INTERNATIONAL JOURNAL OF MODERN TECHNOLOGY AND ENGINEERING RESEARCH
Volume-3, **Issue**-1, 2025
https://caarnjournals.com/index.php/ijmter/index
ISSN (Online): 2581-5792
A Peer Reviewed (Refereed) International Journal

IMPROVED HYBRID MODEL FOR CERVICAL CANCER RISK PREDICTION BASED ON ENSEMBLE LEARNING METHOD

[1]Ogunmodimu Dupe Catherine, [2]C. Ugwu, F. Egbono
[1]Department of Computer Science and Informatics, Federal University Otuoke, Bayelsa State,Nigeria.
[2]Department of Computer Science, University of Port Harcourt, Rivers State, Nigeria
email: dupeogunmodimu@gmail.com

ABSTRACT

One of the prominent health issues faced by women globally, particularly in developing nations, is cervical cancer, also known as cervical magnate. This cancer starts in the cervix, often going unnoticed in its early stages until symptoms manifest later, potentially indicating metastasis. Early detection of cervical cancer significantly increases the chances of treatment and cure. This study focuses on developing a model to help women assess their risk of cervical cancer based on demographic and medical history. It introduces an enhanced hybrid model employing ensemble learning techniques to improve predictive accuracy. The model utilizes a dataset consisting of demographic data, lifestyle habits, and medical histories of 858 patients obtained from the UC Irvine machine learning repository. A pipeline combining a transformer, sampler, and estimator was developed to mitigate overfitting and data leakage while enhancing model performance. This pipeline utilized StandardScaler for transformation, SmoteTomek for sampling, and the XGBoost classifier as the estimating mechanism. A conventional XGBoost classifier was trained to identify the top 12 important features that impact the performance of the classification model. The proposed model successfully identified 100% of at-risk women, achieving a reported accuracy of 99% and a 100% recall rate. Overall, this hybrid model significantly outperforms existing methods in detecting women at risk of developing cervical cancer, yielding superior accuracy, sensitivity, and specificity in cervical cancer risk prediction.

KEYWORDS: Cervical Cancer, Cervical Magnate, Ensemble Learning, Pipeline, SmoteTomek, XGBoost.

1. Introduction

In recent years, there has been a growing interest in using machine learning methods for medical diagnosis, particularly ensemble learning. According to Mahajan et al. [1], ensemble learning is an effective approach that combines multiple learning algorithms and classification models. This method is crucial for addressing the limitations inherent in any single classification technique. The need for ensemble learning arises from the desire to improve overall model performance, minimize the risk of overfitting to the training dataset, and reduce model bias. By utilizing ensemble learning, we can enhance prediction accuracy and overcome the shortcomings of individual classification techniques.

Making decisions based on input from multiple individuals or experts has long been a common practice in human civilization, forming the foundation of a democratic society. In recent decades, researchers in computational intelligence and machine learning have explored methods that utilize a joint decision-making process.

These techniques are called ensemble learning, which seeks to minimize classifiers' variance and improve decision-making systems' robustness and accuracy. Ensemble learning is an overarching approach in machine learning that enhances predictive performance by integrating the predictions from multiple models [2]. Ensembles are predictive models that aggregate predictions from two or more other models [3]. Ensemble learning is a machine learning paradigm where multiple learners are trained to solve the same problem. [4].

Ensemble learning is a machine learning strategy that combines multiple algorithms into a cohesive framework. This approach effectively leverages each algorithm's complementary strengths to improve the model's overall performance [5].

The individual learners of the ensemble, which are combined strategically, are referred to as base learners. These basic models usually do not perform well because they have a high bias or too much variance to be robust. Ensemble methods aim to reduce the bias and/or variance of weak learners by combining multiple learners to create a strong learner or ensemble model that performs better.

Cervical cancer is a type of cancer that originates in the cervix, which is a hollow cylindrical structure connecting the lower part of a woman's uterus to her vagina. Most cervical cancers start in the cells on the surface of the cervix. Cervical cancer is the eighth most common cancer worldwide and the fourth most prevalent among women. It ranks as the second most common type of female cancer globally. In Nigeria, cervical cancer is the second most frequently diagnosed cancer among women. It is the second leading cause of cancer-related deaths for women aged 15 to 44 years, according to [6].

This study explored the XGBoost homogeneous ensemble learning method combined with the SmoteTomek hybrid data resampling technique to reduce bias, which has led to misclassification problems in previous studies. The goal is to correctly identify women who are at risk of developing cervical cancer, i.e., have a better true positive value and minimize false negatives (type II error) to optimize the model's generalization performance and robustness.

This study used the risk factors associated with cervical cancer and applied ensemble learning to predict a woman's risk of developing the disease. Ensemble learning methods are increasingly being used to tackle real-world challenges and play a vital role in the medical field, especially in disease diagnosis.

