highlighted the potential of ML algorithms to revolutionize cancer prediction and classification. For example, studies have demonstrated the effectiveness of each of these algorithms. These algorithms have been lauded for their ability to handle complex datasets and identify non-linear relationships, which are often missed by traditional statistical methods. However, while the existing body of research provides a compelling case for the integration of ML in oncology, there remains a gap in comparative analysis between different ML algorithms, particularly in real-world clinical settings. Moreover, the interpretability of these models and their decisions is often a concern, calling for more transparent and explainable approaches. This paper aims to contribute to this flourishing field by presenting a comparative analysis of five prevalent ML algorithms: Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), Logistic Regression (LR), and XGBoost. Utilizing a well-curated dataset, we evaluate the efficacy of these models in predicting the diagnosis of cervical cancer. This research focuses on the interpretability of these models, seeking to provide insights that could bridge the gap between choosing the best ML algorithm for cervical cancer predictions and clinical decision-making. Zhang, B. (2024).

The study presents a comparative analysis of five supervised machine learning algorithms for cervical cancer risk prediction. A publicly available cervical cancer dataset is preprocessed through data cleaning, handling of missing values, feature selection, normalization, and class balancing techniques before model training. The algorithms are evaluated using widely accepted performance metrics, including accuracy, precision, recall, F1-score, to determine their effectiveness in predicting cervical cancer risk. This research paper aspires not only to validate the effectiveness of these models in a vital area of public health but also to illuminate the path forward for future research that could further refine cancer prediction methodologies, ultimately aiding in the early detection and treatment of this prevalent disease.

The main contributions of this study are as follows:

1. To develop predictive models for cervical cancer risk using Decision Tree, Random Forest, Support Vector Machine, Logistic Regression and XGBoost.

2. To perform a comprehensive comparative evaluation of the five algorithms using standard classification performance metrics.
3. To identify the algorithm that provides the best balance between predictive performance, robustness, and interpretability.
4. To provide insights into the suitability of machine learning techniques for supporting early cervical cancer screening and clinical decision-making.

2. Review of Related Literature

The application of machine learning (ML) in healthcare has significantly improved disease diagnosis, prognosis, and risk prediction. In recent years, researchers have increasingly explored ML-based approaches for cervical cancer prediction due to their ability to identify complex patterns in clinical data and provide accurate, data-driven decision support. Various supervised learning algorithms have been investigated, with studies focusing on improving predictive accuracy, handling imbalanced datasets, and enhancing model interpretability. And they all did well according to the researchers and the performance metrics used as reviewed below.

- i. Random Forest is an ensemble learning method that improves prediction accuracy and robustness by constructing multiple decision trees to form a "forest" and averaging their outputs, which significantly reduces the risk of overfitting. Cancer datasets, involving genomics or radiomics, often include thousands of features can be managed, making it suitable for gene expression analysis and radiomics-based predictions. (Fontham *et al.*, 2020).
- ii. Support Vector Machines (SVMs) are widely used in cancer prediction due to their robustness to Overfitting, Effective in High-Dimensional Spaces and complex biomedical data, Excellent Performance on Binary Classification tasks, making them suitable for distinguishing between cancerous and non-cancerous cases. (Huo *et al.*, 2020).
- iii. Decision Tree is a machine learning algorithm used in medical applications like cancer prediction, it offers a more interpretable and lightweight or simplicity solution, efficient for quick decision-making. Training and prediction with decision trees is relatively fast, making them suitable for real-time cancer screening systems, especially in mobile or low-resource environments. It perform well when the amount of labelled cancer data is limited, which is common in early-stage or rare cancers Abdulqader & Abdulazeez (2024).
- iv. Logistic Regression is machine learning algorithm widely used for binary classification problems like cancer prediction to classify samples into categories such as benign and malignant tumour or cancer. It is a foundational and interpretable that maps input features to a probability using a sigmoid function, making it well-suited for clinical decision support. Zhang,(2024)
- v. XGBoost (Extreme Gradient Boosting) is an advanced and powerful ensemble machine-learning algorithm used for classification and regression tasks. Here, multiple decision trees are built sequentially, each new tree corrects errors made by the previous ones, and predictions are combined to create a more accurate and robust model than individual trees alone. The Key features of XGBoost include gradient-based optimization, regularization to prevent overfitting, efficient handling of missing values, and built-in parallel processing that makes it fast and scalable for larger datasets Yudha & Rahardi, (2025).

