INTERNATIONAL JOURNAL OF RESEARCH AND INNOVATION IN APPLIED SCIENCE (IJRIAS)
ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue VI June 2026

An Explainable SMOTE-Enhanced Fuzzy Ensemble Decision Tree Framework for Intelligent Cancer Prediction and Classification with Hyperparameter Turning

Sylvester I. Ele[1*] and Oko Sunday Adi[2]

[1]Department of Software Engineering, University of Calabar, Nigeria

[2]Department of Computer Science, University of Calabar, Nigeria

DOI: https://doi.org/10.51584/IJRIAS.2026.11060252

Received: 27 June 2026; Accepted: 02 July 2026; Published: 15 July 2026

ABSTRACT

Early detection of cervical cancer remains a critical challenge in healthcare, particularly in low- and middle-income countries where limited access to screening and diagnostic facilities contributes to high mortality rates. Although machine learning techniques have shown considerable promise in supporting cancer diagnosis, their effectiveness is often hindered by class imbalance, data uncertainty, and limited model interpretability. This study proposes an explainable SMOTE-enhanced fuzzy ensemble decision tree framework with hyperparameter optimization for intelligent cervical cancer prediction and classification. The framework integrates Synthetic Minority Oversampling Technique (SMOTE), fuzzy logic, multiple fuzzy decision trees (Fuzzy ID3, Fuzzy C4.5, and Fuzzy CART), and ensemble learning models, including Random Forest, Gradient Boosting, XGBoost, LightGBM, and AdaBoost, to improve predictive accuracy, robustness, and clinical interpretability. The Cervical Cancer (Risk Factors) dataset obtained from the UCI Machine Learning Repository, comprising 858 patient records and 36 clinical and demographic attributes, was employed for experimentation. Data preprocessing involved missing-value imputation, normalization, fuzzification of selected clinical variables using triangular membership functions, and class balancing through SMOTE. Hyperparameter tuning was performed using Bayesian optimization, grid search, and randomized search techniques. Model performance was evaluated using accuracy, precision, recall, F1-score, ROC-AUC, confusion matrices, and ROC curves. Experimental results demonstrated that ensemble methods consistently outperformed conventional decision tree algorithms. The proposed framework achieved its best performance with Gradient Boosting, attaining an accuracy of 97.09%, precision of 87.50%, recall of 63.64%, F1-score of 73.68%, and a ROC-AUC of 94.35%, while XGBoost and Random Forest achieved the highest discriminative capabilities with ROC-AUC values of 97.35% and 97.21%, respectively. The incorporation of SMOTE significantly improved minority-class detection, whereas fuzzy logic enhanced the handling of uncertainty inherent in medical data. The findings confirm that the proposed explainable framework provides an effective, robust, and interpretable decision-support mechanism for early cervical cancer diagnosis and has strong potential for broader applications in intelligent healthcare systems.

Keywords: Cervical cancer prediction, explainable artificial intelligence, SMOTE, fuzzy decision trees, hyperparameter optimization.

INTRODUCTION

Cervical cancer remains one of the leading causes of cancer-related deaths among women worldwide, particularly in low- and middle-income countries where access to early diagnosis and effective screening remains limited. According to the World Health Organization (WHO, 2023), cervical cancer accounts for a significant proportion of female cancer mortality, with delayed diagnosis contributing heavily to poor survival outcomes. Early detection and accurate prediction of cervical cancer risk are therefore essential for improving patient survival and reducing healthcare burdens. The increasing availability of medical datasets and advances in artificial intelligence have encouraged the adoption of machine learning techniques for disease prediction and clinical decision support systems. Machine learning models can analyze complex medical data and identify hidden patterns associated with cancer development and progression (Kourou et al., 2015). Traditional classification algorithms such as Decision Trees, Random Forest, and Gradient Boosting have shown promising results in healthcare prediction tasks due to their ability to model nonlinear relationships and handle multidimensional data. However, medical datasets frequently contain uncertainties, missing values, imbalanced class distributions, and noisy information that reduce predictive performance. Cervical cancer datasets are particularly affected by class imbalance because positive cancer cases are significantly fewer than negative cases. This imbalance often causes machine learning models to become biased toward the majority class, thereby reducing sensitivity and increasing false negative predictions (Chawla et al., 2002). To address this problem, the Synthetic Minority Oversampling Technique (SMOTE) has emerged as one of the most effective data balancing techniques in machine learning. SMOTE generates synthetic minority class samples to improve class representation and enhance classifier learning performance (Chawla et al., 2002). Several studies have shown that integrating SMOTE with ensemble learning methods significantly improves classification accuracy and minority class detection in healthcare applications (Fernández et al., 2018).