2. REVIEW OF RELATED WORKS

The review of related works provides a concise overview of existing research on ensemble learning techniques for detecting cervical cancer and other relevant studies related to the techniques employed in this study. The review also narrates the various algorithms and approaches used.

Jagwani [7] developed a model to identify patients at risk of developing stroke using electronic health records obtained from hospitals and medical institutions. It included nine classification models, including Extreme Gradient Boost (XGBoost). Random Forest, Support Vector Classifier (SVC), Neural Network, Naïve Bayes, Logistic Regression, K-Nearest Neighbours (KNN), Decision Tree Classifier, AdaBoost, along with the Ensemble Voting Classifier, are developed and evaluated using three different resampling techniques, such as SMOTE, Tomek Links, and SMOTE + Tomek. Their findings showed that combining the Ensemble Voting Classifier and the hybrid sampling technique (SMOTE + Tomek) achieved the best results.

Jiang and Chen [8] proposed an anomaly detection and classification method specifically designed for industrial control systems (ICSs). This method relies on network traffic data from industrial field protocols, including Modbus TCP and S7 Communication. To address the data imbalance problem, they employed the Synthetic Minority Oversampling Technique (SMOTE) and the Tomek link (T-Link) mechanism for oversampling and undersampling data, respectively. Extreme Gradient Boosting (XGBoost) was utilized to leverage ensemble learning, avoiding overfitting and ensuring robust performance. The proposed method was evaluated using a real-world dataset from the railway industry's ICS called Electra. The performance results were compared with those of other related methods. The findings indicated that the proposed approach achieved 100% precision, recall, and F1 score in anomaly detection, along with relatively high performance in anomaly classification. Their study concluded that integrating SMOTE, T-Link, and XGBoost techniques leads to exceptionally high performance in ICS anomaly detection and classification based on network traffic data.

Fang et al. [9] conducted a comparison study between the autoregressive integrated moving average (ARIMA) model and the Extreme Gradient Boosting (XGBoost) model to determine which was more accurate for forecasting the occurrence of COVID-19 in the USA. Three accuracy metrics were used to evaluate the performance of the two models: mean absolute error (MAE), root mean square error (RMSE), and mean absolute percentage error (MAPE). For both the training and validation sets, the MAE, RMSE, and MAPE values for the XGBoost model were lower than those for the ARIMA model. Their results showed that the XGBoost model can help improve the prediction of COVID-19 cases in the USA over the ARIMA model.

Kanitkar et al. [10] developed a system that used demographic data to detect whether a patient has a cervical abnormality. They compared the performance of ten different ML algorithms. The authors used Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Tree, Artificial Neural Network (ANN), Naïve Bayes, Random Forest, Stochastic Gradient Descent (SGD), Voting Classifier, and Ada Boost. SVM with PCA (Principal Component Analysis) and RFE were successfully used to extract highly correlated risk factors. The results of their binary classification in terms of accuracy, precision, recall, and F1-score showed that the Artificial Neural Network (ANN) performed better than other algorithms with an accuracy of 75%, precision of 0.78, recall of 0.74, and F1-score of 0.75, respectively.

Rahman [11] proposed a classification model that can diagnose cervical cancer by combining the predictions of three classifiers: The voting technique was used with Tree, Logistic Regression, and Random Forest. Principal Component Analysis (PCA) was used to reduce dimensions, while the SMOTE technique was applied to balance the cervical cancer dataset obtained from the UCI Machine Learning Repository. Their study showed that balancing the dataset with SMOTE improved the model's performance.\

Lu et al. [12] proposed a voting ensemble strategy to predict the risk of cervical cancer. A filling technique with a correction mechanism was adopted to address missing values. Logistic regression, decision tree, support vector machine, multilayer perception, and k-nearest neighbour were used for the voting strategy with grid search. The performance measure was based on accuracy, recall, precision, and F1 score. The result of the study showed that the proposed voting method had an accuracy of 83.16%, a recall of 28.38%, a precision of 51.75%, and an f-1 score of 32.80%. However, their result can be improved for better performance.

Kumar [13] explored a range of predictive models on the breast cancer Wisconsin dataset. They developed an ensemble model utilizing several algorithms, including k-nearest neighbours, random forest, logistic regression, support vector machine, decision tree, AdaBoostM1, gradient boosting, stochastic gradient boosting, XGBoost classifier, and CatBoost. The model's performance was evaluated using recall, precision, and F-measure metrics. The proposed model achieved a maximum accuracy of 99.45%.

Srivastava et al. [14] proposed a hybrid Extreme Gradient Boosting (XGBoost) model with hyperparameter tuning to improve early detection, diagnosis, and risk reduction for liver disease. Their study utilized a dataset comprising 416 individuals with liver issues and 167 without such a history. The results showed that the chi-square automated interaction detection model achieved an accuracy of 71.36%, while the classification and regression trees model reached an accuracy of 73.24%. Both models significantly outperformed conventional methods.

