Efficient Staging of Cervical Cancer Using Machine Learning Approach

BY

Shorove Tajmen
ID: 191-15-2370
AND

Al-Amin Dhaly
ID:191-15-2539

This Report is Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

Aliza Ahmed Khan
Senior Lecturer
Department of CSE
Daffodil International University

Co-Supervised By

Al Amin Biswas
Lecturer
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

FEBRUARY 2023

APPROVAL

This Project titled **"Efficient Staging of Cervical Cancer Using Machine Learning Approach"**, submitted by Shorove Tajmen, ID No: 191-15-2370 and Al-Amin Dhaly, ID No: 191-15-2539 to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 01 February 2023.

BOARD OF EXAMINERS

Chairman

4/2/23

Dr. Touhid Bhuiyan
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Tania Khat

Tania Khatun (TK)
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

lamia 01.2.23

Ms. Lamia Rukhsara (LR)
Senior Lecturer
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

External Examiner

Dr. Mohammad Shorif Uddin
Professor
Department of Computer Science and Engineering
Jahangirnagar University

DECLARATION

We hereby declare that, this project has been done by us under the supervision of **Aliza Ahmed Khan, Lecturer, Department of CSE,** Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:

Aliza Ahmed Khan
Lecturer
Department of CSE
Daffodil International University

Co-Supervised by:

31.01.2023

Al Amin Biswas
Lecturer
Department of CSE
Daffodil International University

Submitted by:

Shorove

Shorove Tajmen
ID: 191-15-2370
Department of CSE
Daffodil International University

Alamin Dhaly 4.02.2023

Al-Amin Dhaly
ID: 191-15-2539
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First we express our heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the final year project/internship successfully.

We really grateful and wish our profound our indebtedness to **Aliza Ahmed Khan**, **Lecturer**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of "*Machine Learning*" to carry out this project. Her endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism , valuable advice ,reading many inferior draft and correcting them at all stage have made it possible to complete this project.

We would like to express our heartiest gratitude to Professor Dr. Touhid Bhuiyan, Head**,** Department of CSE, for his kind help to finish our project and also to other faculty member and the staff of CSE department of Daffodil International University.

We would like to thank our entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, we must acknowledge with due respect the constant support and patients of our parents.

ABSTRACT

Cervical cancer is one of the deadliest diseases, causing a significant number of premature deaths in under-developed countries. Several risk factors are responsible for causing cervical cancer. Several organizations and individuals have proposed numerous approaches, as employing machine learning classifiers has become a very common practice in recent years. This study includes a sophisticated predictive model for classifying cervical cancer stages, as well as traditional machine learning classifiers for comparative analysis. This study used a highly imbalanced data, and missing values are present for a number of attributes. The missing value imputation technique, along with the Synthetic Minority Oversampling Technique (SMOTE), was applied to resolve the data imbalance issue. Several feature selection techniques, like Univariate Feature Selection(UFE) and Recursive Feature Elimination (RFE) were employed to determine the most important attributes for the classification outcomes. A comparison of the performance of various machine learning classifiers such as Decision Tree Classifier (DTC), Random Forest Classifier (RFC), Logistic Regression Classifier (LRC), Gaussian Naive Bayes (NBC), K Nearest Neighbors (KNNC), Gradient Boosting Classifier (GBC), AdaBoost Classifier (ABC), XGBoost Classifier (XGBC), and Support Vector Classifier (SVC) before and after the application of sampling and using feature selection methods to exhibit the effectiveness of the classifiers. In the same manner, ensemble methods like Bagging, Boosting, Stacking and Voting Classifier were employed with a view to obtaining an improved score. The application of Hyper Parameter Tuning does the job of getting the best set of parameters for classification. Thus, this work shows a marginal downfall in outcomes after the application of feature selection techniques and significant improvement in ensemble methods. RFC achieved the highest accuracy score of 99.60% after employing the feature selection technique (RFE).

Keywords: Cervical Cancer, Hyper Parameter Tuning, Recursive Feature Elimination, Ensemble Methods, Runtime Calculation

TABLE OF CONTENT

LIST OF FIGURES

LIST of TABLES

CHAPTER 1
Introduction

1.1 Overview

The cancer of cervix uteri has been the seventh most common cancer causing a significant number of deaths per year. Among the various possible difficulties women face in life, the most critical ailment is cervical cancer. Cervical cancer is mostly caused by a prolonged infection with the Human Papillomavirus (HMV)[1]. Women's unawareness about the early detection to eradicate the disease is also responsible for growing cervical cancer [2]. Cervical cancer has an extremely alarmingly high fatality rate which has made it one of the most dangerous cancer types. According to World Cancer Research Fund International, cervical cancer has affected about 604,127 women having a ratio of 13.3 women per one hundred thousand in 2020 where the mortality rate comes 7.3 women per one hundred thousand causing 341,81 deaths worldwide [3]. As HPV virus cause almost no symptoms, it is very necessary to predict early and take precaution as WHO declares the early detection and effective management can make cervical cancer treatable. Estimated 444,500 new cases are being added to the stat annually as per the research statistics of WHO [4]. More than 80% deaths are taking place in the under-developed countries and the well-developed countries are also in risk as the number of patients with cervical malignancy are increasing with the time being [5]. With the arrival of new technologies and innovative ideas, several models and ideas are proposed in recent years to predict, cure and stage cervical cancer. Medical data being accessible to the researchers has paved a new path of analyzing and innovating. Using ML techniques has been very effective way to the diagnosis processes of cervical cancer. With the advent of ideas in each of the previous works, they contain some gaps as well. Our goal is to fill the previous research gaps and building a sophisticated model which can make efficient prediction with less computational complexity and higher prediction outcome comparing to the recent works. This article goes through employing a very careful sequence of procedures stated in the following-

- The dataset needed to go through the process of missing value imputation for several attributes.
- Numerous data preprocessing techniques were employed to make smooth classification such as SMOTE, Standard Scalar, manual partitioning of categorical and numerical data.
- The traditional machine learning classifiers employed in the work for efficient staging are DTC, RFC, LRC, NBC, KNNC, GBC, ABC, XGBC and SVC
- To determine the most significant features and compare classifiers, two feature selection approaches were used: Univariate Feature Selection and Recursive Feature Elimination.
- Hyper Parameter Tuning was added to the techniques to find out best set of parameters.
- Ensemble methods to boost the classification performance were also employed where the methods include- Bagging, Boosting, Stacking and Voting Classifier.
- Evaluation study was performed to show the comparison of the classification performance with other existing works.

The embellishment of this article includes six more sections. The background study is contained in section 2. Section 3 contains the research methodology including classification algorithms, ensemble classifiers and other implemented techniques. Section 3 includes the analysis of the experimental result for comparison of scores. Section 5 and 6 contains conclusion and future possibility of expanding the work and references respectively.

1.2 Motivation

The alarming rate of getting affected to cervical cancer and the gradual increasing in the number of deaths worldwide has been a major issue in recent years. The unawareness of the women in developing countries about the severity of persistent infection of HMV in cervix uteri has made them the worst victim of getting affected and face the consequence.

According to World Health Statistics, every year 7.75 women per 100,000 face the consequence of death due to Cervical Cancer [6]. Centers for Disease Control and Prevention states, 13,000 new cases and 4,000 deaths are counted in the USA each year[7]. Many researches and diagnosis methods are taking place, yet the mortality rate is not getting decreased. Research gaps, lack of usage in optimal methods are not letting the rate to come down to zero. Being a citizen of a developing country and witnessing the disease in the surroundings motivates us to propose sophisticated approaches to generate efficient outcomes and to fill the gaps exists in previous researches.

1.3 Research Objectives

- To make early staging of cervical cancer.
- To analyze smooth and efficient solution.
- To assist in detection and diagnosis process.
- To spread awareness to people about the threat of persistent HPV virus.

1.4 Research Questions

- What approaches make the classification approach better?
- What key features the dataset contain?
- What algorithms are employed in this research?
- What preprocessing techniques are applied to make the dataset ready for efficient prediction?
- What differs this research from other existing work?