In addition to class imbalance, medical data often involve vagueness and uncertainty that cannot be effectively modeled using traditional crisp logic systems. Fuzzy logic provides a robust mechanism for handling imprecise medical information by representing variables using linguistic terms such as low, medium, and high (Zadeh, 1965). Fuzzy Decision Trees (FDTs) combine fuzzy logic with decision tree learning to improve interpretability and classification performance in uncertain environments. Recent advances in ensemble learning have further improved predictive modeling by combining multiple weak learners to generate more accurate and stable predictions. Algorithms such as Random Forest, Gradient Boosting Machine (GBM), XGBoost, and LightGBM have demonstrated superior performance in medical diagnosis and cancer prediction tasks (Chen & Guestrin, 2016). Despite these advances, limited studies have integrated SMOTE, fuzzy logic, ensemble learning, and optimized hyperparameter tuning into a unified framework for cervical cancer prediction. Existing studies often focus on individual models without adequately addressing uncertainty and class imbalance simultaneously.

This study therefore proposes a SMOTE-Enhanced Fuzzy Ensemble Learning Framework for Cervical Cancer Prediction using optimized decision tree and ensemble learning models. The framework mixes fuzzy decision trees, SMOTE balancing, ensemble learning techniques, and optimized hyperparameter tuning to improve prediction accuracy, sensitivity, and robustness in cervical cancer diagnosis. The study aims to develop a fuzzy decision tree framework for cervical cancer prediction by applying SMOTE to address class imbalance, integrating ensemble learning algorithms such as Random Forest, Gradient Boosting, XGBoost, and LightGBM, optimizing model hyperparameters, and evaluating performance using standard classification metrics.

Machine learning has become increasingly important in medical diagnosis because of its ability to process large-scale healthcare datasets and identify predictive patterns associated with disease occurrence and progression. Kourou et al. (2015) observed that machine learning algorithms significantly improve cancer prognosis and survival prediction by extracting meaningful information from clinical datasets. Several machine learning techniques have been applied to cervical cancer prediction, particularly decision tree-based models such as ID3, C4.5, and CART, due to their simplicity, interpretability, and rule-based structure (Quinlan, 1993). However, traditional decision tree models often experience overfitting and poor generalization when dealing with noisy or imbalanced medical datasets. To overcome these limitations, ensemble learning techniques have been introduced to improve predictive performance by combining multiple classifiers. Random Forest employs bagging and random feature selection to produce robust predictions (Breiman, 2001), while Gradient Boosting sequentially corrects previous classification errors to enhance accuracy (Friedman, 2001). Chen and Guestrin (2016) further developed XGBoost as an optimized boosting algorithm that incorporates regularization, parallel processing, and second-order optimization to improve efficiency and prediction accuracy. Similarly, LightGBM enhances computational performance through histogram-based learning and leaf-wise tree growth (Ke et al., 2017). These ensemble methods have achieved state-of-the-art performance in several medical prediction applications.

Despite these advances, class imbalance remains a major challenge in medical data classification, particularly in cervical cancer datasets where positive cancer cases are often underrepresented. This imbalance causes predictive models to become biased toward the majority class, leading to poor minority class detection and increased false negatives. To address this issue, Chawla et al. (2002) proposed the Synthetic Minority Oversampling Technique (SMOTE), which generates synthetic minority samples by interpolating between neighboring minority instances. SMOTE improves classifier sensitivity and minority class recognition by balancing the class distribution. Fernández et al. (2018) demonstrated that integrating SMOTE with ensemble learning significantly enhances classification performance in healthcare prediction tasks, while Haixiang et al. (2017) reported notable improvements in recall and F1-score for imbalanced medical datasets.

In addition to class imbalance, medical datasets are often characterized by uncertainty, vagueness, and imprecise information, which conventional machine learning methods may not adequately handle. To address this challenge, Zadeh (1965) introduced fuzzy logic as a mathematical framework for modeling uncertainty through partial membership representation. Unlike binary logic systems, fuzzy logic allows variables to belong to multiple classes simultaneously using membership functions. Janikow (1998) integrated fuzzy logic with decision tree learning to develop Fuzzy Decision Trees (FDTs), which improve classification under uncertain conditions while maintaining interpretability. In healthcare applications, fuzzy decision trees effectively transform numerical variables into linguistic categories such as low, medium, and high using triangular or trapezoidal membership functions, thereby enabling more flexible decision boundaries and improved predictive robustness. Recent studies have further demonstrated the effectiveness of combining intelligent learning techniques for cancer prediction and prognosis. Abbasi et al. (2024) applied ensemble models including Random Forest, Decision Trees, and Gradient Boosting for skin cancer survival analysis and achieved an impressive accuracy of 99%, highlighting the strength of ensemble learning in medical prognosis. Apasiya and Agbedemnab (2024) proposed a Type-2 fuzzy logic-based Enhanced Mobile-Based Fuzzy Expert System (EMFES) for breast cancer progression prediction, emphasizing adaptive rule generation and dynamic decision-making. Similarly, Zhu, Wang, and Zhang (2024) developed an improved fuzzy decision tree model integrating fuzzy rough sets and Z-numbers to enhance classification performance under uncertain conditions, while Ghantasala et al. (2024) introduced a Temporal Graph Neural Network model for ovarian cancer survival prediction to address limitations in temporal medical data analysis.