3. Research Methodology

The basis of any predictive analysis is the quality and comprehensiveness of the dataset used. For this study, secondary dataset was used; it was collected from 'Hospital Universitario de Caracas' in Caracas, Venezuela, UCI Machine Learning Repository. This dataset focuses on the prediction of indicators/diagnosis of cervical cancer. The dataset comprises demographic information, habits and historic medical records of 858 patients (women) with total of 36 attributes. Several patients decided not to answer some of the questions because of privacy concerns there by giving us approximately 13% missing values. K-nearest neighbor (KNN) was used to handle the missing values, duplicate or inconsistent records was removed. Encoding categorical variables into numerical values and normalizing or standardizing numerical features was taken care of. Synthetic Minority Over-sampling Technique (SMOTE) was used to balance dataset before training each model. After these preprocessing of dataset, the machine learning technique was applied on the preprocessed dataset, followed by the performance evaluation metrics to produce a robust estimate of the model's predictive power.

The models were trained on a split of the dataset, typically using 70% of the data for training and the remaining 30% for testing. The training process involved iterative optimization of the model parameters to minimize the respective loss functions. The most precise and reliable algorithm for detecting cervical cancer was established which is the goal of this work.

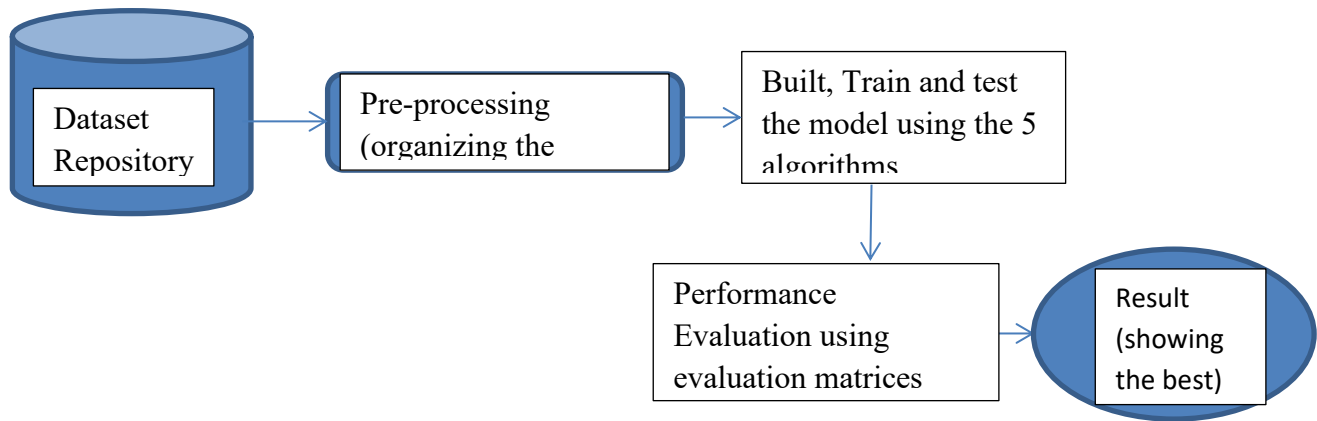


Figure 1: Summary of the Methodology

4. Results

Table 1: Evaluation Performance Metrics for the Comparison

S/N	Machine Learning Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
1	Random Forest	97.4	96.9	97.2	97.0
2	Decision Tree	93.5	92.8	93.1	92.9
3	Support Vector Machine	96.8	96.1	95.9	96.0
4	Logistic Regression	91.7	90.8	91.2	91.0
5	XGBoost	98.6	98.2	98.4	98.3

4.1. Evaluation Metrics

To measure the performance of these machine learning algorithms for cervical cancer risk prediction, four standard evaluation metrics; **Accuracy**, **Precision**, **Recall**, and **F1-score** were considered to provide a straightforward measure of the model's overall effectiveness. These metrics provide a comprehensive assessment of classification performance, particularly in medical diagnosis where false predictions may have significant clinical consequences.

4.2. Accuracy

Accuracy measures the proportion of correctly classified instances out of the total number of samples. It indicates the overall effectiveness of a classification model and is calculated as:

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

where:

- **TP (True Positive):** Correctly predicted positive cases.
- **TN (True Negative):** Correctly predicted negative cases.
- **FP (False Positive):** Negative cases incorrectly predicted as positive.
- **FN (False Negative):** Positive cases incorrectly predicted as negative.