1.5 Research Layout

- Background
- Research Methodology
- Experimental Results and Discussion
- Conclusion & Future Work
- Reference

CHAPTER 2

Background Study

2.1 Related Works

Since cervical cancer being a very common disease in the recent decades, huge number of researches have taken place contributing to the medical sector continuously to initialize the diagnosis process earlier. By fusing the relief feature technique with the wrapper method of a genetic algorithm, B. Nithya et al. [8] offered an optimum classification method that generates a revised feature subset of data. The dataset used in this study was collected from SEER database which needed to go through few preprocessing techniques including missing value imputation and removing the rows containing excessive missing data. The incorporation of filter and wrapper feature selection methods were employed to obtain the optimal feature subset. The algorithms applied with 10-fold cross validation are C5.0, Random Forest and KNN. The experimental result shows scores for different stages of cervical cancer and ovarian cancer where the proposed approach gained a significant accuracy of 96.98% for cervical cancer and 97.5% accuracy for ovarian cancer where the highest accuracy of 97.96% was gained using traditional RF classifier. This work lacks the implementation of hybrid classifier to boost the score and the proposed technique was not compared with only two traditional machine learning classifiers. Jesse Jeremiah Tanimu et al. [9] developed a predictive model for generating the classification outcomes of cervical cancer. This study's data was culled from the UCI Machine Learning repository. The dataset containing missing values for several attributes and is highly imbalanced. It has gone through missing value imputation through dropping and omitting. The only ML classifier employed here is Decision tree (DT). The classification performance before and after employing Recursive Feature Elimination was observed to have an improved accuracy after employing. For resolving the imbalance issue, SMOTETomek technique was employed combining under and oversampling technique. To validate the classification result, K-fold cross validation technique was employed. The DT classifier gained the accuracy of 96% before employing feature selection technique and balancing technique which was improved to 98% after employing this approaches. Thus, the developed model has gained an improved accuracy comparing to the traditional classification approach. The comparative analysis shows the improvement after sampling and feature reducing was employed. This work also lacks the implementation of ensemble techniques for further boosting of the classification scores. Naif Al Mudawi et al. [10] presented a sharp way to predict cervical cancer. The dataset was collected from UCI Machine Learning Repository. Data preprocessing section in this work includes data cleaning, data transformation and data reduction, dimension reduction, normalization, discretization and concept hierarchy generation. The machine learning classifiers employed in this study are support vector machine (SVM), decision tree classifier (DTC), random forest (RF), logistic regression (LR), gradient boosting (GB), XGBoost, adaptive boosting (AB) and K-Nearest neighbor (KNN). The highest accuracy of 99% was gained by employing Decision Tree Classifier on the dataset. Computational complexity was calculated to assess the efficacy of the ML classifiers employed in this work. For achieving better residuals, cross validation technique was employed. Exploratory and survey data analysis was employed to make a summarized discussion on this research. No application of ensemble techniques and feature selection techniques were employed putting all focus on preprocessing technique to gain improved score. A combination of the feature selection and stacked generalization approaches was developed by Avijit Kumar Chaudhury et al. [11]. The research data utilized in this article was obtained from the UCI Machine Learning Repository. In this research, we use a three-stage hybrid feature selection strategy in combination with a stacked classification model. In the initial step of the selection process, we use a Genetic Algorithm and Logistic Regression to find a set of 12 characteristics that are highly linked with the class. In the second round of selection, five characteristics are used using a Genetic Algorithm and a Logistic Regression Architecture. Classifiers like Logistic Regression (LR), Naive Bayes (NB), Support Vector Machine (SVM), and Extra Trees (ET) are used with the aforementioned 5 characteristics in the final stage selection to produce accurate cervical cancer classifications. 10-fold cross validation technique was employed to demonstrate performance improvement of LR, NB, SVM, ET, RF & GDB classifiers reducing the features. Splitting takes place in three different partitions, 50-50, 66-34 and 80-20 to record the classification performance. With a very high accuracy, this work doesn’t contain the attempt of employing several ensemble classifiers for the further improvement of the classification performance of cervical cancer. Laboni akter et al. [12] employed several machine learning classifiers to get the classification score of cervical cancer. The dataset was collected from UCI machine learning repository. The preprocessing technique only include Min-Max scaling approach. The machine learning classifiers employed in this work are Decision Tree, Random Forest and XGBoost. The highest accuracy of 93.33% was generated for all three classifiers. This work lacks any significant or unique approach of employing preprocessing, feature selection or the application of ensemble classifiers. Surbhi Gupta et al. [13] demonstrated a new approach of employing ensemble technique to machine learning classifiers for automatic diagnosis of cervical cancer. The dataset used in this study was collected from University of California Irvine database repository. The dataset needed to go through missing value imputation and several preprocessing techniques. Missing values where handled using K-Nearest Neighbors (KNN) classifier. ROS technique was implemented to balance the data. Another preprocessing technique include Data Standardization and Dimension Reduction. Both 80-20 splitting and 10-fold cross validation technique was performed. Extremely Randomized Trees Feature Selection and Random Forest Feature Selection technique was performed to reduce the number of features to improve the classification performance. The traditional machine learning classifiers implemented in this work are K-Nearest Neighbors (KNN), Gradient Boosting Classifiers (GBC), Random Forest Classifier (RF), Multi-Layer Perception (MLP) Classifier. The ensemble techniques employed are stacking voting classifier which comes in two different approach as Majority Voting Ensemble and Weighted Voting Ensemble. Traditional RF outscored both Majority and Weighted Voting Classifiers. The highest accuracy of 99.6% score was gained employing Stacking classifier. Only two ensemble technique were employed in this study. Irfan Ullah Khan et al. [14] proposed a study for early detection of cervical cancer using feature selection technique and ensemble based machine learning classifiers. The dataset used in this study was collected from Hospital Universitario de Caracas which is also available in UCI repository. The preprocessing technique includes missing value imputation technique done by using the most frequent value technique. The imbalance issue was resolved by applying SMOTE which follows the approach of oversampling the minority class. Firefly feature selection technique was employed for the reduction of features to make the classification performance more efficient. Ensemble-based machine learning classifier like Random Forest, Extreme Gradient Boosting and AdaBoost Classifier were applied to train the model. The classification was performed on several attributes. For every distinct target attribute, XGB classifier gained the highest accuracy score for all features and selected features. This work also didn't come out of implementing traditional machine learning classifier and it lacks the application of ensemble classifiers like Bagging, Boosting, Voting and Stacking classifier. Jesse Jeremiah Tanimu et al. [15] comes up with another predictive model performance for classifying cervical cancer. The cervical cancer dataset from UCI Machine Learning Repository was used in this work. The only machine learning classifier employed in this work is Decision Tree (DT) classifier. The preprocessing section includes missing value imputation and resolving the imbalance issue. The feature selection implementation section includes filter method, wrapper method and embedded method to select the most important feature and reduce the number of features. The imbalanced class resampling techniques applied is a hybrid technique SMOTEomek which includes both Undersampling and Oversampling technique to balance the distribution. The traditional DT classifier gained the accuracy of 95.29% which was improved to 97.65% after employing RFE method and 96.47% after implementing LASSO method. The further improvement of the classification was done employing the SMOTEomek to balance the issue after the implementation of RFE feature selection method generating the highest accuracy of 98.82%.

CHAPTER 3

Research Methodology

3.1 Overview

This study goes forward implementing the dataset gathered from UCI Machine Learning Repository. The methods employed in this study are of several categories. The overall research methodology includes- dataset collection, dataset preprocessing, model fitting, feature selection, feature selection ensemble, hyper parameter tuning, computational complexity calculation and experimental result analysis. The machine learning models employed in this study are Logistic Regression Classifier (LRC), Random Forest Classifier (RFC), Decision Tree Classifier (DTC), Gaussian Naïve Bayes (NBC), KNeighbors Classifier (KNNC), Gradient Boosting Classifier (GBC), AdaBoost Classifier (ABC), XGBoost Classifier (XGBC) and Support Vector Classifier (SVC) before and after the implementation of sampling technique. In order to generate further improved score for this study, several ensemble methods like Bagging, Boosting, Voting and Stacking were employed. The best parameter set was found out employing hyper parameter tuning. Then the runtime calculation for each applied models were performed.

