PREDICTION OF CERVICAL CANCER IN BANGLADESH USING HYBRID MACHINE LEARNING ALGORITHMS

by

Fahima Khanam

MASTER OF SCIENCE IN INSTITUTE OF INFORMATION AND COMMUNICATION TECHNOLOGY

Institute of Information and Communication Technology
BANGLADESH UNIVERSITY OF ENGINEERING AND TECHNOLOGY
October 2021

The thesis titled "PREDICTION OF CERVICAL CANCER IN BANGLADESH USING HYBRID MACHINE LEARNING ALGORITHMS" submitted by Fahima Khanam, Roll No. 0417312045, Session: April 2017 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of Master of Science in Institute of Information and Communication Technology on 10th October 2021.

BOARD OF EXAMINERS

1. Dr. Md. Rubaiyat Hossain Mondal Professor IICT, BUET, Dhaka.	Chairman (Supervisor)
2. Dr. Md. Rubaiyat Hossain Mondal Professor & Director IICT, BUET, Dhaka.	Member (ex-Officio)
3. Dr. Md. Saiful Islam Professor IICT, BUET, Dhaka.	Member
4. Dr. Hossen Asiful Mustafa Associate Professor IICT, BUET, Dhaka.	Member
5. Dr. Md. Mahbubur Rahman Professor Dept. of CSE, Military Institute of Science and Technology, Dhaka-1216.	Member (External)

CANDIDATE'S DECLARATION

It is hereby declared that this thesis or any part of it has not been submitted elsewhere for the award of any degree.

Signature of the Candidate

fahima

Fahima Khanam

DEDICATION

This thesis work is dedicated to my dearest and honorable parents Mohammad Salimullah and Shahnaz Begum, who have always taught me to work hard and believe in Almighty Allah (SWT).

Table of Contents

List of Pseudocode

List of Figures

List of Tables

List of Abbreviations of Technical Terms

HPV	Human Papillomavirus	1.1
IUD	Intrauterine Device	1.2
UCI	University of California, Irvine	1.2
CIN	Cervical Intraepithelial Neoplasia	2.1.1
RF	Random Forest	2.2
LR	Logistic Regression	2.2
SVM	Support Vector Machine	2.2
SVC	Support Vector Classifier	2.2
KNN	K- Nearest Neighbor	2.2
DT	Decision Tree	2.2
ET	Extra Tree	2.2
ELM	Extreme Learning Machine	2.2
SA	Simulated Annealing	2.2
RFE	Recursive Feature Elimination	2.2
CNN	Convolutional Neural Network	2.2
GNRBA	Gauss-Newton Representation Based Algorithm	2.2
HMM	Hidden Markov Model	2.2
PNN	Probabilistic Neural Network	2.2
SMOTE	Synthetic Minority Over-Sampling Technique	2.2
ROC	Receiver Operating Characteristic	2.2
STD	Sexually Transmitted Diseases	3.2
DX	Diagnosis	3.2
OCP	Oral Contraceptive Pills	3.2
IUCD	Intra Uterine Contraceptive Device	3.2
VIA	Visual Inspections with Acetic acid	3.3
AUC	Area Under Curve	3.7.5

Acknowledgment

First and Foremost praise is to Almighty Allah, the greatest of all, on whom eventually we depend for sustenance and guidance. I would like to thank Almighty Allah for giving me the opportunity, determination, and strength to do my research. I would like to thank Kelwin Fernandes, Jaime S. Cardoso, and Jessica Fernandes for uploading their dataset to the UCI machine learning repository. I would also like to show my heartiest gratitude to Dr. Yearul Kabir, Professor and Chairman of the Department of Biochemistry and Molecular Biology at the University of Dhaka, for the guidance in the collection of this invaluable data. I would like to thank Elias Shawon Bhuiyan, a former student of Biochemistry and Molecular Biology, for his support in the collection of cervical cancer data from different hospitals of Bangladesh. I would also like to thank Dhaka Medical College (DMC) Hospital, Shaheed Suhrawardy Medical College and Hospital (SSMC), Bangladesh Cancer Society Hospital & Welfare Home in Bangladesh for allowing the collection of Bangladeshi cervical cancer data. Moreover, I would like to thank Dr. Lutfa Akther, Former program Coordinator, RHSTEP, Dhaka Medical College and Hospital for her unlimited cooperation in collecting the cervical cancer datasets.