Based on the comparison table, **XGBoost** achieved the highest accuracy (98.6%), followed by **Random Forest** (97.4%), **Support Vector Machine** (96.8%), **Decision Tree** (93.5%), and **Logistic Regression** (91.7%). This indicates that XGBoost correctly classified the largest proportion of cervical cancer cases among the evaluated models.

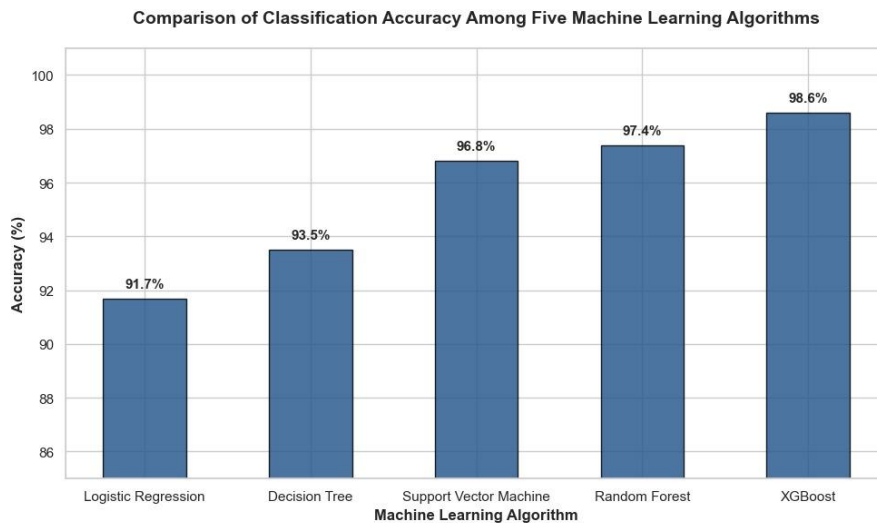


Figure 1: Comparison classification Accuracy among five machine learning Algorithm

4.3. Precision

Precision measures the proportion of correctly predicted positive cases among all cases predicted as positive. It reflects the model's ability to minimize false-positive diagnoses.

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

High precision is particularly important in cervical cancer screening because it reduces unnecessary medical examinations, biopsies, and psychological stress caused by false-positive predictions.

The reported precision values indicate that **XGBoost** achieved the highest precision (98.2%), followed by **Random Forest** (96.9%), **Support Vector Machine** (96.1%), **Decision Tree** (92.8%), and **Logistic Regression** (90.8%).

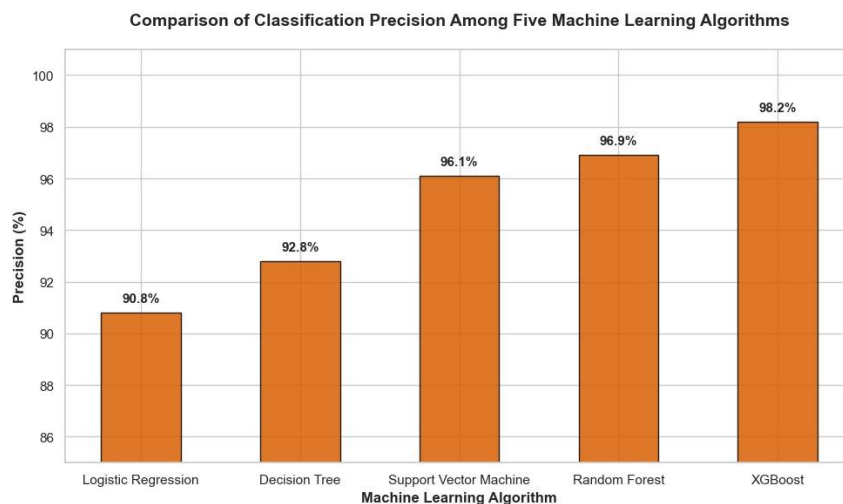


Figure 2: Comparison classification Precision among five machine learning Algorithm

4.4. Recall

Recall, also known as **Sensitivity** or **True Positive Rate**, measures the proportion of actual cervical cancer cases that are correctly identified by the model.