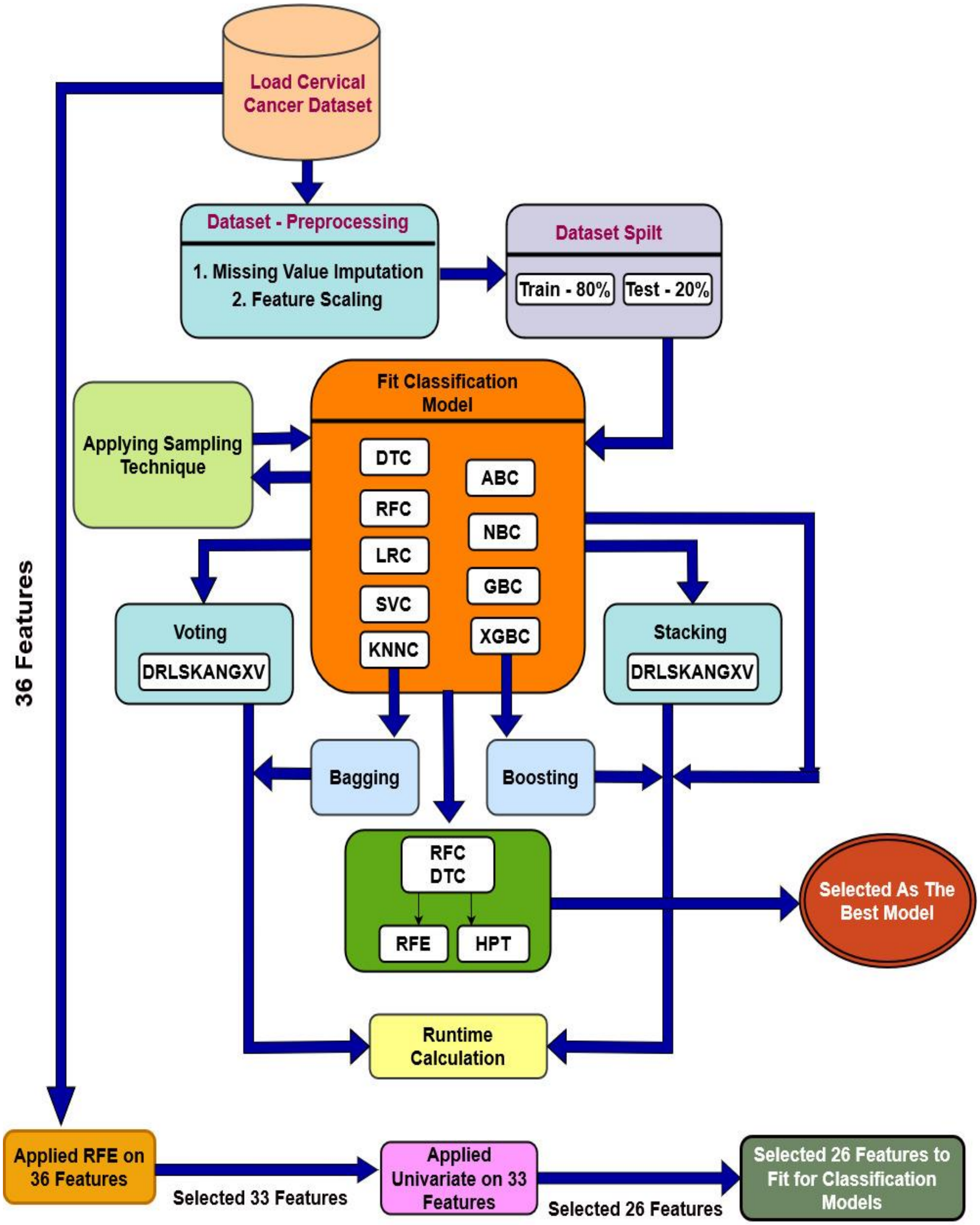


Figure 1: Overall Process Diagram

3.2 Pre-processing

3.2.1 Missing Value Imputation:

3.2.1 Missing Value Imputation: The missing value imputation technique employed went through several procedures. The column which needed to be imputed was removed from the independent column list. The rest of the independent columns that contained null values were filled with either mean or median depending on their dataset. Then the vacant column was considered as the class and the rest of the columns were considered as training set. Thus the training set was trained to predict and generate the missing values of the considered class using Decision Tree Classifier (DTC) model. This was how the missing values were imputed.

3.2.2 Handling Imbalanced Data: Handling imbalanced data indicates equalizing the distribution of class data so that the classification outcome is improved with better score. The technique employed for equalizing the class distribution is Synthetic Minority Oversampling Technique (SMOTE). It manages data by systematically adding more cases to the dataset [16]. While using the entire dataset as input, it boosts the data of minorities.

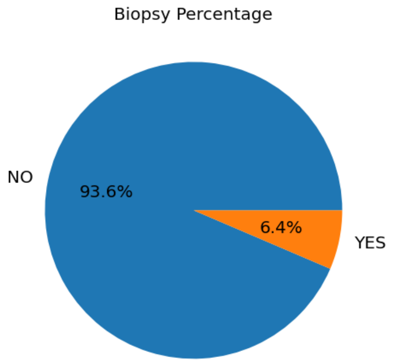


Figure 3.2.2: Pie Diagram of Class Values

3.2.3 Feature Scaling: The range of independent features in data are normalized using the feature scaling method. The technique implemented to scale the data in a definite range is standard scalar. Data must be scaled to meet a conventional normal distribution as part of standardization. A standard normal distribution is 1 with a mean of 0 and a standard deviation of 1. [17].

3.3 Correlation Between Working Attributes

The interdependence between two variables and how one variable varies in response to the change in another is determined by the correlation subplot. Higher correlations between variables suggest successful prediction of one variable from another [18]. Additionally, the dataset can be understood better and key factors can be found with the aid of the depiction of the correlation subplot. All the features that are connected with the predicted property "Biopsy" are shown in Fig. 2's illustration. The correlation is stronger when the value is higher and the color is darker.

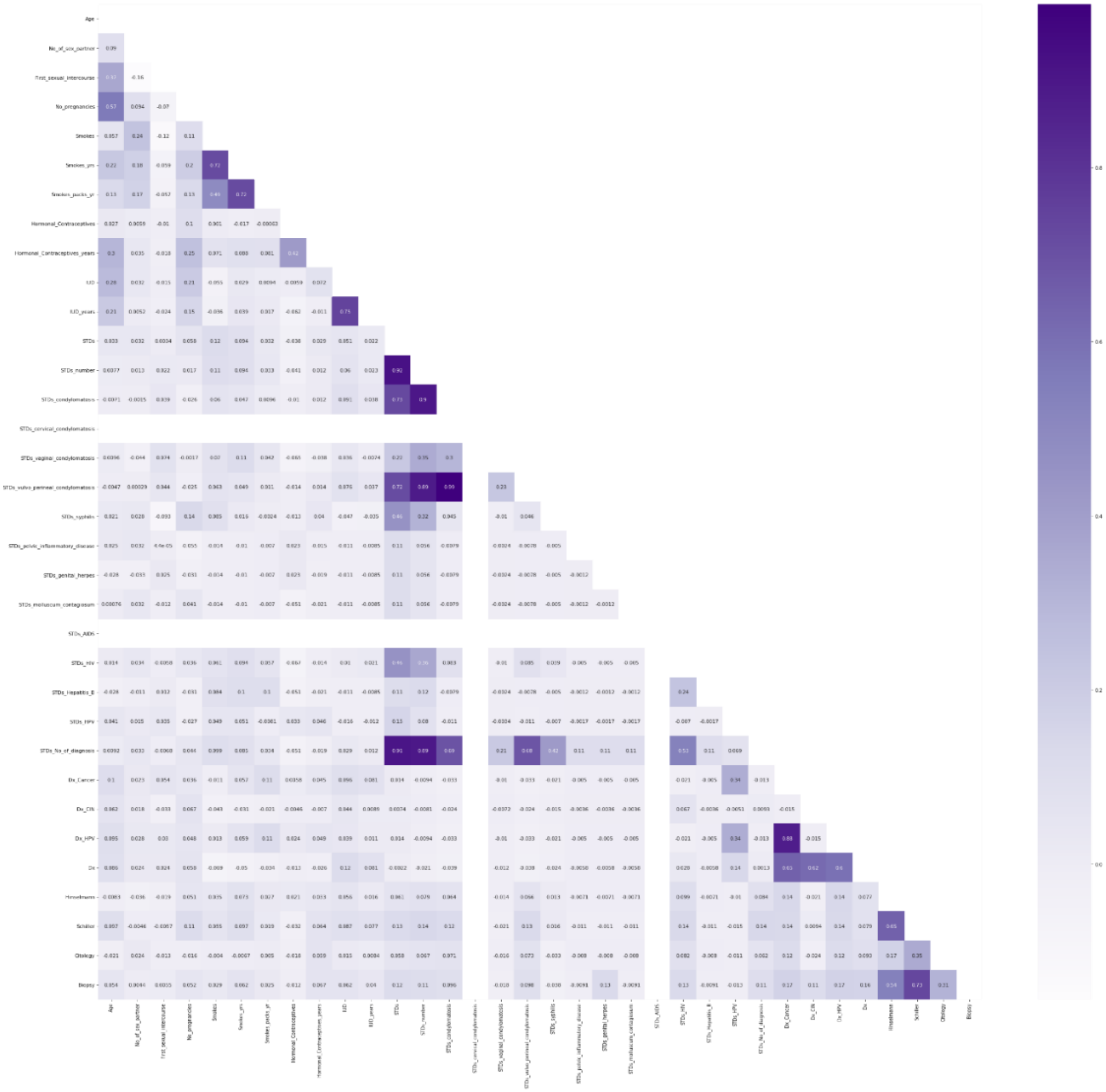


Figure 3.3: Correlation between the attributes

3.4 Classification Algorithm

The conventional machine learning classifiers implemented in this study with certain set of parameters are Logistic Regression (LRC), Random Forest Classifier (RFC), Gaussian Naïve Bayes (NBC), Decision Tree Classifier (DTC), K-Nearest Neighbors Classifier (KNNC), Gradient Boosting Classifier (GBC), AdaBoost Classifier (ABC), XGBoost Classifier (XGBC) and Support Vector Classifier (SVC) to generate the classification outcome for the dataset.

3.4.1 Logistic Regression (LRC): Using the correlation between independent and dependent variables, logistic regression models can predict the dependent variable of interest. In some cases, logistic regression may be used to estimate the probability of a discrete result in light of one or more input factors. The most often used logistic regression models provide a binary output. When there are more than two possible outcomes, multinomial logistic regression can be used as a modeling tool. While attempting to find which category a fresh sample most closely resembles, Logistic regression is a useful method of statistical analysis[19]. The parametric form of the distribution $P(Y|X)$ is logistic regression where the parametric model can be expressed as (see equation 1,2) [21],

$$P(Y = 1|X) = \frac{1}{1 + \exp\left(w_0 + \sum_{i=1}^{n} w_i x_i\right)} \qquad (1)$$

$$P(Y = 1|X) = \frac{\exp\left(w_0 + \sum_{i=1}^{n} w_i x_i\right)}{1 + \exp\left(w_0 + \sum_{i=1}^{n} w_i x_i\right)} \qquad (2)$$

Here $X = \{x1, x2, \ldots, xn\}$ indicates a vector containing either discrete or continuous values where Y is a discrete value. W denotes the likelihood that the observed Y values in the training data will occur.