Win et al. [15] proposed a framework for detecting and classifying cervical cancer from pap smear images. They combined K-means clustering with morphology operations to obtain good

segmentation for cell nuclei and cytoplasm. Random Forest was used for feature selection, while majority voting was used to aggregate the results of five classifiers: SVM, Linear Discriminant (LD), Boosted Trees, Bagged Trees, and K-nearest neighbour. Their proposed method obtained an accuracy of 81.54% with sensitivity and specificity of 77.43% and 90.59%, respectively.

Asadi et al. [16] used a supervised machine-learning algorithm to predict cervical magnates. They used data from 145 patients with 23 attributes. Their study showed that a decision tree can be used to identify the most relevant features for predicting cervical cancer disease.

Rahman et al. [17] proposed a voting method integrating three classifiers: Decision Tree, Logistic Regression, and Random Forest. To address the issue of an imbalanced dataset, they utilized the Synthetic Minority Oversampling Technique (SMOTE). Additionally, they employed Principal Component Analysis (PCA) to reduce dimensions that do not influence model accuracy. To prevent overfitting, they implemented a stratified 10-fold cross-validation technique.

Zhice et al. [18] have introduced four heterogeneous ensemble learning techniques to predict landslide susceptibility. They combined several classifiers, such as Support Vector Machine, Convolutional Neural Network, Logistic Regression, and Recurrent Neural Network; the result of their study shows that the ensemble learning method shows higher prediction accuracy than individual classifiers; the blending classifier method achieves the highest overall accuracy at 80.70%, outperforming other ensemble learning methods.

Even though the efficiency of machine learning in disease diagnosis has improved in recent years, it is important to identify the gaps in previous studies. Most of the models previously proposed by several researchers used conventional machine learning methods such as K-Nearest Neighbor (KNN), Decision Tree, Support Vector Machine (SVM), Multilayer Perceptron, Logistic Regression, Convolutional Neural

Network (CNN), and Naïve Bayes, which are prone to bias, overfitting, and low generalization performance.

High classification error (type II error) is another problem of most previously developed models for cervical cancer detection. Type II error is a misclassification problem where the model predicts that a patient is not at risk of a disease, but in reality, they are at risk. This results in misleading information (false alerts) about the patient. Misclassification problems affect the generalization performance of machine learning models.

This research intends to adopt an ensemble learning method with a hybrid data resampling technique to address the misclassification problem in the previous study to correctly identify the patients who are actually at risk of developing cervical cancer. i.e., have a better true positive value and minimize false negatives to optimize the model's generalization performance and robustness.

3 MATERIALS AND METHODS

3.1 System Analysis

The cervical cancer risk prediction model is an enhanced hybrid model that predicts whether a woman is at risk of developing cervical cancer. To overcome the shortcomings of the existing models, the proposed system adopts the following strategies.

1. The SmoteTomek hybrid resampling technique was adopted at the data level to handle model bias and misclassification.

2. The XGBoost ensemble learning method was adopted at the algorithm level to boost the model's overall performance.

3. A pipeline was constructed during model training to prevent data leakage, which could lead to overfitting.

The proposed framework for this research study is divided into various components, as shown in the architecture of the proposed system in Figure 1.

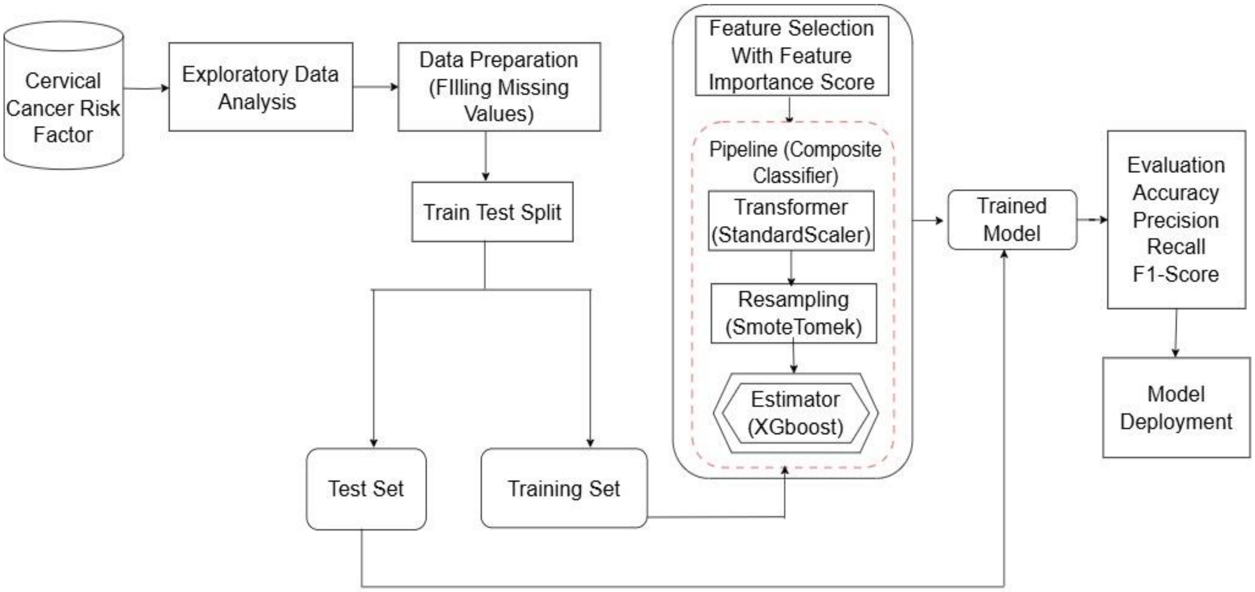
Cervical Cancer Risk Factor
Exploratory Data Analysis
Data Preparation (FIlling Missing Values)
Train Test Split
Test Set
Training Set
Feature Selection With Feature Importance Score
Pipeline (Composite Classifier)
Transformer (StandardScaler)
Resampling (SmoteTomek)
Estimator (XGboost)
Trained Model
Evaluation Accuracy Precision Recall F1-Score
Model Deployment