Although previous studies have demonstrated the effectiveness of machine learning, fuzzy logic, SMOTE, and ensemble learning in medical prediction tasks, several limitations still exist. Many existing studies fail to simultaneously address uncertainty and class imbalance within a unified predictive framework. Furthermore, limited research has integrated fuzzy decision trees with SMOTE-enhanced ensemble learning and optimized hyperparameter tuning for cervical cancer prediction. Therefore, this study proposes a comprehensive framework that combines fuzzy logic, SMOTE balancing, ensemble learning, and hyperparameter optimization to improve cervical cancer prediction performance, robustness, and clinical decision support.

METHODOLOGY

Dataset Description

The study utilized the Cervical Cancer (Risk Factors) dataset obtained from the UCI Machine Learning Repository. The dataset contains 858 patient records and 36 clinical and demographic attributes associated with cervical cancer risk factors. The features include age, smoking behavior, number of sexual partners, hormonal contraceptive usage, sexually transmitted diseases (STDs), and biopsy outcomes. The target variable was the Biopsy feature, indicating the presence or absence of cervical cancer.

Data Preprocessing

Data preprocessing involved handling missing values, normalization, and numerical conversion. Missing values represented by "?" symbols were replaced with NaN values and subsequently imputed using column mean substitution. The dataset was divided into input features (X) and target labels (y). To improve model learning, the dataset was normalized and transformed into fuzzy membership values using triangular membership functions. To address class imbalance, the Synthetic Minority Oversampling Technique (SMOTE) was applied to the training dataset. SMOTE generated synthetic minority class samples by interpolating between nearest-neighbor minority instances. The dataset was subsequently divided into 80% training data and 20% testing data.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue VI June 2026

Fuzzy Logic Implementation

Selected numerical variables including Age, Smoking Years, Number of Pregnancies, and STDs were transformed into fuzzy linguistic variables such as Low, Medium, and High using triangular membership functions. The general fuzzy membership function was defined as:

$$\mu_A(x) = \begin{cases} 0, & x \leq a \text{ ot } x \geq c \\ \dfrac{x-a}{b-a}, & a \leq x \leq b \\ \dfrac{c-x}{c-b}, & b \leq x \leq c \end{cases} \qquad 1$$

where (a, b, c) represent the parameters defining the fuzzy membership function. Each feature X_i is partitioned into three fuzzy linguistic categories: Low (a, b, c), Medium (d, e, f), and High (g,h,i). for the purpose of this study, Smoking Years, and Number of Pregnancies, and STDs numerical variables were transformed into fuzzy linguistic variables, as follows.

i. **Number of Pregnancies Membership Function**

Linguistic Variables: Few, Moderate, and Many.

Few:

$$\mu_{Few}(x) = \begin{cases} 1 & x \leq 1 \\ \dfrac{3-x}{3-1} & 1 < x \leq 3 \\ 0 & x \geq 3 \end{cases} \qquad 2$$

Moderate:

$$\mu_{Noderate}(x) = \begin{cases} 0 & x \leq 2 \\ \dfrac{x-2}{5-2} & 2 < x \leq 5 \\ \dfrac{8-x}{8-5} & 5 < x \leq 8 \\ 0 & x \geq 8 \end{cases} \qquad 3$$

Many:

$$\mu_{Many}(x) = \begin{cases} 0 & x \leq 6 \\ \dfrac{x-6}{10-6} & 6 < x \leq 10 \\ 1 & x \geq 10 \end{cases} \qquad 4$$

V. **Smoking Years Membership Function**

Linguistic Variables: Non-smoker, Moderate Smoker, and Heavy Smoker.

Non-smoker:

$$\mu_{Non-Smoker}(x) = \begin{cases} 1 & x \leq 0 \\ \dfrac{5-x}{5} & 0 < x \leq 5 \\ 0 & x \geq 5 \end{cases} \qquad 5$$

Moderate Smoker:

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue VI June 2026

$$\mu_{Noderate\ Smoker}(x) = \begin{cases} 0 & x \leq 3 \\ \frac{x-3}{10-3} & 3 < x \leq 10 \\ \frac{20-x}{20-10} & 10 < x \leq 20 \\ 0 & x \geq 20 \end{cases} \quad 6$$

Heavy Smoker:

$$\mu_{Heavy\ Smoker}(x) = \begin{cases} 0 & x \leq 15 \\ \frac{x-15}{30-15} & 15 < x \leq 30 \\ 1 & x \geq 30 \end{cases} \quad 7$$