3.4.2 Decision Tree Classifier (DTC): Decision trees organize instances into categories by branching them out from a central node to a set of "leaf" nodes that provide the classification. To assign a category to an instance, we examine the attribute pointed out by the node at the tree's root, and then follow the branch of the tree that corresponds to the attribute's value. The Decision Tree technique, which uses only two numClasses, is one of the most powerful and well-known prediction techniques available. A decision tree is a hierarchical data structure in which each leaf node stands for a distinct class and each inside node stands for a separate attribute test. Commonly utilized is a tree structure that progresses backwards in "learning" based on decision trees (DT). Problems involving classification and regression can both be solved using the technique. The "splitting" process is then used to identify the "Best Feature" or "Best Attribute" from the set of accessible characteristics as the tree develops from the root node. It is common practice to determine the "Best Attribute" by calculating two additional metrics, "Entropy," as shown in (3), and "Information Gain," as shown in (4) [20]. The feature that offers the most valuable data is the "best characteristic." Dataset homogeneity is measured by entropy, while the rate of change in entropy of characteristics is measured by information gain.

$$E\ (D) = -P\ (positive) log2\ P\ (positive) - P\ (negative) log2\ P\ (negative) \quad (3)$$

$$Gain\ (Attribute\ X) = Entropy\ (Decision\ Attribute\ Y\ ) - Entropy(X, Y\ ) \quad (4)$$

Here The Entropy E of a dataset D that contains both positive and negative "Decision Attributes" is determined by Equation (3).

3.4.3 Random Forest Classifier (RFC): An ensemble approach, the Random Forest (RF) classifier is used. This suggests that it has multiple algorithms. Usually It consists of various DT algorithms in this instance. During the training portion, RF constructed a complete forest out of several unrelated and random Decision Trees. In ensemble learning techniques, many learning algorithms are combined to build a single, superior predictive model. Although RF uses more features than a solitary DT, computational complexity may rise as a result, it typically performs more accurately when working with unknown datasets. The outcome of the Random Forest algorithm is the average outcome of all Decision Tree algorithms combined. In order to achieve the optimal outcome, the Random Forest ensemble classifier constructs and combines a number of decision trees. Assembling bootstraps is mostly used to understand trees. Assuming the provided data $X = \{x_1, x_2, x_3, \dots, x_n\}$ along with the response $Y = \{x_1, x_2, x_3, \dots, x_n\}$ having the lower limit b=1 to upper limit B. So, the prediction for sample x' consists of averaging the prediction $\sum_{b=1}^{B} f_b(x')$ for every single tree x' as shown below (see equation 5) [21],

$$j=\frac{1}{B}\sum_{b=1}^{B} f_b(x') \qquad (5)$$

In large data research, the Random forest (RF) classifier is often employed since it is a combination of numerous different tree predictors. It's a technique for learning useful in ensemble classification and regression.

3.4.4 K-Nearest Neighbors Classifier (KNNC): K-Nearest Neighbors (n neighbors = 5) is a popular classification strategy in machine learning. It has previously been used to treat several diseases. KNN is called nonparametric since it makes no assumptions about data distribution. KNN takes into account the similarities between the new and old data and assigns it to the group that it most closely resembles. KNN is utilized for both regression and recognition issues. It is called the lazy learner algorithm because it takes time to learn from training data. Using the equation, KNN computes the Euclidean distance between new $A\ (x_1, y_1)$ data and previously accessible $B(x_2, y_2)$ data (see equation 6)[22].

$$\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \qquad (6)$$

In two-dimensional space, the Euclidean formula can be used to calculate the distance between two data points (x_2, x_1) and (y_2, y_1). KNN assigns new data to the class with the shortest Euclidean distance to the new data.

3.4.5. AdaBoost Classifier (ABC): Adaptive Boosting, or AdaBoost for short, is a Boosting strategy for binary classification that combines many poor classifiers into a single, more reliable one. The approach generates the expected precision with a sample size of 1000 As illustrated in, the training dataset instances are weighted with a starting weight (7)

$$Weight(xi) = \frac{1}{N} \qquad (7)$$

For each input variable, the decision stump produces an output. where N is the total number of instances used for training and xi is the i^{th} example. An output is generated by the decision stump for each input value. The rate of misclassification is then computed using equation (8).

$$Error = \frac{Correct - N}{N} \qquad (8)$$

where N denotes the number of training instances. Boosting is just using many basic trainers intanded to get a more precise prediction. AdaBoost (Adaptive Boosting) adjusts the weights of samples and classifiers. As a result, the classifiers focus on findings that are rather difficult to reliably classify [23]. Equation depicts the final classification formula (9).

$$h_k(\mathrm{p}) \ = \ +/-(\, f(x) = a_0 + (\sum_{k=1}^{k} (a_k h_k(\mathrm{p})) \qquad (9)$$

A linear combination of all weak classifiers (simple learners) is shown in Equation (9), where k denotes the total number of weak classifiers.$h_k(\mathrm{p})$ is the output of the weak classifier t (which can be 1 or 1). The weight of classifier k is denoted by a_k.

3.4.6. Gradient Boosting Classifier (GBC): For classification and regression issues, the Boosting method known as Gradient Boosting only needs 100 samples. An enhanced loss function, a weak learner to produce predictions, and an additive model to combine weak learners in order to minimize the loss function make up Gradient Boosting. In order to make algorithms more effective, Gradient Boosting can be used to eliminate overfitting.

The 'Grabit' model, which is the result of applying gradient tree Boosting to the Tobit model, improves accuracy in situations when there is a mismatch between the numbers in each class. Despite it requires prior knowledge of a specific area, boosting over base techniques, also known as tree based learners, can improve prediction accuracy across a wide variety of datasets [24]. Unlike conventional machine learning, Boosting does not include optimizing the function space. After m^{th} iterations, the optimal function $F(X)$ is attained (see equation 10).

$$\mathrm{F\ (X)} = \sum_{i=0}^{m} \mathrm{fi(x)} \qquad (10)$$

where $fi\ (x)\ I\ =\ 1, 2, \ldots, M)$ denotes feature increments and $fi\ (x)\ =\ I\ x\ gm\ (X)$. The most recent base-learner has the highest loss function that is connected with negative gradients.

The m^{th} iteration for the negative gradient is(11),

$$\mathrm{gm}\ =\ -\left[\frac{\partial \mathrm{L\ (y, F\ (X)}}{\partial \mathrm{F\ (X)}}\right] \mathrm{F(X) = Fm - 1(X)} \qquad (11)$$

where gm is the path along which the loss function reduces the fastest when $F(X) = Fm - 1(X)$. A new decision tree seeks to repair the error caused by its predecessor. The T model is then altered to (12).

$$F_m\ (X) = F_m\ -\ 1\ (X) +\ \rho_m x h_m(X, \alpha_m) \qquad (12$$

3.4.7. XGBoost Classifier (XGBC): To put it simply, XGBoost is a decision tree with a gradient boost. Following this method, decision trees are generated one after the other. When using XGBoost, weights are essential. The decision tree, which makes predictions based on a number of factors, is fed information about the various weights assigned to the independent variables. Variables that were not initially considered by the tree are given a larger weight and are used to train a second decision tree. A more robust and accurate model is constructed by combining many independent classifiers. It can do regression, classification, ranking, and custom prediction tasks [25].

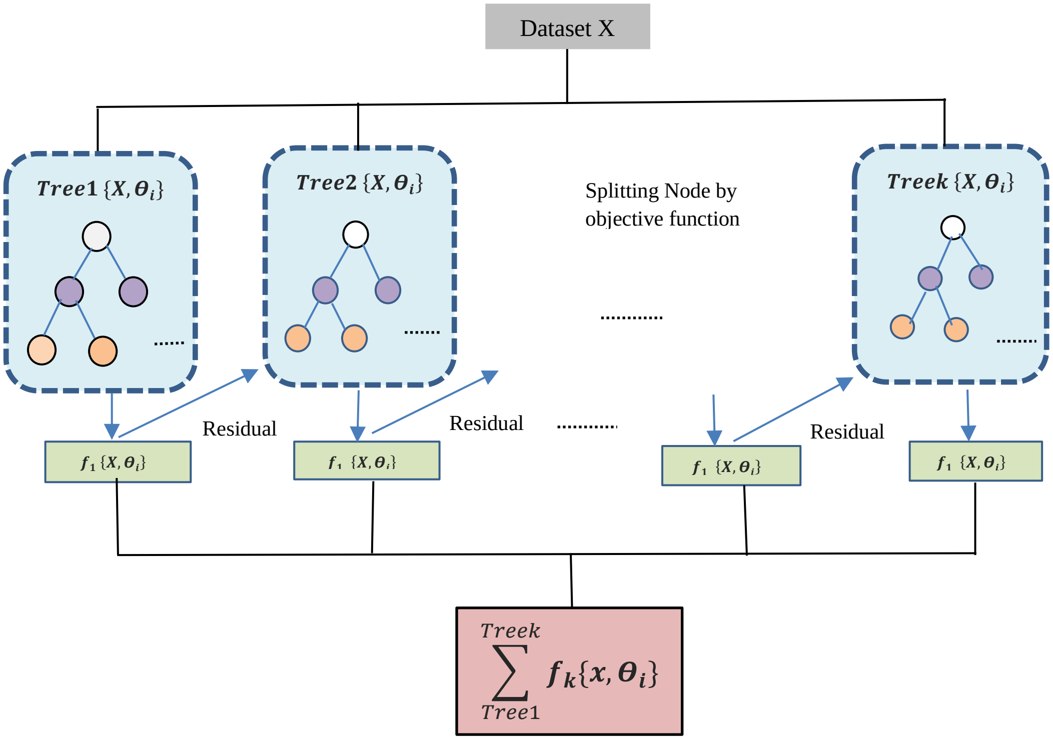


Figure 3.4.7: Structure of XGBoost Classifier

3.4.8. Gaussian Naïve Bayes (NBC): Naive Bayes classifier refers to a group of classification algorithms based on Bayes' Theorem which calculates the likelihood of an event occurring given the chance of another event occurring as expressed in equation (20). It is a family of algorithms that all share a fundamental premise, which is that every pair of features being categorized is independent of each other (see equation 13)[26].

