ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS | Volume VIII Issue VI June 2023

Prediction of Cervical Cancer Using Boosting Techniques.

Ramoni Tirimisiyu Amosa[1*], Adekiigbe Adebanjo[1], Olawale Olaniran Kayode[2], Fabiyi Aderanti Alifat[1], Olorunlomerue Adam Biodun[1], Oluwatosin Adefunke Oluwatobi[1], Adejola Aanu Adeyinka & Fakiyesi Favour[1]

[1]Department of Computer Science, School of Applied Sciences, Federal Polytechnic Ede, Osun State, Nigeria.

[2]Department of Science Laboratory Technology, School of Applied Sciences, Federal Polytechnic Ede, Osun State, Nigeria.

**Corresponding Author*

DOI: **https://doi.org/10.51584/IJRIAS.2023.8605**

Received: 06 May 2023; Revised: 02 June 2023; Accepted: 06 June 2023; Published: 03 July 2023

Abstract: Cancer of the cervix, commonly called cervical cancer, is a type of cancer that develops in the cells of the cervix, which is the lower portion of the uterus that attaches to the vagina. It hardly shown symptoms in its early stage. To detect the disease, regular is required, however larger population of women not aware of this approach while many shy away and refuse to take the test. Hence cervical cancer spread like wild fire among women and being the most common cause of cancer disease it result to untimely death among women in our society today. In this research, the performance of a few sophisticated ensemble models, such as Bagging Classifier and Adaptive Boosting (AdaBoost) Classifier, is shown for the purpose of predicting a diagnosis of cervical cancer based on recorded cancer risk factors and target variables. Accuracy, sensitivity, and specificity were the measures that were used in the evaluation of the models. Python library was adopted for the classification and the cervical cancer dataset used for the experiment was acquired from UCI (University of California at Irvine), the classification was carried using voting approach by combining three classifiers: Decision Tree (DT), K-N Neighbour(KNN) and Random Forest (RF). The results indicated that the proposed model was highly accurate in predicting the risk of cervical cancer, with 119 instances classified as 'class zero' and only three instances classified as 'class one' based on the predictions.

Keywords: Accuracy, Cervical Cancer, Experiment, Machine Learning (ML), Model, AdaBoost Classifier, Bagging Classifier.

I. Introduction

Cancer is a perilous illness characterized by the abnormal growth and division of cells, which occurs in a disorderly and uncontrolled manner, disregarding the normal regulations of cell division. (Hejmadi, 2014).Cancer is caused by the abnormal growth of cells in a specific part of the body, which multiply uncontrollably. It is a significant cause of death worldwide, with approximately 8.2 million people dying from cancer each year, accounting for 13% of all deaths globally. Screening services for cancer are not widely available, particularly in underdeveloped countries, with only 26% reporting screening services available for the public in 2017. While 90% of developed countries offer treatment services, less than 26% of low-income countries have access to such services. It is estimated that the number of cancer cases will increase to 22 million by 2030. (Ferlay et al., 2021).Lung and breast cancer are responsible for millions of premature deaths among women, but cervical cancer is considered to be the most perilous because it affects only females. The female reproductive system includes the cervix, uterus, vagina, and ovaries, and cervical cancer develops in the opening of the uterus from the vagina, which is known as the cervix. HPV, often known as the human papillomavirus, is the virus responsible for the development of cervical cancer (Shah &Itzkowitz, 2022)

Cervical cancer is more prevalent in low- and middle-income countries, as reported by Siegel (2018). Screening is crucial for detecting cervical cancer. An effective screening test should be minimally invasive, easy to perform, acceptable to the patient, affordable, and capable of diagnosing the disease in its early stages when treatment is most effective (Cohen et al, 2019). The four main screening methods for the disease are; cervical cytology, biopsy, Schiller test, and Hinselmann test (Desaiet. al., 2021). Researchers have used decision tree for the prediction and some other approaches (Fahad, 2019; Keller et al., 2019; Hejmadi, 2014; Bray et al., 2018). However, previous results shows that further investigation on how to improve the result and accuracy of the prediction is imperative.