Model Development

The study implemented several machine learning models for cervical cancer prediction, including decision tree algorithms such as ID3, C4.5, and CART, as well as ensemble learning models including Random Forest, Gradient Boosting Machine (GBM), XGBoost, LightGBM, and AdaBoost. To further enhance predictive performance and improve the handling of uncertainty in medical data, the outputs generated from the Fuzzy Decision Trees were integrated as meta-features into the ensemble learning models. The model builds a tree using:

i. Fuzzy ID3 uses entropy

$$H(Y) = -\sum plog_2 p \quad 8$$

ii. C4.5 Uses fuzzy gain ratio:

$$GR(X) = \frac{IG(X)}{SI(X)} \quad 9$$

where $SI(X)$ is the split implementation

$$SI(X) = -\sum P(X_i = \cup)\ Log_2\ P(X_i = \cup) \quad 10$$

iii. Fuzzy CART Uses Gini impurity as a measure:

$$Gini(X) = 1 - \sum_{j=1}^{m} P(Y_i|X_i)^2 \quad 11$$

iv. **Fusion of Multiple FDTs:** The aggregate predictions from Fuzzy ID3, Fuzzy C4.5, and Fuzzy CART using a weighted voting scheme: X_i

$$P(Y_j|x) = \sum_{t=1}^{T} w_t P_t(Y_j|x) \quad 12$$

Where $P(Y_j|x)$ is the class probability from the tree t. w_t is the weight assigned to tree t.

Hyperparameter tuning was applied using Bayesian Optimization, Grid Search, and Randomized Search techniques to improve model performance and robustness. The optimization process focused on key parameters

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue VI June 2026

such as learning rate, tree depth, number of estimators, and regularization coefficients to achieve optimal predictive accuracy and generalization capability.

Model Frameworks

The model framework for this study is illustrated in figure 1. The framework presents a hybrid fuzzy ensemble machine learning architecture designed for intelligent prediction and classification tasks. The process begins with a patient dataset in CSV format, which undergoes data preprocessing involving feature engineering, data cleaning, and transformation to generate a refined dataset. The cleaned data is then passed through a fuzzification stage where fuzzy membership functions are applied to convert input features into fuzzy sets capable of handling uncertainty and imprecision. The fuzzified dataset is used to build multiple fuzzy decision tree models, including Fuzzy C4.5, Fuzzy CART, and Fuzzy ID3, each generating independent predictions based on fuzzy conditions and decision rules. These predictions are integrated using an ensemble learning pipeline that combines bagging and boosting techniques to improve robustness, reduce classification errors, and enhance predictive performance. The final hybrid ensemble model undergoes training and hyperparameter optimization to achieve optimal performance. The effectiveness of the model is evaluated using standard performance metrics such as Accuracy, Precision, Recall, F1-Score, and AUC-ROC. Overall, the framework integrates fuzzy logic, decision trees, and ensemble learning to produce a robust and accurate predictive system suitable for complex and uncertain datasets in domains such as healthcare, cybersecurity, and intelligent systems.