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \qquad (13)$$

For each feature in Gaussian Naive Bayes, the continuous value is assumed to have a Gaussian distribution. The term "Normal distribution" is often used interchangeably w$_i^{th}$ "Gaussian distribution." The resulting histogram looks like a bell curve, with all points being roughly equal distance from the curve's center. The conditional probability is provided by (see equation 14)[26] if the feature likelihood is Gaussian.

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y{}^2}} exp\left(-\frac{(x_i - \mu_y)^2}{2\pi\sigma_y{}^2}\right) \quad (14)$$

3.4.9. Support Vector Classifier (SVC):

One of the most well-known methods of Supervised Learning, Support Vector Machine (SVM) may be applied to both Classification and Regression problems. Still, Machine Learning for Classification issues is where it is most often applied. In order to classify new data points efficiently, the SVM method seeks to locate a line, or decision boundary, that divides the space into classes in the most optimal way possible over all n dimensions. This bound of maximized utility is a hyperplane. The hyperplane may be built with the help of SVM, which chooses the most extreme points and vectors. Consequently, the name of the technique, Support Vector Machine[27], comes from the term "support vector," which is used to describe these extreme situations.

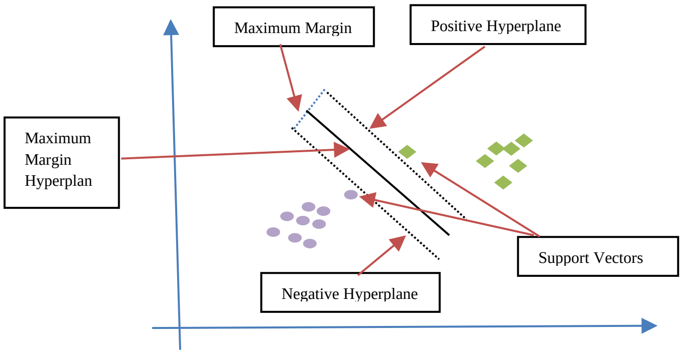


Figure 3.4.9: Concept Diagram of Support Vector Machine

3.5 Ensemble Classifier

Ensemble techniques are a methodology for machine learning that integrates multiple base models into one best predictive model. In order to get improved and more efficient score ensemble classifiers are also employed in this study so that an analysis of comparison can me made between the classifiers [28]. The ensemble classifier employed in this study are Bagging, Boosting, Stacking and Voting Classifier.

3.5.1. Bagging:

To lower volatility, handle dimensionality, and handle missing data, Bootstrap Aggregating is a model averaging approach. While the Bagging approach improves TA and stability for many different kinds of algorithms, DT algorithms are where it really shines. Two ensemble hybrid models based on RFC, DTC, KNNC, and NBC were created using Bagging methods and employed in both the training and testing stages. Different types of hybrid models including RFBM, NBBM, DTBM, and KNNBM have been developed. Equation (15) [29] provides the classification formula for the Bagging technique.

Here, $f'(x)$ is the mean of $f_i\ (x)$ for $i = 1, 2, 3, \ldots, T$.

$$f'(x) = sign(\sum_{i=1}^{T} f_i\ (x) \qquad (15)$$

3.5.2. Boosting:

To increase the efficiency of individual models that produce loss functions, "boosting" uses weighted average to turn several weak learners into strong ones [30]. In order to create the hybrid models ABBM, XGBBM, and GBBM, the authors apply the boosting method to the training and testing phases of GBC, XGBC, and ABC, respectively. The equation gives the boosting formula (16). Let $\gamma_i = 1/2 - €_t$ denote how much better f_i is on the weighted sample than flipping a coin.

$$\frac{1}{n}\sum_{i=1}^{n} I(\,y_i g(x_i) < 0) \leq \prod_{t=1}^{T} \sqrt{1 - 4\gamma_i{}^2} \qquad (16)$$

3.5.3. Stacking (DRLSKANGXS): Stacking, often known as Stacked Generalization, is an alternative paradigm. The purpose of stacking is to investigate many models for the same problem. The concept is that you can approach a learning problem using several sorts of models that are able to learn a portion of the problem but not its entirety. A separate intermediate prediction may be built for each learnt model, making it feasible to build numerous unique learners. In this way, the intermediate predictions may be used to train a second model that will eventually learn the same objective.

Hence the name, this model is supposed to be stacked on top of the others. Thus, you can enhance your total performance, and you frequently end up with a model that is superior to each intermediate model. Eventually, stacking trains a single model that integrates the predictions of numerous algorithms and makes a new prediction. Stacking yields a more efficient result than any single model. Using a joiner algorithm, it can provide a representation of any ensemble approach [31]. In this study, we employed a stacking technique that utilized Logistic regression as a joiner algorithm to merge all conventional classifiers into a final prediction.

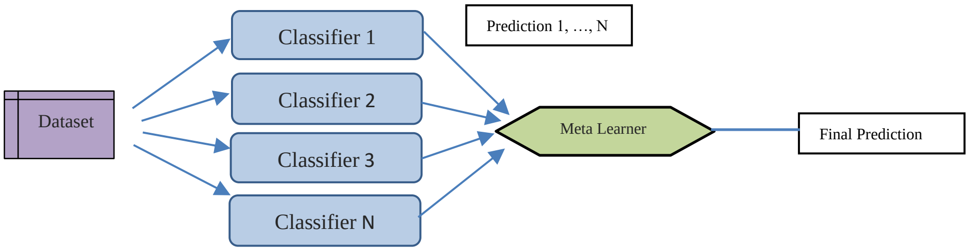


Figure 3.5.3: Concept Diagram of Stacking Classifier

3.5.4. Voting Classifier (DRLSKANGXV): The voting classifier takes into account the findings of many classifiers to make a prediction about which class has the most votes. In other words, several models work together to train a single model to predict output by tallying up the votes for each class. Answers were obtained by using a soft voting classifier (DRLSKANGSV) to merge all traditional classifiers. Using the classifier's anticipated probability p, we make predictions for the class labels in soft voting. In order for this strategy to be effective, the classifier must be properly calibrated [20]. The average probability score was determined via soft voting since it takes into consideration the uncertainty of each classifier (see equation 17) [32]. ST denotes the weight that can be assigned to the j^{th}classifier.

$$y' = argmax(\sum_{j=1}^{m} w_i p_{ij} \qquad (17$$

3.6 Dataset Description

The dataset used in this study is publicly available on 'UCI Machine Learning Repository' [33]. The dataset consists of 858 samples and 36 attributes. The class attribute is 'biopsy' which indicates if the test result is positive or negative denoting by 0 and 1. The dataset can be considered as imbalanced as the class attribute holds the data for negative and positive biopsy result of 93.6% and 6.4% respectively. The dataset contains both numerical and categorical data. There exist missing values for several classifiers which needed to be imputed and gone through the implementation of several preprocessing techniques before being ready to fit for the classification models. The details of the dataset for each attribute is stated in Table 1 below.

TABLE 3.6: DETAILS OF THE DATASET

Attribute Names	Value Range	Attribute Types
Age	Age between 13 years and 84 years	Numerical (int)
Number of sexual partners	Ranging between 1 partner to 28 partners	Numerical (int)
First sexual intercourse	10 to 32	Numerical (int)
First sexual intercourse	0 to 11	Numerical (int)
Number of pregnancies	0 and 1	Numerical (int)
Smokes	Ranging between 0 to 37	Categorical (bool)
Smokes (years)	Ranging between 0 to 37	Numerical (int)
Smokes (packs/years)	0 and 1	Numerical (int)
Hormonal Contraceptives	Ranging between 0 to 30	Categorical (bool)
Hormonal Contraceptives (years)	0 and 1	Numerical (int)
IUD	0 and 1	Categorical (bool)
IUD(years)	Ranging between 0 to 19	Numerical (int)
STDs	0 and 1	Categorical (bool)
STDs (number)	Ranging between 0 to 4	Categorical (bool)
STDs:condylomatosis	0 and 1	Categorical (bool)
STDs:cervical condylomatosis	Only 0	Categorical (bool)
STDs:vaginal condylomatosis	0 and 1	Categorical (bool)
STDs:vulvo-perineal condylomatosis	0 and 1	Categorical (bool)
STDs:syphilis	0 and 1	Categorical (bool)
STDs:pelvic inflammatory disease	0 and 1	Categorical (bool)
STDs:genital herpes	0 and 1	Categorical (bool)

STDs:molluscum contagiosum	0 and 1	Categorical (bool)
STDs:AIDS	Only 0	Categorical (bool)
STDs:HIV	0 and 1	Numerical (int)
STDs:Hepatitis B	0 and 1	Numerical (int)
STDs:HPV	0 and 1	Categorical (bool)
STDs: Number of diagnosis	Ranging between 0 to 3	Numerical (int)
STDs: Time since first diagnosis	Ranging between 1 to 22	Numerical (int)
STDs: Time since last diagnosis	Ranging between 1 to 22	Numerical (int)
Dx:Cancer	0 and 1	Categorical (bool)
Dx:CIN	0 and 1	Categorical (bool)
Dx:HPV	0 and 1	Categorical (bool)
Dx	0 and 1	Categorical (bool)
Hinselmann	0 and 1	Categorical (bool)
Schiller	0 and 1	Numerical int)
Citology	0 and 1	Categorical (bool)
Biopsy	0 and 1	Categorical (bool)

3.7 Feature Selection Method:

Each column of the dataset represents a feature. To train an optimal model, we must ensure that only the most important features are utilized. If we have too many features, the model can learn from noise and capture insignificant patterns if there are too many. Feature Selection is the process of determining which data attributes are the most significant. The feature selection methods employed in this study are Recursive Feature Elimination (RFE), Univariate Feature Selection (UFS) to do the ensemble of these two feature selection methods.