II. Literature Review

As a form of artificial intelligence, machine learning (ML) includes teaching algorithms to recognize patterns in data and applying those patterns to forecast outcomes for fresh data. Decision trees, logistic regression, artificial neural networks (ANNs), support vector machines (SVMs), and random forests are some of the most widely used machine learning (ML) algorithms for cervical cancer prediction (Siegel et al., (2018).

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS | Volume VIII Issue VI June 2023

Machine learning algorithms have been utilized in several research to forecast cervical cancer. For instance, an ANNwas employed to forecast cervical cancer based on gene expression data in a study by Robilotti et al. (2020). The study showed that ANNs are useful for predicting cervical cancer, with an accuracy of 96.9% and a sensitivity of 97.3%, One of the significant challenges here is the limited availability of high-quality data. The data used for training and testing the models are often incomplete, imbalanced, or biased, which can affect the accuracy and generalizability of the models. Additionally, there is a lack of standardized data collection methods, which can lead to inconsistencies in the data.

SVMwas employed in a study by Galluzzi et al. (2017) to forecast cervical cancer based on clinical data. The study showed that SVMs are useful for predicting cervical cancer, with an accuracy of 89.1% and a sensitivity of 88.8%. Similar to this, Zheng et al. (2020) employed an SVM to predict cervical cancer based on clinical data, and they were able to do so with 92.6% accuracy and 95.2% sensitivity.In another study by Chen et al. (2021), a random forest algorithm was used to predict cervical cancer based on medical imaging data. The study reported an accuracy of 88.8% and a sensitivity of 86.9%, demonstrating the effectiveness of random forests for cervical cancer prediction. Similarly, a decision tree was used by Desai et al. (2021) to predict cervical cancer based on clinical data, achieving an accuracy of 83.6% and a sensitivity of 91.3%.

Overall, these investigations demonstrate the efficiency of ML techniques for forecasting cervical cancer. ML algorithms have shown promising results when applied to diverse types of data, including gene expression data, clinical data, and medical imaging data (Desantis et al., 2019). ML techniques can identify important features that are associated with the probability of developing cervical cancer, and improve the accuracy of predictions by combining these features (Ferlay et al., 2021), the major drawback here is overfitting and generalizability. Overfitting is a common challenge in machine learning where a model becomes too familiar with the training data and may fail to accurately predict new data. This can result in poor performance and reduced accuracy when the model is tested on new data.Boosting is a machine learning ensemble technique that uses multiple weak classifiers to construct a powerful classifier (Hejmadi, 2014).The weak classifiers are iteratively trained on subsets of the training data, with each subsequent classifier focusing on the samples that were misclassified by the previous classifier. This iterative process continues until a stopping criterion is met, typically when a predefined number of classifiers have been trained.

In the case of cervical cancer, boosting techniques can be used to analyze different types of data, including gene expression data, clinical data, and medical imaging data. For instance, gene expression data can provide insights into the molecular mechanisms of cervical cancer and help identify important biomarkers for early detection. Clinical data, on the other hand, can include demographic information, medical history, and other relevant factors that can be used to predict the risk of developing cervical cancer. Medical imaging data, such as computed tomography (CT) scans and magnetic resonance imaging (MRI), can be used to detect abnormalities in the cervix and surrounding tissues.Boosting algorithms, such as AdaBoost, Gradient Boosting, and XGBoost, have been shown to be effecient for the prediction of cervical cancer based on these different types of data. These algorithms can identify important features, such as gene expression patterns, clinical variables, or imaging characteristics, that are associated with the risk of developing cervical cancer. By combining these features into a single prediction model, boosting algorithms can improve the accuracy of cervical cancer prediction.In conclusion, the use of boosting techniques for predictingng cervical cancer is a promising area of research that has the potential to improve early detection and treatment outcomes. With further development and validation, these methods may become an important tool for clinicians and researchers working in the field of cervical cancer. Other drawback of previous researches include feature selection and class imbalance. Cervical cancer prediction requires the identification of relevant features that can be used to build predictive models. However, there is a wide range of potential features, including demographic, clinical, and genomic data, and selecting the most relevant ones is a challenging task. Class imbalance occurs where the number of samples in the minority class (i.e., patients with cervical cancer) is significantly lower than the majority class (i.e., patients without cervical cancer). Class imbalance can affect the performance of the models, leading to biased results. Researchers can address this challenge by using appropriate sampling techniques, such as oversampling, undersampling, or synthetic data generation.