Figure 1: Architecture of the Proposed System

3.2 Dataset

The cervical cancer risk factor dataset used in this research study is from Hospital University de Carcus and was obtained from the UC Irvin Repository. The raw data is the real electronic health record of patients. It contains 858 cases (patients). All instances of the dataset have 36 features, including the target variables. The target variables are the decision-makers that confirm whether the person is diagnosed with cervical (1) or not (0). The dataset contains various patient attributes, including demographic data, sexual behaviour, smoking habits, contraceptive usage, history of sexually transmitted diseases (STDs), and diagnostic outcomes related to cervical cancer. The dataset is highly imbalanced; the challenge of the imbalanced dataset is addressed using SmoteTomek hybrid resampling technique.

3.3 Exploratory Data Analysis

Exploratory Data Analysis (EDA) was conducted on the dataset to uncover underlying patterns and trends. This process involved visualizing the data through various graphs and charts, as well as performing statistical analyses to summarize key characteristics. By examining distributions, correlations, and potential outliers, we aimed to gain a deeper understanding of the dataset. Figure 2 illustrates how the target variables are distributed in the cervical cancer risk factor dataset.

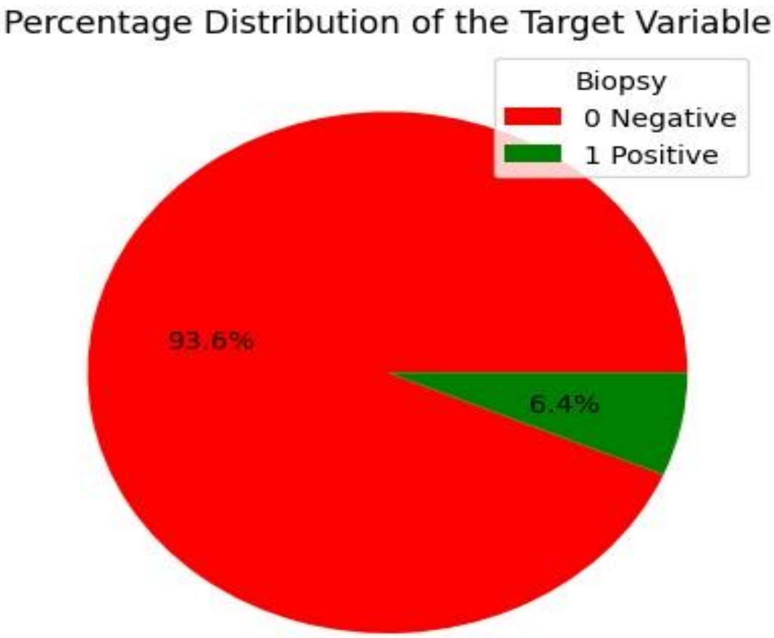

Figure 2: Percentage Distribution of the Target Variable

Figure 2 shows that 93.6% of the patients tested positive for cervical cancer, while 6.4% tested negative based on the outcome of their biopsy results. This target variable distribution shows that the dataset is highly imbalanced.

4. DATA PREPROCESSING

4.1 Filling Missing Values

There are a lot of missing values in the dataset due to patients not given answers to certain private questions. If the missing values/data in the dataset are not handled properly, we might build a biased model, which can lead to incorrect results. Hence, we adopted two techniques for handling missing values: these techniques are;

4.1.1 Dropping Technique

This technique deletes rows or columns with missing data; if any columns have more than half of their values as null, the entire column can be deleted. In the same way, rows can also be dropped if they have one or more null column values. Due to the size of the dataset, instead of dropping all the columns with missing values, we dropped only the two columns with 91.72% missing values. Then we used the imputation method to handle the remaining missing values.