3.7.1 Recursive Feature Elimination (RFE): A technique for choosing among attributes Optimizing a model using recursive feature elimination (RFE) involves gradually reducing the number of features until the desired number is obtained. RFE prioritizes features using the model's coef_ or feature importances_ attributes to reduce dependencies and collinearity by recursively deleting a small number of features in each loop. RFE necessitates keeping certain features, but it's not always clear which ones are real. Several feature subsets are scored utilizing RFE's cross-validation method, and the highest-scoring set of features is selected as the optimal set of features. The RFECV visualizer shows how many features were included in the model, how well they performed in cross-validation tests, and how much variation there was in the results [34].

In this study, the implementation of RFE took place for several classifiers. The best set of features were selected based on the score generated by fitting Decision Tree Classifier (DT) & Random Forest Classifier (RFC). A set of 21 features with a score of 79.87% were selected using Decision Tree Classifier (DTC) & Logistic Regression (LRC). For Random Forest Classifier (RFC), a set of 19 features were selected from the loop of several set of attributes having a score of 68.39%. The traditional approach of RFE fitted on Logistic Regression Classifier (LRC) selected 33 best features for efficient classification performance.

3.7.2 Univariate Feature Selection (UFS): Univariate Feature Selection selects the best features comparing each attribute to the dependent variable to see whether or not they have a statistically significant relationship using univariate statistical test. It is also known as variance analysis (ANOVA). When analyzing the link between one characteristic and the dependent variable, the other characteristics were disregarded. This is why it is referred to as "univariate." Each element has its own test score. Finally, all test scores are evaluated, and the highest-scoring features are chosen [35]. The ensemble of two feature selection methods employed in this study were also done to reduce a significant number of features for an improved score generation. The selection of 33 features while implementing RFE using Logistic Regression Classifier (LR) were merged with Univariate Feature Selection Method to reduce the number of features and make it down to 26 features which lead to the generation of scores those drop marginally with less computational complexity with less runtime.

3.8 Hyper Parameter Tuning (HPT)

Hyper parameter tuning is the process of identifying the optimal hyper parameter values for a base classifier and then deploying that algorithm on any given dataset. Using this set of hyperparameters, the model's performance is maximized by minimizing a certain loss function, leading to more accurate predictions. Note how the learning algorithm strives for the best possible answer within the given constraints, optimizing the loss depending on the input data. However, hyperparameters provide a fine-grained description of this environment. GridSearchCV is the most elementary strategy for adjusting hyperparameters. The procedure entails constructing a model for each feasible value of each hyperparameter, analyzing the outcomes of each model, and finally picking the best-performing architecture. [36]. In this study, the base classifiers to go through all techniques are Decision Tree Classifier (DT) and Random Forest Classifier (RF). The best parameter set selection using hyper parameter tuning method was done on these classifiers as well as shown in Table 2.

TABLE 3.8: HYPER PARAMETER TUNING

Model	Parameter Set	Best Parameter
Decision Tree Classifier (DTC)	'criterion':['gini','entropy, 'splitter':['best','random'], 'max_depth':[3,4,5,6],'max_features':['auto','log2'],'random_state':[123]	'criterion':'gini', 'max_depth':4,'max_features':'auto','random_state':123,'splitter':'best'
Random Forest Classifier(RFC)	'n_estimators': range(10,100,10),'max_depth' : range(2,10,1),'criterion' : ['gini','entropy'],'max_leaf_nodes' : range(2,10,1),'max_features' : ['auto','log2']	'n_estimators': range(10,100,10),'max_depth' : range(2,10,1),'criterion' : ['gini','entropy'],'max_leaf_nodes' : range (2,10,1),'max_features' :['auto','log2']

3.9. Performance Matrix Evaluation

Using performance metrics, the effectiveness and precision of the machine learning process may be evaluated. A person is positively classified as having Cervical Cancer when they are identified as having the disorder. When an individual is not diagnosed with Cervical Cancer, he has a negative categorization. The equations stated below from (18) to (21) was applied to arrive at these results:

TP = A model is considered to have a true positive result when it is appropriately identified as having Cervical Cancer.

TN = True Negative (where the model accurately identified the opposing class, like patients without cervical cancer problem).

FP = False Positive, this occurs when the model mistakenly classifies cervical Cancer patients as non-cervical cancer patients.

FN = False Negative, when the model mistakenly classifies one class as the other, for as when it classifies cervical cancer patients as normal patients.

Accuracy: It refers to the proportion of test data for which predictions were accurate. Where precision outperforms the availability of measurements with real measurements. There is only one aspect involved. Systematic inaccuracies are addressed through accuracy.

$$\frac{TP + TN}{TP + TN + FP + FN} \qquad (18)$$

Precision: It refers to the proportion of positive observations that were accurately predicted. Precision identifies the actual true portion of the total number of occasions in which the prediction was accurate.

$$\frac{TP}{TP + FP} \qquad (19)$$

Recall: It represents the proportion of accurately predicted positive observations.

$$\frac{TP}{TP+FN} \qquad (20)$$

F1-score: In its most basic form, it is the harmonic average of the recall and the precision.

$$\frac{2(Precision\ X\ Recall)}{(Precision + Recall)} \qquad (21)$$

roc_auc score: The ROC-AUC indicates the degree of distinction between the predictions of the two classes. The higher the score, the greater the differentiation and the smaller the overlapping of the forecasts of the two classes.

CHAPTER 4

Experimental Results and Discussion

4.1 Experimental Result Analysis

This section contains a comparative analysis on the basis of experimental results for several classifiers employed in this study on cervical cancer dataset. The dataset needed to go through several preprocessing techniques to analyze the performance of conventional machine learning classifiers and ensemble classifiers. The most important features for an improved score generation were done implementing feature selection techniques. The best parameter set was found out using hyper parameter tuning. The Table 3. Below conclude the scores come out for several classifiers where the performance evaluation measures were Training Accuracy (TRA), Testing Accuracy (TA), F1-Score (FS), Recall (R), Precision (P), roc_auc Score (RA) before (BS) and after (AS) implementing sampling technique.

TABLE 4.1: PERFORMANCE OF MODELS BEFORE AND AFTER EMPLOYING SAMPLING TECHNIQUE

Models	Sampling	TRA	TS	FS	R	P	RA
DTC	BS	96.41%	95.23%	71.74%	93.33%	58.33%	94.55%
	AS	100%	94.44%	61.11%	73.33%	52.38%	84.55%
RFC	BS	97.44%	93.65%	20%	13.33%	40%	56.03%
	AS	100%	95.24%	68.42%	86.67%	56.52%	91.22%
LRC	BS	97.26%	95.23%	57.14%	53.33%	61.53%	75.61%
	AS	96.06%	94.04%	63.41%	86.67%	50%	90.59%
KNNC	BS	94.53%	93.65%	27.27%	20%	42.85%	59.15%
	AS	93.96%	91.27%	8.33%	6.67%	11.11%	51.64%
GBC	BS	96.07%	95.23%	40%	26.67%	80%	63.12%
	AS	97.25%	94.44%	46.15%	40%	54.54%	68.94%
ABC	BS	98.97%	94.44%	58.88%	66.67%	52.63%	81.43%
	AS	99.36%	95.63%	66.67%	73.33%	61.11%	85.18%
XGBC	BS	96.59%	95.24%	40%	26.67%	80%	63.12%
	AS	98.53%	94.44%	99%	6.67%	99%	53.33%

SVC	BS	98.46%	93.25%	45.16%	46.67%	43.75%	71.14%
	AS	84.19%	94.04%	0%	0%	0%	50%

The base model considered in this study while employing several techniques except fitting the classifier models before and after implementing the sampling techniques synthetic minority oversampling technique (SMOTE). Thus, both feature selection techniques and hyper parameter tuning was done on the base models. Table 4. Below demonstrate the generated scores after the implementation of RFE and hyper parameter tuning on RFC and DTC.

TABLE 4.1(A): IMPLEMENTATION OF HYPER PARAMETER TUNING AND RFE ON BASE CLASSIFIERS

Models	**Technique Applied**	**TRA**	**TS**	**FS**	**R**	**P**	**RA**
DTC	Hyper Parameter Tuning	98.23%	96.03%	80%	95.23%	68.96%	95.67%
	RFE	100%	99.60%	97.67%	100%	95.45%	99.78%
RFC	Hyper Parameter Tuning	98.26%	96.82%	83.33%	95.24%	74.07%	96..10%
	RFE	100%	99.20%	95.24%	95.24%	95.24%	97.40%

In order to improve the classification performance, ensemble methods like bagging, boosting, stacking and voting classifier were employed after the performance of the conventional ML classifiers were employed. The classification performance of ensemble methods is shown in Table 5. stated below.