III. Methodology

This research built and trained a model to predict the risk of cervical cancer in 858 patients using data obtained from a hospital in Caracas, Venezuela. The dataset contained medical records, demographic data, and habits of the patients, with factors such as excessive sexual activity, HPV, oral contraceptives, large family size, and smoking being identified as major risk factors for cervical cancer. The model used an eggy boost algorithm and input features such as age, number of pregnancies, smoking habits, and medical history to predict whether the patient has cancer or has a high risk of cancer. The gradient boost method was used to develop the deep learning framework model, and the critical factors for predicting cervical cancer risk were visualized using the dataset. The study concluded by summarizing and visualizing the model's performance, which was found to be better than other studies in the same domain due to the use of the gradient boost algorithm.

The research methodology proposed in this study is divided into four phases: research dataset, data preprocessing, predictive model selection (PMS), and training method. An architectural diagram of the proposed research is provided in Figure 1, which

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS | Volume VIII Issue VI June 2023

illustrates the essential tasks that are performed in each phase. The research dataset is described in the Research Dataset section, while the Data Preprocessing section explains how noise is removed from the dataset to make it suitable for machine learning. The type of predictive model selected for cervical cancer prediction is discussed in the PMS portion, while the training methods section covers the requirements for model training. The Python programming language was used to design the platform for an overall pipeline of cervical cancer prediction. The algorithm implemented in this research is well-suited for clinical use, particularly for categorizing negative and positive cervical cancer diagnoses. Several algorithms, including decision tree, logistic regression, K-nearest neighbours (KNN), adaptive boosting, Gradient boosting, random forest, and Gaussian NB, can be used to diagnose cervical cancer. The sequence and consequences of these algorithms are presented in the following sections.

Table 1: Dataset Description

1	Age	Number o	First sexua	Num of pr	Smokes	Smokes (y	Smokes (p	Hormonal	Hormonal	IUD	IUD (years	STDs	STDs (num	STDs:cond	STDs:cervi	STDs:vagir	STDs:vulvc	STDs:syph	STDs:pelvi	ST
2	18	4	15	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
3	15	1	14	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
4	34	1	?	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
5	52	5	16	4	1	37	37	1	3	0	0	0	0	0	0	0	0	0	0	
6	46	3	21	4	0	0	0	1	15	0	0	0	0	0	0	0	0	0	0	
7	42	3	23	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
8	51	3	17	6	1	34	3.4	0	0	1	7	0	0	0	0	0	0	0	0	
9	26	1	26	3	0	0	0	1	2	1	7	0	0	0	0	0	0	0	0	
10	45	1	20	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
11	44	3	15	?	1	1.266973	2.8	0	0	?	?	0	0	0	0	0	0	0	0	
12	44	3	26	4	0	0	0	1	2	0	0	0	0	0	0	0	0	0	0	
13	27	1	17	3	0	0	0	1	8	0	0	0	0	0	0	0	0	0	0	
14	45	4	14	6	0	0	0	1	10	1	5	0	0	0	0	0	0	0	0	
15	44	2	25	2	0	0	0	1	5	0	0	0	0	0	0	0	0	0	0	
16	43	2	18	5	0	0	0	0	0	1	8	0	0	0	0	0	0	0	0	
17	40	3	18	2	0	0	0	1	15	0	0	0	0	0	0	0	0	0	0	
18	41	4	21	3	0	0	0	1	0.25	0	0	0	0	0	0	0	0	0	0	
19	43	3	15	8	0	0	0	1	3	0	0	0	0	0	0	0	0	0	0	
20	42	2	20	?	0	0	0	1	7	1	6	1	2	1	0	0	1	0	0	
21	40	2	27	?	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	
22	43	2	18	4	0	0	0	1	15	0	0	0	0	0	0	0	0	0	0	
23	41	3	17	4	0	0	0	1	10	0	0	1	1	0	0	0	0	1	0	