4.1.2 Imputation Method

Missing value Imputation, or replacement techniques, assist machine learning models in learning from incomplete data. This approach is utilized in this study because removing data from the dataset each time is not practical and can significantly reduce the dataset's size. Such a reduction raises concerns about bias and can lead to inaccurate analysis. After dropping the two columns with more than 90% missing values, we used the imputation method (replacing missing data with mean values) to handle the rest of the missing values because our dataset size is not very big, and removing many parts of it can have a significant impact on the final model.

4.2 Data Splitting Process

The Stratified Train-Test sampling technique is adopted in this study to split the dataset into the training set and the test set in a ratio of 80:20. This technique was used to create an unbiased dataset during the data-splitting process. Since the dataset is skewed, the stratified train-test split ensures that the training and the test sets have the same proportion of target classes as in the original dataset. The technique also plays a crucial role in correctly classifying skewed datasets. The scikit-learn train-test split with the stratified parameter is used to split the dataset into training and test sets.

4.3 Feature Scaling Process

A standardization technique known as Standard Scaler was adopted in this research study to transform the dataset features into the same scale while building the model. The idea behind Standard Scaler is that it will transform the data such that its distribution will have a mean value of 0 and a standard deviation of 1. The standardization is calculated as follows:

Standardization:

$$z = \frac{x - \mu}{\sigma} \qquad 1$$

$mean$:

$$\mu = \frac{1}{N} \sum_{i=1}^{N} (x_i) \qquad 2$$

$standard\ deviation$:

$$\sigma = \sqrt{\frac{1}{N} \sum_{i-1}^{N} (x_i - \mu)^2} \qquad 3$$

Using this formula, we replaced all the input values with the Z-score for every value. Hence, we get values ranging from -1 to +1, keeping the range intact.

Standardization performs the following:

- Converts the Mean (μ) to 0
- Converts the S.D. (σ) to 1

4.4 Feature Selection Process

The XGboost feature importance ranking for feature selection was used to identify the most relevant features in the dataset that can be used to train the proposed model to best identify women who are at risk of developing cervical cancer. The performance measure used for feature importance ranking in this study is the Gini index. when training a tree, it is possible to compute how much each feature decreases the impurity. The more a feature decreases the impurity, the more important the feature. Table 1 shows the 12 most important features used in this study for cervical cancer risk prediction according to their feature importance score using XGbbost classifier.

Table 1: 12 most important features used in this study for cervical cancer risk prediction

Features	Feature Importance Score
Age	0.181411
Number of sexual partners	0.157368
First sexual intercourse	0.133497
Num of pregnancies	0.111059
Smokes (years):	0.098403
Smokes (packs/year	0.055555
Hormonal Contraceptives (years)	0.054618
IUD (years)	0.051533
Dx:Cancer	0.029722
Dx:HPV	0.018297
STDs:HIV	0.016007
Dx:CIN	0.013990

4.6 Data Resampling with SmoteTomek

This research study used a hybrid approach to handle class imbalance. A resampling technique known as SmoteTomek was utilized to handle the class imbalance, as 93.6% of the target variable belongs to the majority class, while 6.4% belongs to the minority class. Training the proposed model without handling class imbalance, the model might end up learning how to predict the majority class and fail to learn how to predict the minority class, which is more important for this research study.

SmoteTomek is a hybrid resampling technique that combines the SMOTE ability to generate synthetic data for the minority class and Tomek Links's ability to remove the data that are identified as Tomek links from the majority class (that is, samples of data from the majority class that are closest to the minority class data). SMOTE (Synthetic Minority Oversampling Technique) is used to oversample the minority class. after oversampling, Tomek Links is applied to the oversampled data to remove noisy and overlapping data generated by SMOTE. The steps involved in the process of SMOTE-Tomek are shown below:

Step 1: Start of SMOTE: choose random data from the minority class.
Step 2: Calculate the distance between the random data and its k nearest neighbors.
Step 3: Multiply the difference with a random number between 0 and 1, then add the result to the minority class as a synthetic sample.
Step 4: Repeat steps 2–3 until the desired proportion of minority class is met.
End of SMOTE.
Step 5: Start of Tomek Links: choose random data from the majority class
Step 6: If the random data nearest neighbour is the data from the minority class (i.e., create the Tomek Link), remove the Tomek Link.
End

5. SYSTEM IMPLEMENTATION

The model was implemented in an Anaconda 3 environment using Jupyter Notebook with Python 3.9.11. Scikit-learn was used for data manipulation using pandas, data visualization using matplotlib and seaborn, mathematical operations/array processing using numpy, and also for the implementation of the XGBoost algorithm. The model was deployed into production using Flask framework.