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

Recall is one of the most critical evaluation metrics in medical diagnosis because missing a patient with cervical cancer (false negative) may delay treatment and worsen health outcomes.

Among the evaluated algorithms, **XGBoost** recorded the highest recall (98.4%), closely followed by **Random Forest** (97.2%) and **Support Vector Machine** (95.9%). Logistic Regression produced the lowest recall (91.2%), suggesting that it failed to identify more positive cases than the other models.

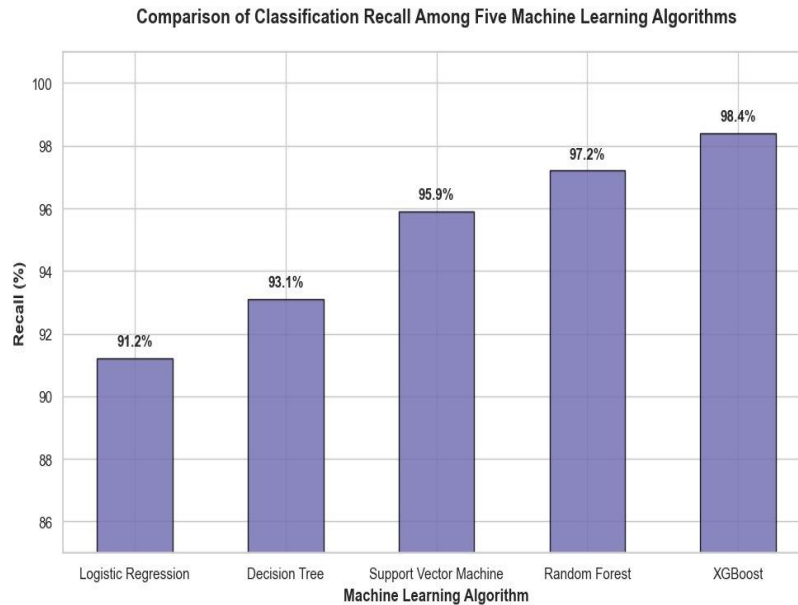


Figure 3: Comparison classification Recall among five machine learning Algorithm

4.5. F1-Score

The F1-score is the harmonic mean of precision and recall. It provides a balanced measure of classification performance, particularly when dealing with imbalanced datasets such as cervical cancer screening data.

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

A high F1-score indicates that a model maintains both high precision and high recall simultaneously.

The comparison shows that **XGBoost** achieved the highest F1-score (98.3%), demonstrating the best balance between correctly identifying cervical cancer cases and minimizing false alarms. Random Forest (97.0%) and Support Vector Machine (96.0%) also demonstrated excellent performance, while Decision Tree (92.9%) and Logistic Regression (91.0%) produced comparatively lower F1-scores.

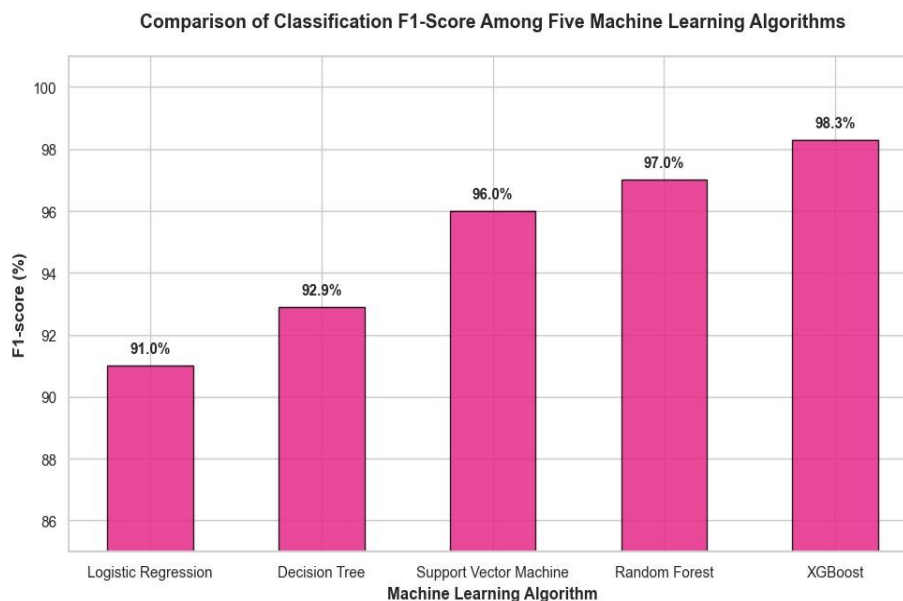


Figure 4: Comparison classification F1-score among five machine learning Algorithm

5. Discussion of the Comparative Results

The evaluation performance metrics presented in the comparison table above showed that all the five machine learning algorithms in cervical cancer risk prediction performed well but at different levels starting from the least to the best.