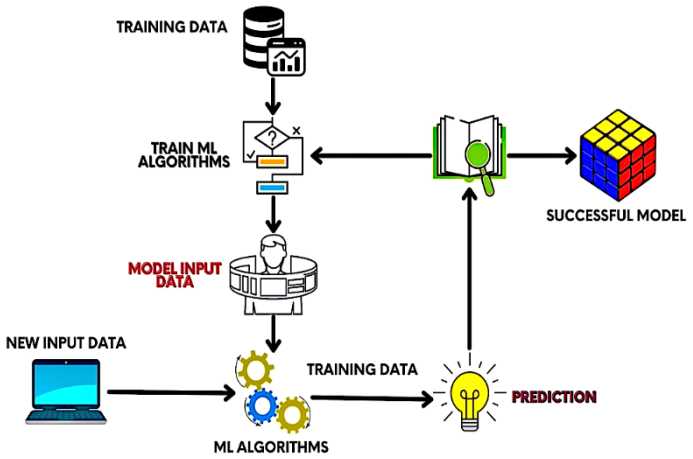

Figure 1: Research Framework

Figure 1 illustrates the proposed machine learning (ML) model. The model training starts by feeding the training data into the system. Then, ML algorithms are utilized, followed by the application of model input data and new input data to train the architecture effectively. Lastly, prediction is carried out on the newly collected data.

3.1. Research Dataset

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS | Volume VIII Issue VI June 2023

The dataset "Cervical Cancer Risk Factors for Biopsy" was acquired from UCI repository, it contains information about 858 records and 32 attributes which include age, IUD, smoking, STDs, and so on. Table 1 shows the structure of the dataset.

For the analysis of the dataset and implementation of the aforementioned algorithm, Weka machine learning tool was used. Weka is a well-known open-source machine learning software package that offers a combination of algorithms and tools for data preprocessing, classification, regression, clustering, association rules mining, and visualisation. Weka was developed by the Apache Software Foundation. Java is the programming language that was used in its creation at the University of Waikato, which is located in New Zealand. Weka is an acronym that stands for "Waikato Environment for Knowledge Analysis," and it has amassed a large amount of popularity in the field of machine learning and data mining among researchers, students, and practitioners. It has a graphical interface that is straightforward and easy to navigate, making it suitable for users with a wide range of experience levels. In general, Weka is a tool that is useful for a wide variety of machine learning and data mining jobs due to its versatility and capability. Because of its user-friendliness, broad algorithm collection, and adaptability, it is a tool that is frequently selected by novice as well as seasoned professionals working in the relevant industry. Figure 2 shows the details of the dataset in Weka interface.