6. RESULTS AND DISCUSSION

We implemented a conventional XGBoost classifier as a base model and an XGBoost classifier as an estimator combined with SmoteTomek in a pipeline. The performance of the proposed model is based on accuracy (Acc), precision (Pre), recall (Rec), and F1-score, as shown in the classification reports. The results obtained on the available data showed that by the application of the conventional XGBoost model in the prediction of cervical cancer risk, the accuracy value of 93% was obtained while for the proposed model based on ensemble learning method using XGboost combined with SmoteTomek resampling technique in a pipeline, an accuracy of 99% was obtained. Table 2 and Figure 3 show the classification report and the confusion matrix of the conventional XGBoost model used in the existing study. Table 3 and Figure 4 show the classification report and the confusion matrix of the proposed model after implementing the XGboost classifier with SmoteTomek resampling algorithm in a pipeline. We can see from the results how the existing model suffered from huge misclassification (False Negative; Type II Error) despite having an accuracy of 93%. Combining Xgboost classifier with SmoteTomek drastically reduces this misclassification in the proposed system. From the statistics obtained from the classification reports, it is possible to get other statistically significant information, such as recall, precision, and F1 values, as presented in Tables 2 and 3

Table 2: Classification Report of the Conventional XGboost Classifier

	Precision	Recall	F1-Score	Support
0	0.96	0.97	0.96	164
1	0.17	0.12	0.14	8
Accuracy			0.93	172
Macro avg	0.56	0.55	0.55	172
Weighted avg	0.92	0.93	0.93	172

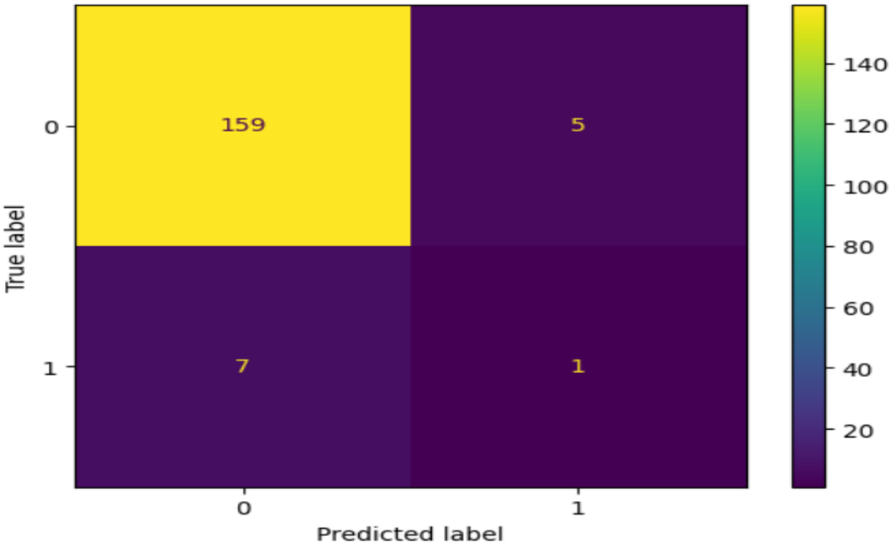

Figure 3: Confusion Matrix of the Conventional XGboost Classifier

Table 3: Classification Report of the Proposed System

	Precision	Recall	F1-Score	Support
0	1.00	0.99	1.00	164

1	0.89	1.00	0.94	8
Accuracy			0.99	172
Macro avg	0.94	1.00	0.97	172
Weighted avg	0.99	0.99	0.99	172

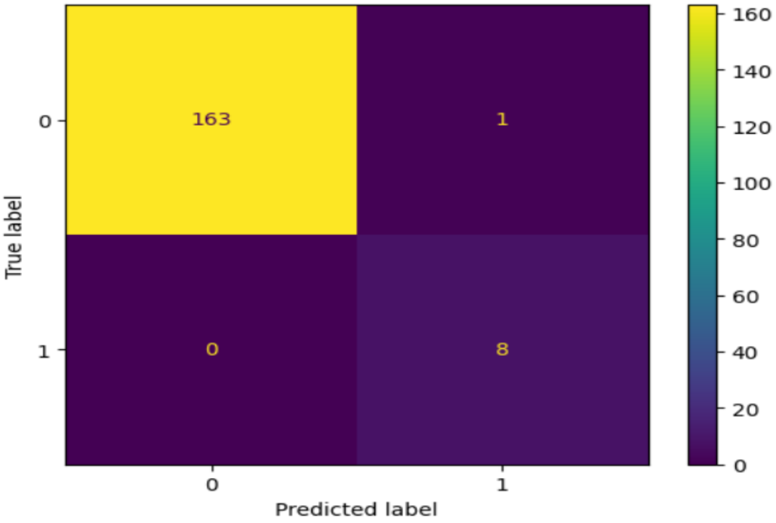

Figure 4: Confusion Matrix of the Proposed System

The bar chart in Figures 5 and 6 compares the performance of the existing and the proposed system in terms of accuracy, precision, recall, and F1-Score for the classification of negative and positive cases of cervical cancer.