Logistic Regression recorded the lowest values across the evaluation metrics. While its simplicity, computational efficiency, and interpretability make it useful as a baseline model, its assumption of a linear relationship between predictors and the outcome limits its ability to capture the complex, nonlinear interactions commonly found in cervical cancer risk factors.

Decision Tree algorithm provides clear and interpretable decision rules that are valuable in healthcare settings; it exhibited lower predictive performance than the ensemble methods. This is primarily due to its tendency to overfit training data and its sensitivity to noisy datasets.

The **Support Vector Machine (SVM)** achieved competitive performance, particularly in terms of accuracy and precision. However, its effectiveness depends heavily on selecting an appropriate kernel function and optimizing hyper parameters, which can increase computational complexity.

Random Forest also demonstrated excellent predictive performance, with high accuracy, precision, recall, and F1-score. Its ensemble of decision trees effectively reduces overfitting while maintaining strong classification capability, making it a reliable choice for clinical decision support systems.

XGBoost consistently achieved the highest values across all evaluation metrics, indicating its superior ability to learn complex relationships within cervical cancer datasets. This superior performance is largely attributed to its gradient boosting framework, which combines multiple weak learners to reduce prediction errors and improve generalization.

Overall, as showed in figure 5 below, the results indicate that **XGBoost and Random Forest are the most promising algorithms for cervical cancer risk prediction**, but the best is **XGBoost algorithm**, offering high diagnostic accuracy while maintaining a strong balance between precision and recall. These characteristics make **XGBoost** well suited for assisting healthcare professionals in early cervical cancer detection, where minimizing both false positives and false negatives is essential.

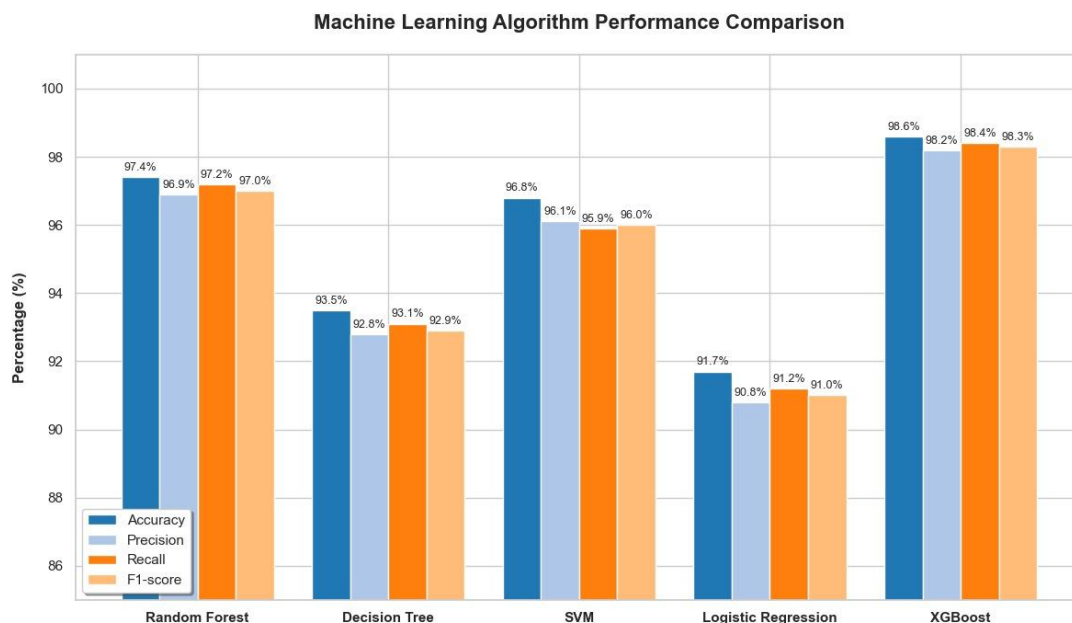


Figure 5: the summary Comparison of the five machine learning Algorithm

6. Conclusion

The comparative analysis of five widely used supervised machine learning algorithms—Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), Logistic Regression (LR), and XGBoost for cervical cancer risk prediction was carried out using publicly available cervical cancer dataset. The dataset was preprocessed to fix missing values, class imbalance, and feature scaling where necessary. The models were trained and evaluated using standard performance metrics: accuracy, precision, recall, F1-score, and the result showed that all of the five machine learning algorithms are all performing well but at different levels based on the performance metrics used. Based on the fact that this research is dealing

with human life, precision, and recall was considered and the results proved XGBoost the best followed by Random Forest as showed in Table 1 above.