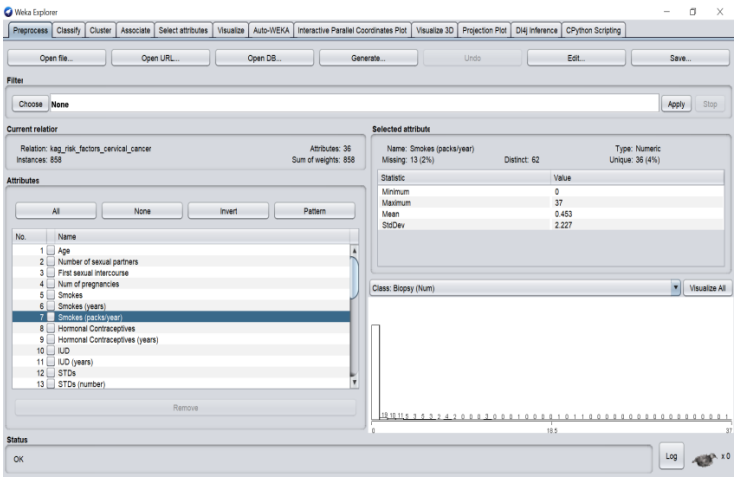

Figure 2: Dataset Information

3.2Data Preprocessing

The process of preparing the data for machine learning involves three main steps which include the cleaning of data, transformation and reduction. Data cleaning is necessary to remove impurities such as noise, outliers, redundant, and missing data, which can negatively affect the analysis. Data transformation techniques were also used to prepare the data for analysis, the major operation here is the normalisation of data and attribute selection. Data reduction techniques were employed to improve storage efficiency and reduce the dimensionality of the data. This was done to overcome difficulties in analyzing large datasets. Overfitting in the model was mitigated by Principal component analysis (PCA).

Step 1: The technique of ignoring instances and attributes with a high ratio of missing values was applied to ensure data consistency. This method was found to be effective, given that the dataset used had several instances and attributes with missing values.

Step 2: The dataset contained some attributes with no value(s) such as "number of pregnancies" and "hormonal contraceptive use". The missing values in the dataset were denoted by a "?" symbol. To address this issue, the missing values were replaced with the median values of their respective class.

$$Median: X = Sort(x), Median = X_{\frac{n}{2}}(Half\ below, Half\ above)$$

Step 3: Identifying outliers in the data was another crucial step. An outlier is a data point that significantly deviates from the rest of the data points. In this research, the attributes of age and number of partners were found to contain outliers. To address this issue, lower and upper threshold limits were defined, and the outliers were replaced with the median value.

Step 4: Normalization is a technique used in data preprocessing to scale the values of different attributes. There are various methods for normalization, such as Min-Max, Z-score, and decimal scaling normalization. In this research, decimal scaling normalization was used,

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS | Volume VIII Issue VI June 2023

3.3. Predictive Model Selection (PMS)

The PredictiveModelSelection utilized a variety of classification algorithms including SVM, DT, RF, logistic regression (LR), gradient boosting (GB), XGBoost, adaptive boosting (AB), and Knearest neighbor (KNN). Among these algorithms, some have shown satisfactory accuracy on the research dataset. This section will focus on explaining the theoretical concepts of these algorithms in the following subsections.

3.3.1. Decision Tree (DT)

The Classification and Regression Tree (CART) algorithm, also known as the Decision Tree (DT), is capable of solving both classification and regression problems. The DT is structured like a tree with branches, hence its name. The root node of the DT is similar to the base of a tree, from which branches spread out based on different decision conditions. These decision nodes are followed by leaf nodes where the final decision is made.

3.3.2. Random Forest (RF)

Ensemble learning is a technique that combines multiple learners to enhance model performance. Random Forest (RF) is one such ensemble learning method that uses a bagging approach to reduce the impact of outliers and can handle both categorical and continuous data. Scaling of data is not required for this algorithm, but a higher number of learners may require

more computational resources for complex models. In RF, the decision is made through voting, which makes it an ensemble learning algorithm. It consists of many trees, just like a forest, and the final decision is based on the majority of decisions made by the trees.

3.3.3. Gaussian Naïve Bayes

Naive Bayes is a machine learning algorithm that assumes that the features of a dataset are independent of each other. This assumption allows the algorithm to use class-specific covariance matrices, similar to the Quadratic Discriminant Analysis (QDA) algorithm, but with diagonal covariance matrices. This assumption is made because Naive Bayes assumes that each feature has no correlation with other features. In a training dataset with N input variables x and corresponding target variables t, Naive Bayes assumes that the class-conditional densities are normally distributed, according to Robert (2023).