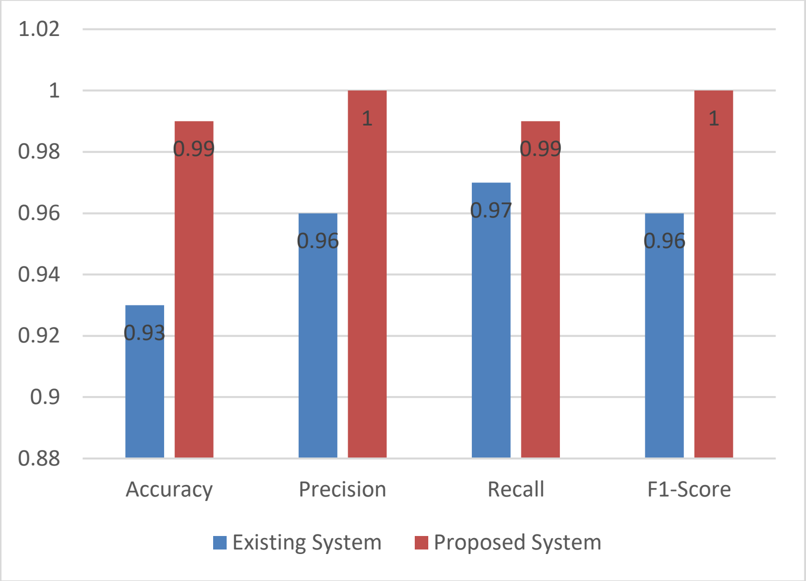

Figure 5: Graph Showing Comparison of the Existing and Proposed System Based on Negative Case Predictions

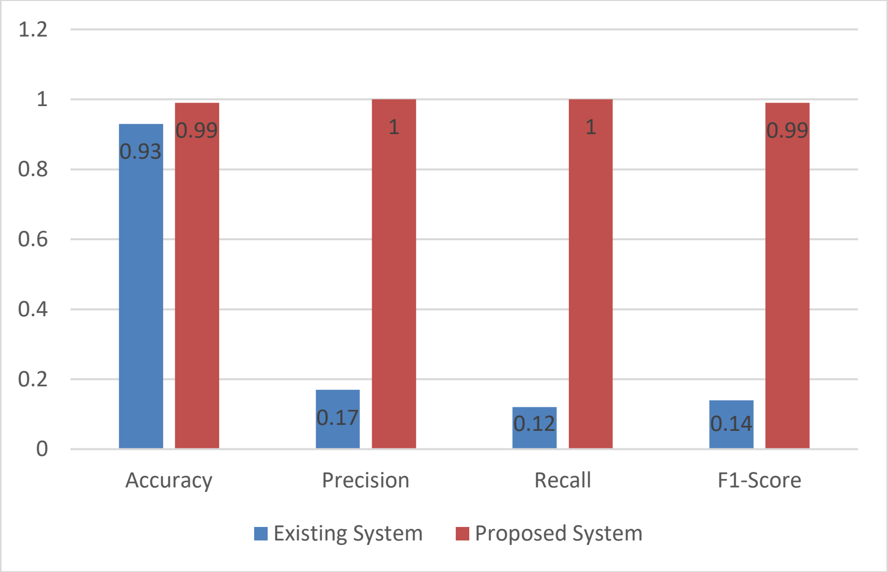

Figure 4.12: Graph Showing Comparison of the Existing and Proposed System Based on Positive Case Predictions

From Table 2, the first metric is the precision, the precision is a measure of the ratio between the number of correct classifications of a given class on the total number of times that the algorithm classifies it. While the recall metric expresses the sensitivity of the model and is represented by the ratio between the correct classifications for a given class, on the total of cases in which the event occurs. Based on these definitions, we can see that for the positive class, the conventional XGBoost model is not sensitive and precise in recognizing instances that present women who are at risk of cervical cancer. In predicting cervical cancer risk the precision and the recall values obtained for the positive class are 0.17 and 0.12 (17% and 12%) respectively. For the negative class, an accuracy of 0.93, precision of 0.96, and recall of 0.97 (93%, 96%, and 97%) were obtained. These results showed that the existing system was able to identify only 21% of the positive class which represents women at risk of cervical cancer. At the same time, for the proposed model, an XGboost ensemble classifier combined with SmoteTomek hybrid resampling technique, the model is sensitive and precise in recognizing instances that present women who are at risk of developing cervical cancer with accuracy, precision, and recall of 0.99, 0.89, 1.00 (99%, 89%, and100%) respectively. This means that on average, out of 172 cases used to evaluate the performance of the proposed model, the model correctly classifies all the patients that are at the risk of developing cervical cancer. The proposed method for the classification of cervical cancer risk is robust and provides significant performances. With the results obtained from this study, the proposed system served as a significant tool to support clinical decisions in the prevention of cervical cancer. Figure 4.13 shows the confusion matrix of the proposed system between 0 and 1, representative of the classification capacity of the proposed system. It showed that out of 164 negative cases, the proposed system was able to identify 163 negative cases (True Negative = 163), and misclassified 1 case as positive (False Positive = 1), this showed that 99% of the negative cases were correctly identified. Likewise, out of 8 positive cases, the proposed system was able to identify all the 8 positive cases (True Positive = 8), and no positive cases were misclassified (False Negative = 0). The results showed that the proposed system was able to identify 99% of the negative cases and 100% of the positive cases.