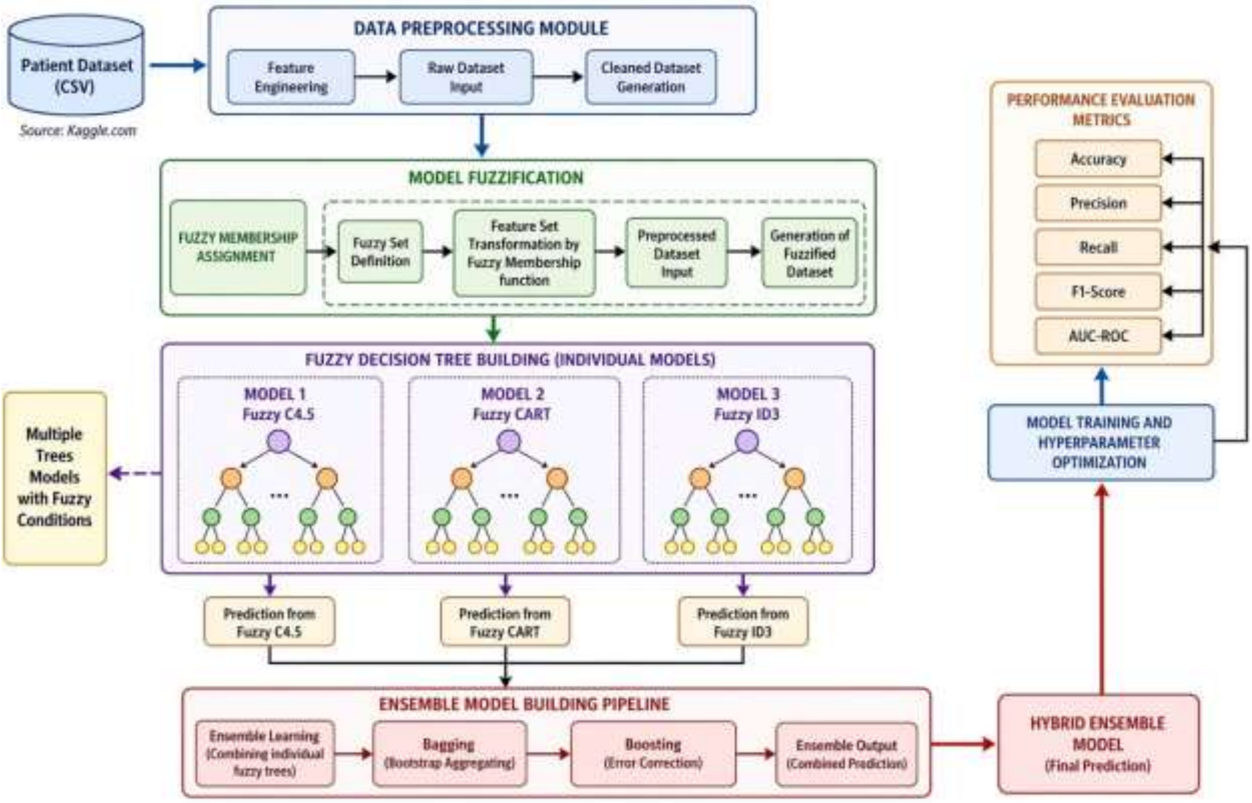

Figure 1: ProposedFuzzy Ensemble Decision Tree Framework for Intelligent Prediction and Classification

Evaluation Metrics

The developed models were evaluated using standard classification metrics including Accuracy, Precision, Recall, F1-Score, and ROC-AUC to comprehensively assess predictive performance. Accuracy measured the overall correctness of the models, while Precision and Recall evaluated the ability to correctly identify positive cervical cancer cases and minimize false predictions. The F1-Score provided a balanced measure of Precision and Recall, particularly for the imbalanced dataset, whereas ROC-AUC assessed the models' discriminative capability across different classification thresholds. In addition, confusion matrices and ROC curves were

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue VI June 2026

generated to provide comparative performance analysis and visualize the classification effectiveness of the models. These metrics are defined in equations (13) – (16).

i. **Accuracy**: Accuracy measures the overall correctness of the model. It is the proportion of total predictions that are correct. The Formula**:**

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \quad 13$$

ii. **Precision**: Precision measures how many of the predicted positive cases are actually positive. It focuses on false positives. The Formula is**:**

$$\text{Precision} = PPV = \frac{TP}{TP + FP} \quad 14$$

iii. **Recall (Sensitivity):** Recall measures how many actual positive cases are correctly identified. It focuses on false negatives. The formula is as follows**:**

$$\text{Sensitivity or Recall} = \text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad 15$$

iv. **F1-Score**: F1-Score is the harmonic mean of Precision and Recall. F1-Score balances both false positives and false negatives. Important when missing a positive case is costly (e.g., disease detection, malware detection). The formula is stated below:

$$\text{F1} - \text{Score} = 2 \text{ x } \frac{\text{Precision x Recall}}{\text{Precision} + \text{Recall}} \quad 16$$

RESULTS

The experimental results demonstrated the effectiveness of combining SMOTE, fuzzy logic, and ensemble learning techniques for cervical cancer prediction. The correlation heatmap analysis, shown in figure 1 revealed low multicollinearity among most variables, indicating that the dataset contained largely independent predictive features. However, strong positive correlations were observed among STD-related variables and smoking-related features, suggesting possible feature redundancy.

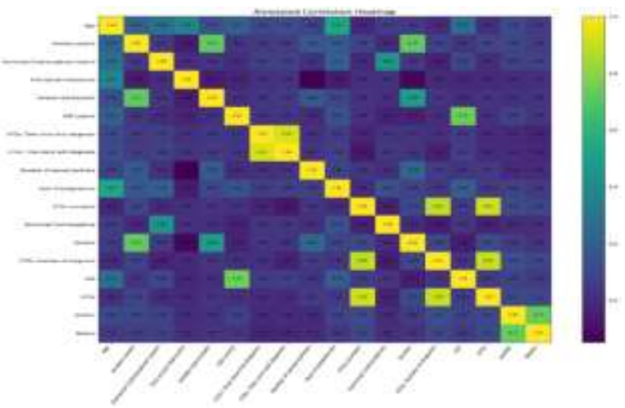

Figure 1. Heatmap of the Cervical Cancer Dataset

The comparative evaluation showed that ensemble learning models outperformed standalone decision tree algorithms. Gradient Boosting achieved the highest classification accuracy of 97.09% and F1-score of 73.68%, indicating strong predictive performance and balanced classification capability. The C4.5 model also demonstrated stable performance with balanced precision and recall values of 72.73%. In contrast, AdaBoost and Random Forest recorded lower recall values, indicating weaker sensitivity toward positive cervical cancer cases. Table 1 illustrate the performance of the models before smote. The visual representation of the models

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue VI June 2026

performance in terms of accuracies before the application of SMOTE is shown in figure 2, while figure 3 show the performance of the models in terms of ROC-AUC.