3.3.4. K Nearest Neighbour (KNN)

K-Nearest Neighbour (K-NN) is a Supervised Learning algorithm and one of the simplest Machine Learning algorithms available. It operates on the assumption that new data is similar to the available data, and it classifies new data based on the similarity between the new data and available data. K-NN stores all the available data and classifies new data based on similarity. This allows new data to be easily classified into the most appropriate category by using K-NN. The algorithm can be used for both Regression and Classification problems, although it is mainly used for Classification. K-NN is a non-parametric algorithm and does not make any assumptions about underlying data. It is also known as a lazy learner algorithm because it does not learn from the training set immediately, but instead stores the dataset and performs an action on it when it is classified. During training, K-NN simply stores the dataset and when new data is received, it classifies it into a category that is similar to the new data.(Apalla et al, 2017).

3.4. Boosting Selection

Gradient boosting and adaptive boosting are two popular ensemble learning algorithms that differ in their approach to reducing error. In adaptive boosting, the weights of incorrectly predicted samples are updated to gradually reduce the error. In contrast, gradient boosting optimizes the loss function and reduces the error by optimizing each individual loss. To achieve this, the weak learner model is updated iteratively to improve the performance of subsequent weak learners. Gradient boosting consists of three essential components, including the weak learner, loss function optimization, and additive model.

3.4.1. Gradient Boosting

The gradient boosting technique makes use of the sequential ensemble learning approach, in which less capable learners are made to grow gradually better than their predecessors through the application of loss optimisation. This indicates that the second underachieving student is better than the first, and that the third underachieving student is better than the second, and so on. As a consequence of this, the amount of error in the model will reduce as weak learners become more proficient, which will lead to the model becoming more robust. The gradient boosting approach performs exceptionally well when used to problems of the regression variety.

3.4.2 Adaptive Boosting (AB)

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS | Volume VIII Issue VI June 2023

Adaptive boosting is a technique that combines multiple weak learners to create a powerful learner. In this approach, all weak learners use the same input, which is often referred to as a training set, and each training data point is given equal importance. The task of correcting faulty predictions made by the first weak learner is then given to the next weak learner, who is given additional weight on the predictions made by the first weak learner. This continues until all of the predictions made by the first weak learner have been corrected. This process is repeated for subsequent weak learners, with the errors made by previous weak learners being passed on to the next with increasing weight. This continues until the number of inaccurate predictions is reduced to an acceptable level. Ultimately, the combined efforts of all weak learners produce a powerful learner with reduced prediction errors. (Hall et al., 2019).

IV. Result and Discussion

Five classifiers were initially implemented but two distinct classifier techniques: DT, and RF were implemented to build the model. The dataset was found to be biased during the initial analysis, so k-fold cross-validation was performed. The feature selection process has been completed, and the features that have been chosen are used for prediction. The data is split 20-80 for Decision Tree and Random Forest. Parameter tuning is carried out get the good predictions with the highest evaluation points. Figure 3 Shows the numerical features of the dataset was analysed to show the number of entities, mean, standard deviation as well as the minimum and maximum values and percentage difference in the dataset. This was done to highlight the similarity of attributes in the dataset.

	Age	No_pregnancies	No_of_sex_partner	STDs_No_of_diagnosis	Hormonal_Contraceptives_years	STDs_number
count	838.000000	838.000000	838.000000	838.000000	838.000000	838.000000
mean	26.812649	2.273667	2.506563	0.084726	2.379033	0.151551
std	8.529209	1.441692	1.586688	0.295293	3.869996	0.521638
min	13.000000	0.000000	1.000000	0.000000	0.000000	0.000000
25%	20.000000	1.000000	2.000000	0.000000	0.000000	0.000000
50%	25.000000	2.000000	2.000000	0.000000	0.580000	0.000000
75%	32.000000	3.000000	3.000000	0.000000	3.000000	0.000000
max	84.000000	11.000000	28.000000	3.000000	30.000000	4.000000

Figure 3: Numerical feature of the dataset

Figure 4 shows the initial result of each classification model before boosting techniques was applied.