7. CONCLUSION

In the field of medical healthcare, early detection of cervical cancer remains challenging due to its asymptomatic nature. Based on the developed classification model, and respective results, it can be concluded that this study has successfully achieved all the objectives set in section 1.3 for the problem of detecting cervical cancer. This study proposes a hybrid approach using a composite classifier that integrates diverse algorithms and techniques in a pipeline to predict the risk of developing cervical cancer. The hybrid approach leverages SMOTETomek for data resampling, XGboost for feature selection and classification, achieving better accuracy and optimal performance than the existing system. The proposed model performed better than the existing model on all the evaluation metrics by correctly identifying 100% of the total positive cases. (Recall = 1.0, i.e., 100%) with an accuracy score of 0.99, i.e., 99%, which indicates that the model can correctly identify 100% of women who are at risk of developing cervical cancer. The proposed model performed well in predicting women at risk of developing cervical cancer based on demographic and behavioral data.

REFERENCES

[1] Mahajan, P., Uddin, S., Hajati, F., & Moni, M. A. (2023). Ensemble Learning for Disease Prediction: A Review. In *Healthcare* (Vol. 11, No. 12, p. 1808). MDPI.

[2] Dong, X., Yu, Z., Cao, W., Shi, Y., & Ma, Q. (2020). A survey on ensemble learning. *Frontiers of Computer Science*, *14*(2), 241-258.

[3] Jason, B (2021). A Gentle Introduction to Ensemble Learning Algorithms in Ensemble Learning

[4] Derrick, M. (2021). A Comprehensive Guide to Ensemble Learning: What Exactly Do You Need to Know. https://neptune.ai/blog/ensemble-learning-guide.

[5] Das, S., and Biswas, D. (2019). Prediction of breast cancer using ensemble learning. *5th International Conference on Advances in Electrical Engineering, ICAEE*, 804–808.

[6] World Health Organization. (2023). WHO recommendations on self-care interventions: human papillomavirus (HPV) self-sampling as part of cervical cancer screening (No. WHO/ SRH/20.12). World Health Organization.

[7] Jagwani, G. (2019). Identifying the Patients at Risk of Stroke Using Anomaly Detection Based Classification Approach. *Doctoral dissertation, Dublin, National College of Ireland.*

[8] Jiang, J. R., & Chen, Y. T. (2022). Industrial control system anomaly detection and classification based on network traffic. *IEEE Access*, *10*, 41874-41888.

[9] Fang, Z. G., Yang, S. Q., Lv, C. X., An, S. Y., & Wu, W. (2022). Application of a data-driven XGBoost model for the prediction of COVID-19 in the USA: a time-series study. *BMJ open*, *12*(7), e056685.

[10] Kanitkar A., Joshi V., Karwa Y., Gindi S., & Kale, G. (*2019).* Comparison of Machine Learning Algorithms for Cervical Abnormality Detection. *International Conference on Contemporary Computing,* 1-6. doi: 10.1109/IC3.2019.8844926

[11] Rahman, S. et al. (2020). Performance analysis of boosting classifiers in recognizing activities of daily living. *Int J Env Res Public Health 17(3):1082*

[12] Lu, J., Song, E., Ghoneim, A., & Alrashoud, M. (2020). Machine learning for assisting cervical cancer diagnosis: An ensemble approach. *Future Generation Computer Systems,* doi: 10.1016/j.future.2019.12.033

[13] Kumar, M., Singhal, S., Shekhar, S., Sharma, B., & Srivastava, G. (2022). Optimized Stacking Ensemble Learning Model for Breast Cancer Detection and Classification Using Machine Learning. *Sustainability*, *14*(21), 13998.

[14] Srivastava, G., Singhal, S., Shekhar, S., Sharma, B., & Kumar, M. (2022). Optimized Stacking Ensemble Learning Model for Breast Cancer Detection and Classification Using Machine Learning. *Sustainability*, *14*(21), 13998.

[15] Win, K. P., Kitjaidure, Y., Paing, M. P., & Hamamoto, K. (2019). Cervical cancer detection and classification from pap smear images. In *Proceedings of the 2019 4th International Conference on Biomedical Imaging, Signal Processing* (pp. 47-54).

[16] Asadi, F., Salehnasab, C., & Ajori, L. (2020). Supervised algorithms of machine learning for the prediction of cervical cancer. *Journal of biomedical physics & engineering*, *10*(4), 513.

[17] Rahman, S. et al. (2020). Performance analysis of boosting classifiers in recognizing activities of daily living. *Int J Env Res Public Health 17(3):1082*

[18] Fang, Z., Wang, Y., Peng, L., & Hong, H. (2021). A comparative study of heterogeneous ensemble-learning techniques for landslide susceptibility mapping. *International Journal of Geographical Information Science*, *35*(2), 321-347.