Table 1: The Performance of the implemented models before SMOTE

Model	Accuracy	Precision	Recall	F1 Score	ROC-AUC
ID3	0.953488	0.666667	0.545455	0.600000	0.763411
C4.5	0.965116	0.727273	0.727273	0.727273	0.854320
CART	0.947674	0.600000	0.545455	0.571429	0.760305
Random Forest	0.947674	0.666667	0.363636	0.470588	0.675607
Gradient Boosting	0.970930	0.875000	0.636364	0.736842	0.815076
AdaBoost	0.936047	0.500000	0.363636	0.421053	0.669396
Optimized RF	0.959302	0.750000	0.545455	0.631579	0.766516
Optimized GB	0.953488	0.714286	0.454545	0.555556	0.721062

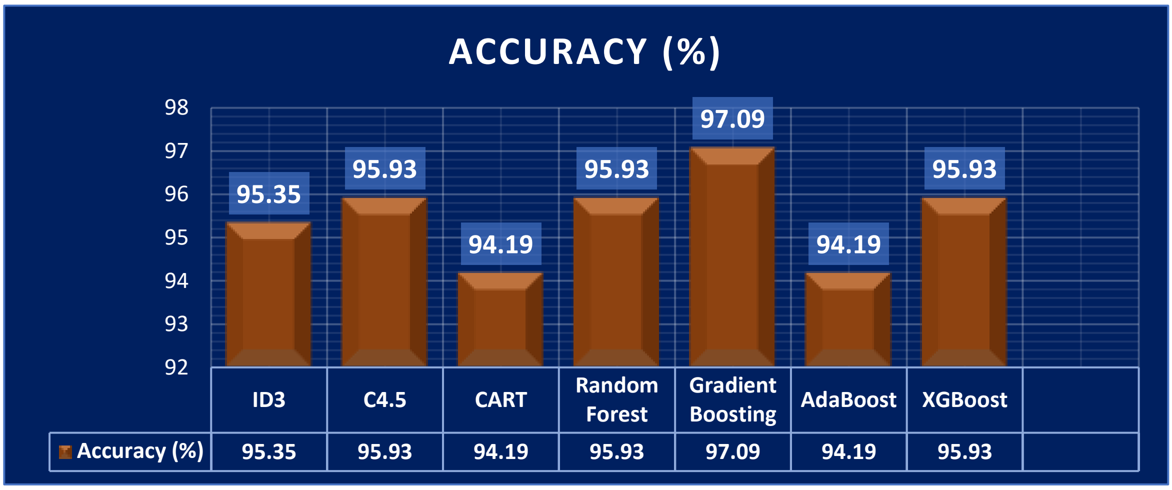

Figure 2: Performance Evaluation of the implemented models before SMOTE

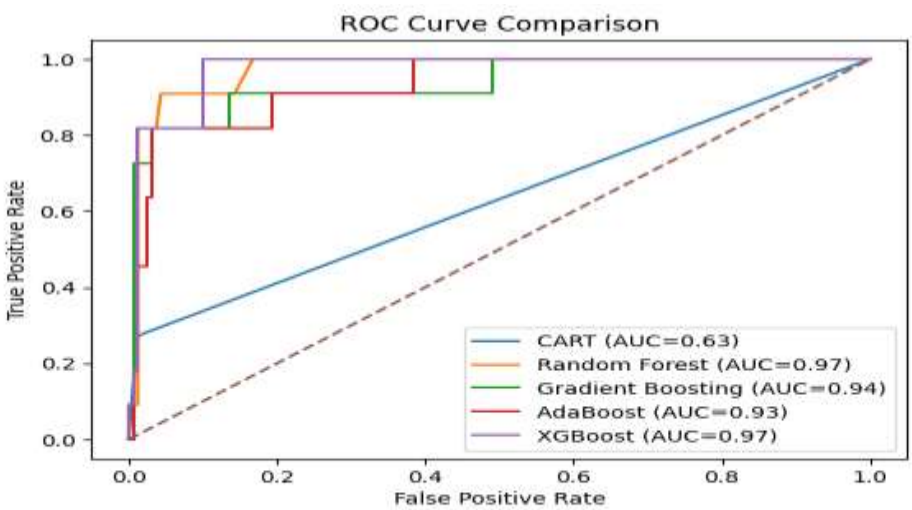

Figure 3. Performance Evaluation of the Models in terms of ROC-AUC

INTERNATIONAL JOURNAL OF RESEARCH AND INNOVATION IN APPLIED SCIENCE (IJRIAS)
ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue VI June 2026

The application of SMOTE significantly improved minority class representation and enhanced recall performance compared to models trained on imbalanced datasets. The integration of fuzzy logic further improved the handling of uncertainty in medical features. The confusion matrix of the benchmarked XGBoost model showed high specificity with very few false positive predictions. However, several minority class instances were still misclassified, indicating the need for further sensitivity improvement. ROC-AUC analysis showed that Random Forest and XGBoost achieved the highest AUC values of 0.97, demonstrating excellent discriminative capability. Gradient Boosting also recorded strong performance with an AUC of 0.94. The performance of the models after applying SMOTE is shown in table 2.