	Model	Train_Score	Test_accuracy	f1score	recall	precision	roc_auc
0	LogisticRegression	0.972696	0.952381	0.571429	0.533333	0.615385	0.756118
1	Decision Tree	1.000000	0.932540	0.564103	0.733333	0.458333	0.839241
2	Random Forest	0.996587	0.944444	0.500000	0.466667	0.538462	0.720675
3	GaussianNB	0.146758	0.095238	0.116279	1.000000	0.061728	0.518987
4	KNN	0.950512	0.936508	0.333333	0.266667	0.444444	0.622785

Figure 4: Result Data Frame

4.1. Evaluation Metrics

The following are the metrics used for evaluation:

- Accuracy: Accuracy is a measure that calculates the proportion of observations that were classified correctly to the total number of observation. It is a useful metric when dealing with datasets that are symmetric, meaning that the number of false positives and false negatives are nearly equal.
- Precision: Precision is a metric that measures the proportion of correctly predicted positive observations to the total number of predicted positive observations.
- Recall (Sensitivity) :Recall, also known as sensitivity, is a performance metric that measures the ratio of correctly predicted positive observations to all the observations in the actual positive class- yes(1).
- F1 Score:The F1 score is a measure that combines both precision and recall into a single score. It is the harmonic mean of precision and recall, and provides a way to balance the two metrics.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS | Volume VIII Issue VI June 2023

- ROC score: The ROC score is a measure of the ability of the model to distinguish between the target classes.
- To ensure accurate detection of cervical cancer patients, the recall score is given higher importance in this sensitive medical data. Therefore, "Decision Tree" and "Random Forest" models are chosen as the base models due to their higher recall and roc_auc scores.

The reason recall is given higher importance is that it is crucial to predict actual cancer patients as cancer patients accurately. False negatives, where a cancer patient is predicted as healthy, can have dangerous consequences and cause chaos in the patient's life.

Note: The base models are built using the entire features with default hyper parameters and before oversampling the data.

4.2. Feature Selection

The feature selection technique used was Recursive Feature Elimination (RFE), which involves iteratively removing features, training a model with the remaining features, and evaluating the model's accuracy. RFE helps to determine the subset of features that are most relevant for predicting the target variable or class.Used RFE on Decision Tree and Random Forest separately and found the best features for both the models individually. The features are chosen based on recall score i.e., which set of features gave the gigher recall score.

4.3. Result Analysis

Feature selection is performed to assist us in selecting the features that will improve prediction. The random forest algorithm generates a decision tree with multiple levels based on the training data set.

Accuracy is one of the measurements used to evaluate the models used in classification. In simple terms, accuracy measures the proportion of correct predictions made by a model, as illustrated in Figure 4.

	Model	Train_Score	Test_accuracy	f1score	recall	precision	roc_auc
1	Decision Tree After Sampling	1.000000	0.948413	0.666667	0.866667	0.541667	0.910127
2	Random Forest After Sampling	0.999086	0.944444	0.611111	0.733333	0.523810	0.845570
3	Decision Tree after Feature Selection	1.000000	0.988095	0.933333	1.000000	0.875000	0.993506
4	Random Forest after Feature Selection	1.000000	0.984127	0.900000	0.857143	0.947368	0.926407
5	Decision Tree after Hyperparameter Tuning	0.982633	0.960317	0.800000	0.952381	0.689655	0.956710
6	Random Forest After Hyperparameter Tuning	0.979890	0.964286	0.816327	0.952381	0.714286	0.958874
7	Bagged Decision Tree with Hyperparameter	0.985375	0.964286	0.800000	0.857143	0.750000	0.915584
8	Decision Tree ADA Boost with Hyperparameter	1.000000	0.992063	0.952381	0.952381	0.952381	0.974026
9	Gradient Boost	0.982633	0.980159	0.888889	0.952381	0.833333	0.967532

Figure 5: Model Accuracy

- Bagging: There is a no improvement in recall score.
- Ada Boost: ADA boost may not be the optimum way as there is overfitting.
- Gradient Boost: Gave the best recall score of 95.2 %

Finally, after applying all optimization techniques, an increase in recall score from 53.3 % to 95.2 % was achieved.