TABLE 4.1(B): CLASSIFICATION PERFORMANCE OF ENSEMBLE CLASSIFIERS

Score Matrices	**RFBM**	**KNNBM**	**DTBM**	**NBBM**	**GBBM**	**XGBBM**	**ABBM**	**DRLSKANGXS**	**DRLSKANGXV**
Accuracy	97.22%	95.63%	96.03%	88.09%	95.63%	95.63%	96.42%	98.41%	96.03%
Recall	87.66%	75.97%	78.54%	91.34%	88.96%	78.13%	89.39%	94.80%	82.28%

Precision	93.37%	93.75%	94.27%	70.16%	84.53%	90.97%	87.77%	94.80%	82.28%
F1 Score	90.27%	82.16%	84.23%	75.11%	86.58%	83.11%	88.55%	94.80%	82.28%

The ensemble of feature selection method RFE and Univariate is a significant approach which covers two goals of reducing a significant number of features and generating an improved score at the same time. The RFE feature selection method which selects 33 best features using Logistic Regression Classifier got merged with univariate feature selection technique to reduce the number of features to 26 features. The scores generated after the implementation of the ensemble of feature selection methods are stated in Table 6.

Table 4(c): Classification performance of traditional classifiers after employing feature selection technique

Models	**TRA**	**TS**	**FS**	**R**	**P**	**AUC**
DTC	96.41%	95.63%	71.79%	93.33%	58.33%	94.55%
RFC	97.44%	95.63%	52.17%	40%	75%	69.51%
LRC	97.44%	95.63%	62.06%	60%	64.28%	78.94%
KNNC	94.53%	93.65%	27.27%	20%	42.85%	59.15%
GBC	94.88%	95.23%	33.33%	20%	100%	60%
ABC	98.46%	95.23%	64.70%	73.33%	57.89%	84.98%
XGBC	95.90%	94.84%	23.52%	13.33%	100%	56.67%
SVC	93.34%	94.04%	0%	0%	0%	50%

The tables and discussion concludes the overall classification performance and the techniques employed in study. According to the overall performance analysis, RF generates the highest accuracy of 99.60% while RFE was employed to the classifier.

TABLE 4(D): COMPARISON BETWEEN THIS WORK AND OTHER RELATED WORKS

Reference	Dataset	Best Model	Accuracy
[8]	Seer Database	RFC	96.98%
[9]	UCI Machine Learning Repository	DTC	98%
[10]	UCI Machine Learning Repository	DTC	99%
[12]	UCI Machine Learning Repository	DTC, RFC, XGBC	93.33%
[15]	UCI Machine Learning Repository	DTC	98.82%
This Work	UCI Machine Learning Repository	RFC	**99.60%**

4.2. Runtime Calculation:

The effectiveness of an algorithm can be characterized by its runtime calculation, which indicates how much more computing power and time are required to execute the method. Runtime analysis of a method is not only essential in order to understand the internal workings of the algorithm, but it also generates a more effective execution. This is because runtime analysis is performed while the algorithm is really being run [20]. Table 7. demonstrates the runtime needed for several classifiers employed in this study.

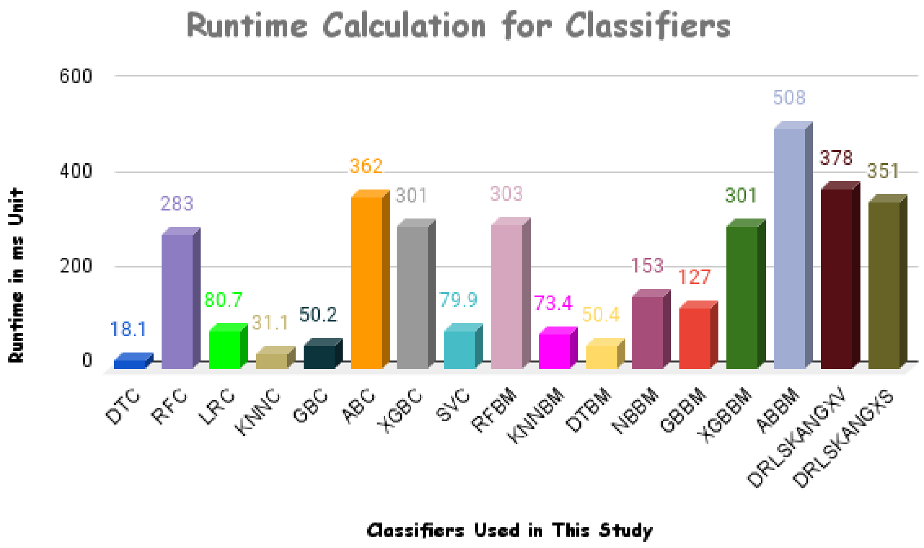


Figure 4.1: Runtime Calculation for All Models

CHAPTER 5

Conclusion and Future Work

5.1 Limitation

The missing value imputation technique employed in this work was based on predicting Decision Tree Classifier (DTC) which must not hundred percent accurate while implementing on healthcare data. Several implementation of pre-processing techniques may put an effect on prediction scores. A very marginal drop took place in classification outcomes for several classifiers while applying classification algorithms on 26 features. This study looks forward to expanding resolving this issues while making further updates.

5.2 Conclusion and Future Work

The applied approach in this study generates high testing scores recognizing the risk factors of cervical cancer. The alarming increasing rate of cervical cancer has become a massive issue to handle in recent years. A new recognizing approach generating good classification performance overall can add new possibility to the goal of reducing mortality rate due to getting affected by cervical cancer. This research takes the conventional methods that other researchers used in recent years and makes some differences from other recent researches while using multiple ensemble techniques for efficient classification performance and the implementation of the ensemble of feature selection was done. This study generates a very high accuracy score of 99.60% while using RFE feature selection technique on RF classifier. Our goal is to implement few more models on the dataset or using a different dataset with higher dimensionality for implementing more techniques to do the preprocessing of the dataset. The inclusion of Deep Learning approaches can also be a significant way to improve the classification performance with more updated techniques come forward to be implemented in future.

Reference:

[1] Burd, Eileen M. “Human papillomavirus and cervical cancer.” Clinical microbiology reviews vol. 16,1 (2003): 1-17. doi:10.1128/CMR.16.1.1-17.2003

[2] Mukama, T., Ndejjo, R., Musabyimana, A., Halage, A. A., & Musoke, D. (2017). Women's knowledge and attitudes towards cervical cancer prevention: a cross sectional study in Eastern Uganda. BMC women's health, 17(1), 9. https://doi.org/10.1186/s12905-017-0365-3

[3] Cervical Cancer, available at << https://www.who.int/news-room/fact-sheets/detail/cervical-cancer>>, last accessed on January 1, 2023, 7:24pm

[4] Getaneh, Alem et al. “Knowledge, attitude and practices on cervical cancer screening among undergraduate female students in University of Gondar, Northwest Ethiopia: an institution based cross sectional study.” BMC public health vol. 21,1 775. 23 Apr. 2021, doi:10.1186/s12889-021-10853-2

[5] Sherris, J et al. “Cervical cancer in the developing world.” The Western journal of medicine vol. 175,4 (2001): 231-3. doi:10.1136/ewjm.175.4.231

[6] WORLDHEALTHRANKINGS, available at << https://www.worldlifeexpectancy.com/cause-of-death/cervical-cancer/by-country/female>>, last accessed on January 1, 2023, 7:31pm

[7] Late-stage cervical cancer still on the rise despite ways to prevent, detect and treat early, available at << https://abcnews.go.com/Health/late-stage-cervical-cancer-rise-ways-prevent-detect/story?id=88704185#:~:text=The%20CDC%20reports%20that%2013%2C000,survival%20rate%20of%20only%2017%25.>>, last accessed on January 1, 2023, 7:34pm

[8] Nithya, B., and V. Ilango. "Optimized machine learning based classifications of staging in gynecological cancers using feature subset through fused feature selection process." *International Journal of Advanced Computer Science and Applications* 11, no. 7 (2020).

[9] Tanimu, J.J., Hamada, M., Hassan, M., Kakudi, H. and Abiodun, J.O., 2022. A Machine Learning Method for Classification of Cervical Cancer. Electronics, 11(3), p.463.

[10] Al Mudawi, N. and Alazeb, A., 2022. A Model for Predicting Cervical Cancer Using Machine Learning Algorithms. Sensors, 22(11), p.4132.

[11] Chaudhuri, A.K., Ray, A., Banerjee, D.K. and Das, A., 2021. A multi-stage approach combining feature selection with machine learning techniques for higher prediction reliability and accuracy in cervical cancer diagnosis. Int. J. Intell. Syst. Appl, 13, pp.46-63.

[12] Akter, L., Islam, M., Al-Rakhami, M.S. and Haque, M., 2021. Prediction of cervical cancer from behavior risk using machine learning techniques. SN Computer Science, 2(3), pp.1-10.

[13] Gupta, S. and Gupta, M.K., 2022. Computational prediction of cervical cancer diagnosis using ensemble-based classification algorithm. The Computer Journal, 65(6), pp.1527-1539.

[14] Khan, I.U., Aslam, N., Alshehri, R., Alzahrani, S., Alghamdi, M., Almalki, A. and Balabeed, M., 2021. Cervical Cancer Diagnosis Model Using Extreme Gradient Boosting and Bioinspired Firefly Optimization. Scientific Programming, 2021.