Table 2: Performance of the Models after applying SMOTE (Synthetic Minority Over-sampling Technique)

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	ROC-AUC (%)
ID3	95.35	71.43	45.45	55.56	72.11
C4.5	95.93	75.00	54.55	63.16	89.89
CART	94.19	60.00	27.27	37.50	63.02
Random Forest	95.93	70.00	63.64	66.67	97.21
Gradient Boosting	97.09	87.50	63.64	73.68	94.35
AdaBoost	94.19	55.56	45.45	50.00	93.28
XGBoost	95.93	75.00	54.55	63.16	97.35

Overall, the results confirm that the proposed SMOTE-enhanced fuzzy ensemble learning framework significantly improves cervical cancer prediction performance and provides a reliable decision-support mechanism for early diagnosis. The visual representation of the models performance in terms of accuracies and ROC-AUC after the application of SMOTE is shown in figures 4 and 5.

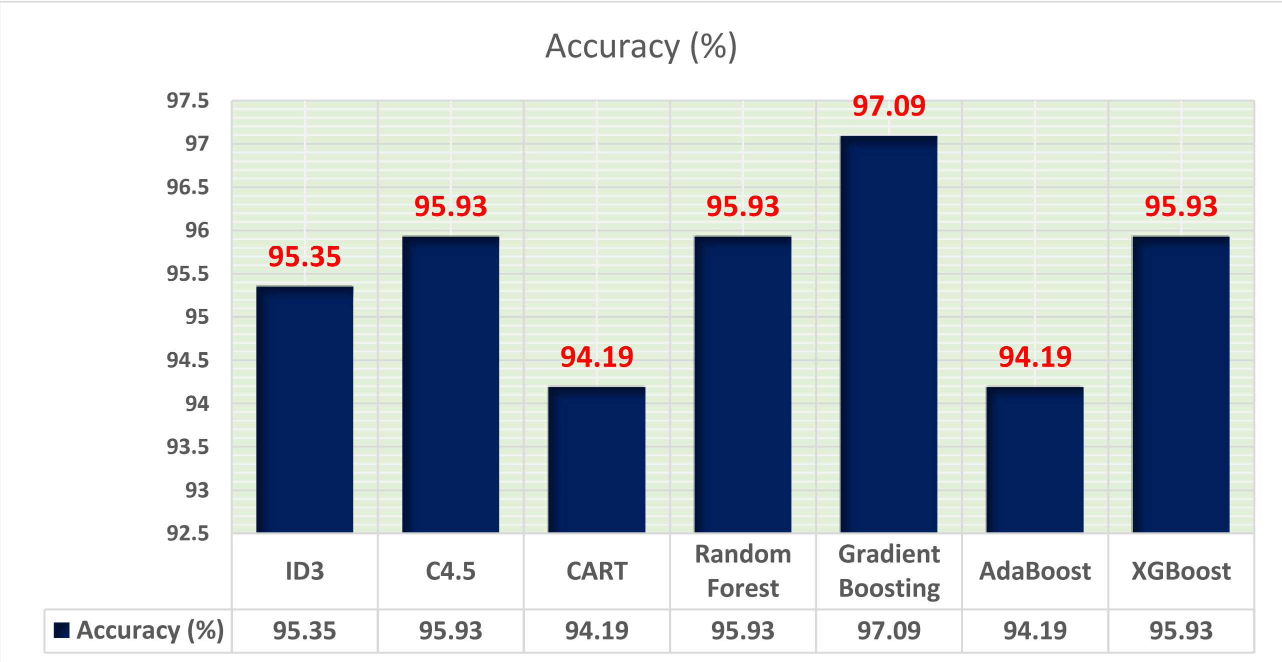

Figure 4. The Models Performance in Terms of Accuracies after Applying SMOTE

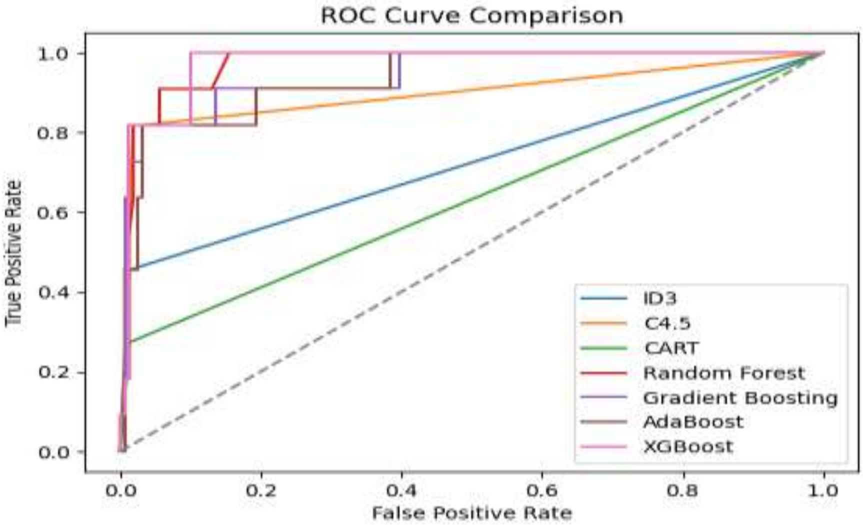

Figure 5: Models Performance in terms of ROC-AUC after Applying SMOTE

The ROC curve in figure 5 shows that the ensemble models, particularly XGBoost, Random Forest, and Gradient Boosting, achieved the best classification performance because their curves are closest to the top-left corner, indicating high true positive rates and low false positive rates. AdaBoost and C4.5 also performed well but slightly below the leading ensemble methods. In contrast, ID3 and CART showed weaker predictive ability, as their curves are closer to the diagonal reference line. Overall, the results indicate that ensemble learning techniques outperform traditional decision tree algorithms in classification accuracy and discriminative capability.

DISCUSSIONS

The experimental findings demonstrate that integrating SMOTE, fuzzy logic, and ensemble learning significantly enhances cervical cancer prediction performance. Correlation analysis revealed low multicollinearity among most variables, indicating that the dataset contains diverse and informative predictors, although strong associations among smoking and STD-related features suggest important clinical relationships. Before applying SMOTE, ensemble models outperformed conventional decision trees, with Gradient Boosting achieving the best performance, recording 97.09% accuracy, 87.50% precision, 63.64% recall, and an F1-score of 73.68%. C4.5 also showed balanced predictive capability, while Random Forest and AdaBoost exhibited lower recall values, highlighting the challenges posed by class imbalance in identifying positive cancer cases. The application of SMOTE substantially improved minority-class representation and enhanced model sensitivity. XGBoost and Random Forest achieved the highest ROC-AUC values of 97.35% and 97.21%, respectively, demonstrating excellent discriminative ability. These results confirm that data balancing techniques play a crucial role in improving cancer prediction models. Fuzzy logic further strengthened the framework by effectively handling uncertainty and imprecision in medical data. The