Final Model: Gradient boosting model has been chosen as the final model as it gave the optimal accuracy score, roc score, f1 score and recall score compared to all the other models, shown in figure 5.The model looks superior while considering the overall evaluation metrics.Therefore, it was finalized as the final model.

V.Conclusion

The focus of this research was to predict the risk of cervical cancer using various predictive models and performance metrics such as precision, recall, f1-score, and support. The study utilized a deep learning model implemented in Python with relevant packages and libraries, and the cervical cancer dataset was analyzed through data profiling, standardization, and visualization. The dataset included attributes related to major risk factors of cervical cancer such as smoking, STDs, STD, AIDS, first sexual intercourse, and cytology. The results indicated that the proposed model was highly accurate in predicting the risk of cervical cancer, with 119 instances classified as 'class zero' and only three instances classified as 'class one' based on the predictions.

References

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS | Volume VIII Issue VI June 2023

1. Apalla, Z., Nashan, D., Weller, R. B., & Castellsagué, X. (2017). Skin cancer: epidemiology, disease burden, pathophysiology, diagnosis, and therapeutic approaches. Dermatology and therapy, 7, 5-19.
2. Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., & Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians, 68(6), 394-424.
3. Cohen, P. A., Jhingran, A., Oaknin, A., & Denny, L. (2019). Cervical cancer. The Lancet, 393(10167), 169-182.
4. Desai, A., Gupta, R., Advani, S., Ouellette, L., Kuderer, N. M., Lyman, G. H., & Li, A. (2021). Mortality in hospitalized patients with cancer and coronavirus disease 2019: a systematic review and meta-analysis of cohort studies. Cancer, 127(9), 1459-1468.
5. De Santis, C. E., Ma, J., Gaudet, M. M., Newman, L. A., Miller, K. D., Goding Sauer, A., ... & Siegel, R. L. (2019). Breast cancer statistics, 2019. CA: a cancer journal for clinicians, 69(6), 438-451.
6. Fahad Ullah, M. (2019). Breast cancer: current perspectives on the disease status. Breast Cancer Metastasis and Drug Resistance: Challenges and Progress, 51-64.
7. Ferlay, J., Colombet, M., Soerjomataram, I., Parkin, D. M., Piñeros, M., Znaor, A., & Bray, F. (2021). Cancer statistics for the year 2020: An overview. International journal of cancer, 149(4), 778-789.
8. Galluzzi, L., Buqué, A., Kepp, O., Zitvogel, L., & Kroemer, G. (2017). Immunogenic cell death in cancer and infectious disease. Nature Reviews Immunology, 17(2), 97-111.
9. Hejmadi, M. (2014). Introduction to cancer biology. Bookboon.
10. Keller, D. S., Windsor, A., Cohen, R., & Chand, M. (2019). Colorectal cancer in inflammatory bowel disease: review of the evidence. Techniques in coloproctology, 23, 3-13.
11. Robilotti, E. V., Babady, N. E., Mead, P. A., Rolling, T., Perez-Johnston, R., Bernardes, M., ... & Kamboj, M. (2020). Determinants of COVID-19 disease severity in patients with cancer. Nature medicine, 26(8), 1218-1223.
12. Saini, K. S., Tagliamento, M., Lambertini, M., McNally, R., Romano, M., Leone, M., ... & de Azambuja, E. (2020). Mortality in patients with cancer and coronavirus disease 2019: a systematic review and pooled analysis of 52 studies. European Journal of Cancer, 139, 43-50.
13. Shah, S. C., & Itzkowitz, S. H. (2022). Colorectal cancer in inflammatory bowel disease: Mechanisms and management. Gastroenterology, 162(3), 715-730.
14. Siegel, R. L., Miller, K. D., & Jemal, A. (2018). Cancer statistics, 2018. CA: a cancer journal for clinicians, 68(1), 7-30.