[15] Tanimu, J.J., Hamada, M., Hassan, M. and Ilu, S.Y., 2021. A contemporary machine learning method for accurate prediction of cervical cancer. In SHS Web of Conferences (Vol. 102, p. 04004). EDP Sciences.

[16] SMOTE for Imbalanced Classification with Python, available at << https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/>>, last accessed on January 1, 2023, 7:52pm

[17]scikitlearn, available at <<https://scikitlearn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html>>, last accessed on January 1, 2023, 7:55pm

[18] How to Calculate Correlation Between Variables in Python, available at << https://machinelearningmastery.com/how-to-use-correlation-to-understand-the-relationship-between-variables//>>, last accessed on January 1, 2023, 8: 01pm

[19] Classification with logistic regression (continued), available at << https://courses.cs.washington.edu/courses/cse446/22wi/schedule/week5.pdf>>, last accessed on January 1, 2023, 8: 24 pm

[20] Ghosh, P., Azam, S., Jonkman, M., Karim, A., Shamrat, F.J.M., Ignatious, E., Shultana, S., Beeravolu, A.R. and De Boer, F., 2021. Efficient prediction of cardiovascular disease using machine learning algorithms with relief and LASSO feature selection techniques. IEEE Access, 9, pp.19304-19326.

[21] Tajmen, S., Karim, A., Hasan Mridul, A., Azam, S., Ghosh, P., Dhaly, A.A. and Hossain, M.N., 2022, July. A Machine Learning based Proposition for Automated and Methodical Prediction of Liver Disease. In Proceedings of the 10th International Conference on Computer and Communications Management (pp. 46-53).

[22] K-Nearest Neighbor(KNN) Algorithm for Machine Learning, available at << https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning >>, last accessed on January 1, 2023, 8: 30 pm

[23] AdaBoost Classifier in Python, available at <https://www.datacamp.com/tutorial/adaboost-classifier-python>, last accessed on January 1, 2023, 9:14pm

[24] Gradient Boosting Algorithm: A Complete Guide for Beginners, available at << https://www.analyticsvidhya.com/blog/2021/09/gradient-boosting-algorithm-a-complete-guide-for-beginners/>>, last accessed on January 1, 2023, 9:19pm

[25] XGBoost, available at << https://www.geeksforgeeks.org/xgboost/>>, last accessed on January 1, 2023, 9:22pm

[26] Naive Bayes Classifiers, available at << https://www.geeksforgeeks.org/naive-bayes-classifiers/>>, last accessed on January 1, 2023, 9:25pm

[27] Support Vector Machine Algorithm, available at << https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm/>>, last accessed on January 1, 2023, 9:25pm

[28] Advanced Ensemble Classifiers, available at << https://towardsdatascience.com/advanced-ensemble-classifiers-8d7372e74e40#:~:text=Ensemble%20learning%20is%20a%20way,representation%20or%20the%20training%20set./>>, last accessed on January 1, 2023, 9:30pm

[29] What is Bagging classifier? available at << https://medium.com/@arch.mo2men/what-is-bagging-classifier-45df6ce9e2a1>>, last accessed on January 1, 2023, 9:33pm

[30] Boosting Algorithms Explained, available at << https://towardsdatascience.com/boosting-algorithms-explained-d38f56ef3f30>>, last accessed on January 1, 2023, 9:37pm

[31] Stacking Classifiers for Higher Predictive Cancer Diagnosis, available at << https://vannguyen-8073.medium.com/stacking-classifiers-for-higher-predictive-cancer-diagnosis-7021075ce21c >>, last accessed on January 1, 2023, 9:39pm

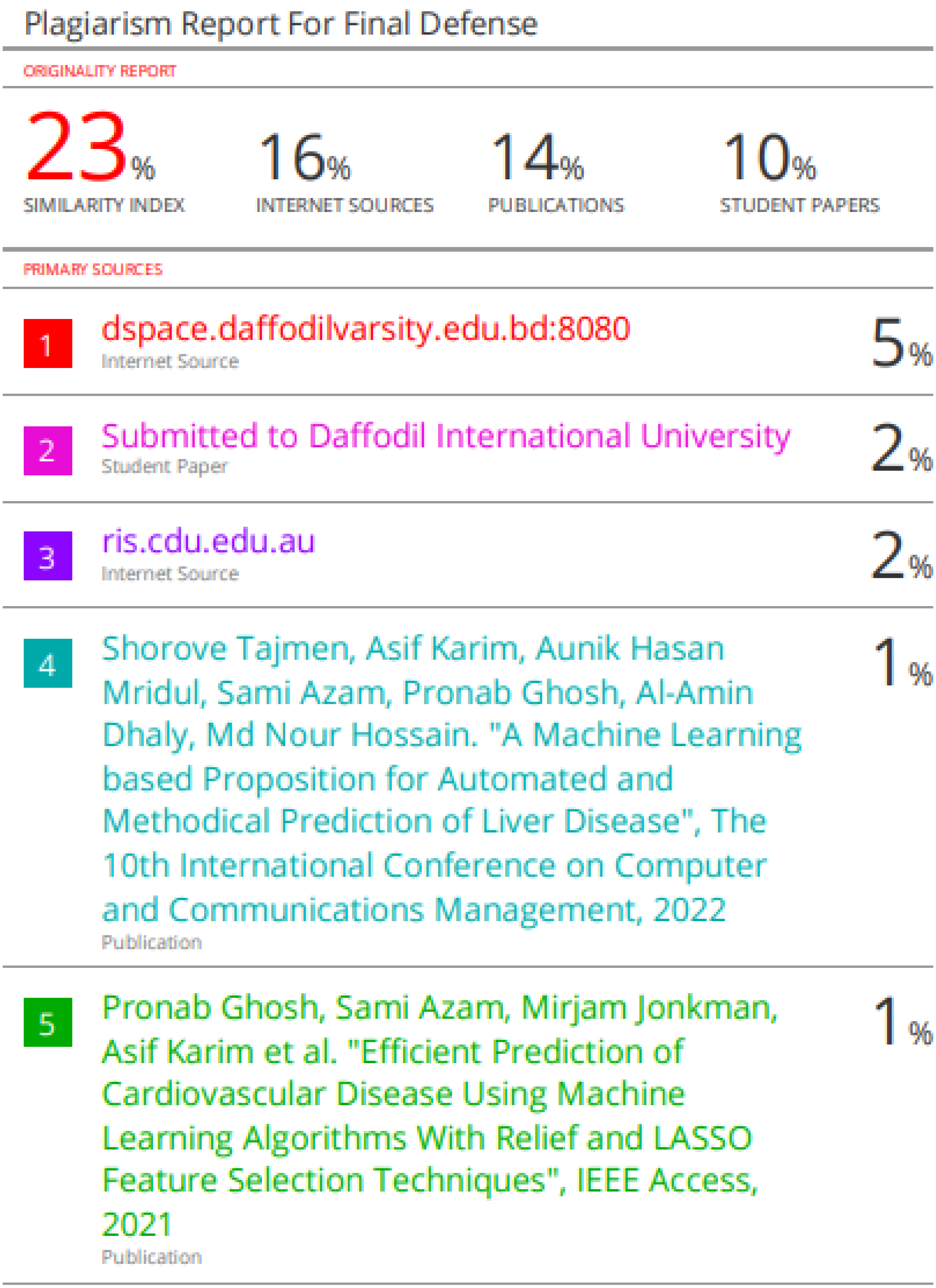
[32] VOTING CLASSIFIER USING SKLEARN, available at << https://prutor.ai/voting-classifier-using-sklearn/>>, last accessed on January 1, 2023, 9:42pm

[33]UCI Machine Learning Repository, available at << https://archive.ics.uci.edu/ml/datasets/Cervical+cancer+%28Risk+Factors%29 >>, last accessed on January 1, 2023, 9:50pm

[34] Recursive Feature Elimination, available at << https://www.scikit-yb.org/en/latest/api/model_selection/rfecv.html#:~:text=Recursive%20feature%20elimination%20(RFE)%20is,number%20of%20features%20is%20reached. >>, last accessed on January 1, 2023, 10:46pm

[35] Feature selection using Python for classification problems, available at << https://towardsdatascience.com/feature-selection-using-python-for-classification-problem-b5f00a1c7028#:~:text=Univariate%20feature%20selection%20works%20by,analysis%20of%20variance%20(ANOVA).>>, last accessed on January 1, 2023, 10:49pm

[36] Hyperparameter tuning for machine learning models. available at << https://www.jeremyjordan.me/hyperparameter-tuning/>>, last accessed on January 1, 2023, 10:58pm



Plagiarism Report For Final Defense

ORIGINALITY REPORT

23%	16%	14%	10%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	5%
2	Submitted to Daffodil International University Student Paper	2%
3	ris.cdu.edu.au Internet Source	2%
4	Shorove Tajmen, Asif Karim, Aunik Hasan Mridul, Sami Azam, Pronab Ghosh, Al-Amin Dhaly, Md Nour Hossain. "A Machine Learning based Proposition for Automated and Methodical Prediction of Liver Disease", The 10th International Conference on Computer and Communications Management, 2022 Publication	1%
5	Pronab Ghosh, Sami Azam, Mirjam Jonkman, Asif Karim et al. "Efficient Prediction of Cardiovascular Disease Using Machine Learning Algorithms With Relief and LASSO Feature Selection Techniques", IEEE Access, 2021 Publication	1%