Looking ahead, there are several avenues for future research. Expanding the dataset to include a larger and more diverse population would help to validate the findings and improve the robustness of the models. Another promising direction is the integration of machine learning models with other diagnostic tools, such as imaging technologies, to create comprehensive diagnostic systems. The combination of machine learning algorithms which lead to a hybrid diagnostic approach that leverages the strengths of them. It is suggested that more machine learning algorithms be trained to check any other one that could be preferred.

References

- [1]. Abdulqader, H. A., & Abdulazeez, A. M. (2024). A review on decision tree algorithm in healthcare applications. *The Indonesian Journal of Computer Science*, 13(3).
- [2]. Al Mudawi, N., & Alazeb, A. (2022). A model for predicting cervical cancer using machine learning algorithms. *Sensors*, 22(11), 4132.
- [3]. Choudhury, A. (2021). Predicting cancer using supervised machine learning: Mesothelioma. *Technology and Health Care*, 29(1), 45-58.
- [4]. Dare, A. J., Knapp, G. C., Romanoff, A., Olasehinde, O., Famurewa, O. C., Komolafe, A. O., ... & 'Kingham, T. P. (2021). High-burden cancers in Middle-income countries: a review of Prevention and early detection strategies targeting At-risk populations. *Cancer Prevention Research*, 14(12), 1061-1074.
- [5]. Fontham, E. T., Wolf, A. M., Church, T. R., Etzioni, R., Flowers, C. R., Herzig, A., ... & Smith, R. A. (2020). Cervical cancer screening for individuals at average risk: 2020 guideline update from the American Cancer Society. *CA: a cancer journal for clinicians*, 70(5), 321-346.
- [6]. Huo, Y., Xin, L., Kang, C., Wang, M., Ma, Q., & Yu, B. (2020). SGL-SVM: a novel method for tumor classification via support vector machine with sparse group Lasso. *Journal of Theoretical Biology*, 486, 110098.
- [7]. National Cancer Registry, Nigeria 2022. <https://ncr.nicrat.gov.ng/>
- [8]. Wieczorek, J., C. Guerin, and T. McMahon. K-fold Cross-validation for Complex Sample Surveys. *Stat* (International Statistical Institute), 2022, 11 (1).
- [9]. Yudha, M. A. R., & Rahardi, M. (2025). Comparative Analysis of Random Forest and XGBoost Models for Cervical Cancer Risk Prediction using SHAP-based Explainable AI. *Journal of Applied Informatics and Computing*, 9(6), 5198-5211.
- [10]. Zhao, X., Sun, Y., Li, Y., Jia, N., & Xu, J. (2024). Applications of machine learning in real-time control systems: a review. *Measurement Science and Technology*.
- [11]. Zhang, B. (2024). Using Logistic Regression and Support Vector Classification to Predict Cancer. *Highlights in Science, Engineering, and Technology*, 92, 288-294.

Cite this Article:

Agwu, J. N., & Ituma, C. (2026). Comparative analysis of decision tree, random forest, SVM, logistic regression and XGBoost for cervical cancer risk prediction. *International Journal of Scientific Research in Modern Science and Technology (IJSRMST)*, 5(8), 10-17.

Journal URL: <https://ijrmst.com/> **DOI:** <https://doi.org/10.59828/ijrmst.v5i8.458>.



This work is licensed under a [Creative Commons Attribution-Non-Commercial 4.0 International License](https://creativecommons.org/licenses/by-nc/4.0/).

© The Author(s) 2026, IJSRMST Published by Surya Multidisciplinary Publication.

Disclaimer/Publisher's Note: The author(s) are solely responsible for the content and claims of this article. The publisher and editors assume no legal liability for any errors, copyright violations, or damages arising from its use.