transformation of numerical variables into linguistic categories enabled the generation of interpretable decision rules, thereby improving explainability and supporting clinical decision-making. The integration of fuzzy decision tree outputs into ensemble models contributed to greater robustness and predictive stability. In general, the results indicate that the proposed SMOTE-enhanced fuzzy ensemble framework successfully addresses class imbalance, uncertainty, and interpretability challenges simultaneously. The combination of fuzzy reasoning, ensemble learning, and hyperparameter optimization provides a reliable and effective decision-support mechanism for early cervical cancer diagnosis and intelligent healthcare applications.

CONCLUSION

This study developed an explainable SMOTE-enhanced fuzzy ensemble decision tree framework for cervical cancer prediction and classification. The integration of SMOTE effectively addressed class imbalance, while fuzzy logic improved the handling of uncertainty in medical data. The ensemble learning models consistently outperformed traditional decision tree algorithms, with Gradient Boosting achieving the highest classification

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue VI June 2026

accuracy and XGBoost producing the best ROC-AUC performance. The findings demonstrate that combining fuzzy reasoning, data balancing techniques, and optimized ensemble methods provides a robust, accurate, and interpretable approach for early cervical cancer diagnosis. Therefore, the proposed framework offers a reliable decision-support tool with strong potential for intelligent healthcare applications and improved patient outcomes.

REFERENCES

1. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P.(2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321–357.
2. Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324.
3. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794.
4. Fernández, A., Garcia, S., Galar, M., Prati, R. C., Krawczyk, B., & Herrera, F. (2018). Learning from imbalanced data sets. Springer. https://link.springer.com/book/10.1007/978-3-319-98074-4.
5. Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. Annals of Statistics, 29(5), 1189–1232.
6. Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G. (2017). Learning from Class-Imbalanced Data: Review of Methods and Applications. Expert Systems with Applications, 73, 220–239.
7. Janikow, C. Z. (1998). Fuzzy decision trees: Issues and methods. IEEE Transactions on Systems, Man, and Cybernetics, 28(1), 1–14.
8. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T. Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems, 30.
9. Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. Computational and Structural Biotechnology Journal, 13, 8–17.
10. Quinlan, J. R. (1993). C4.5: Programs for machine learning. Morgan Kaufmann.
11. World Health Organization. (2023). Cervical cancer. https://www.who.int/news-room/fact-sheets/detail/cervical-cancer
12. Zadeh, L. A. (1965). Fuzzy sets. Information and Control, 8(3), 338–353.