I would like to thank my supervisor Dr. Md. Rubaiyat Hossain Mondal, Professor, IICT, BUET, for the patient guidance, encouragement, and advice he has provided me throughout my time as his student. I have been extremely lucky to have a supervisor who cared much about my work, and who responded to my questions and queries so promptly. In addition to being an admirable supervisor, he is a man of principles and has immense knowledge of research in general and his subject in particular. I am deeply indebted to him for his steady guidance in publishing my research work. I would also like to thank the members of my thesis examination board, Dr. Md. Saiful Islam, Dr. Hossen Asiful Mustafa, and Dr. Md. Mahbubur Rahman for their encouragement and insightful comments.

I wish to express my gratitude to the Bangladesh University of Engineering and Technology (BUET) for providing an excellent environment for research and ICT division, Ministry of Posts, Telecommunications & Information Technology, Bangladesh for financial support in the form of scholarship.

Most importantly, none of this would have been possible without the love and patience of my parents. I would like to express my heartfelt gratitude to my family.

Abstract

The aim of this research work is to apply machine learning algorithms for predicting cervical cancer. Early screening of vulnerable patients is essential to prevent cervical cancer. However, in many developing countries, there is a scarcity of medical facilities for such screening. Hence, research is needed in the field of data-driven diagnosis of cervical cancer. In this thesis, a dataset of cervical cancer patients has been considered, which includes attributes suitable for Bangladeshi patients. Another objective is to classify the patients of the dataset by using a new efficient hybrid algorithm. Firstly, an existing dataset collected from the University of California, Irvine (UCI); a machine learning repository is considered, which consists of 36 attributes and 858 instances. To overcome the imbalance of the data samples, the borderline Synthetic Minority Over-sampling Technique (SMOTE) is used. Next, a new dataset of cervical cancer patients collected from various hospitals in Bangladesh has been introduced. This new dataset consists of 21 attributes and 228 instances. The Recursive Feature Elimination method is applied to both datasets to find the most important attributing to cervical cancer. A number of classifiers, including base, ensemble, and hybrid algorithms, are applied to the datasets. Next, a two-stage hybrid algorithm is proposed where ExtraTreeClassifier is used in the first stage, and a stacking algorithm is used in the second stage. Results show that stacking as a combination of Random Forest, ExtraTreeClassifier, XGBoost, and Bagging exhibits the best classification accuracy of 95.3% for the first dataset. For the second dataset, AdaBoost shows the best classification accuracy of 95.6%. The proposed hybrid method offers classification accuracy of 95.9% and 96.2% for first and second datasets. Hence, the Bangladeshi dataset and the proposed hybrid algorithm can play an essential role in predicting cervical cancer.

CHAPTER 1
Introduction

1.1 Overview

Cervical cancer is a leading malignancy in terms of mortality and morbidity and the second deadliest cancer in women, next to breast cancer, and it is now thought that cervical cancer is curable in its early stages [1-2]. Cancer is a prominent cause of death, resulting in over 13% of all deaths worldwide, according to the World Health Organization (WHO) [3]. WHO also reported that cervical cancer is the fourth most common malignancy worldwide, with 5,70,000 new cases reported in 2018, contributing to 7.5 percent of all women's cancer losses. In low and middle-income nations, about 3,11,000 cervical cancer losses are reported each year, with an estimated 85 percent survival rate [4]. The rate of cancer incidence is increasing at an alarming rate. Despite the fact that cancer can be prevented and treated in its early stages, the great majority of people get a late diagnosis. Cervical cancer screening facilities are scarce in low and middle-income countries due to a lack of skilled and educated healthcare personnel and insufficient healthcare funds to fund screening systems [5]. Because of insufficient physical infrastructure and inadequate services, developing and underdeveloped countries face a significant challenge in treating cervical cancer. Access to equipment, uniformity of screening tests, sufficient supervision, and identification and treatment of lesions discovered are all factors that have defined screening effectiveness [5]. Despite significant medical and scientific advances, this disease is not totally curable, especially if it is discovered when the patient is still in a developing state. As a result, prevention and screening programs are critical in the war against cervical cancer. Cervical cancer screening follows a standard procedure: HPV testing, cytology or PAP smear testing, colposcopy, and biopsy. Several technologies were developed to support the workflow in order to make it more productive, practical, and cost-effective. The PAP smear image screening is most commonly used for the treatment of cervical cancer, but it requires a greater number of microscopic examinations to diagnose cancer and noncancer patients. It also requires time and trained professionals. Pap smears and HPV tests are both relatively expensive treatments with little sensitivity [6]. Hence, it is crucial to determine if a woman is at risk of having cervical cancer, considering the risk factors. A substitution incentive for doctors to increase patient diagnosis is the development of medical data collection. Practitioners have expanded their use of computer technology in recent years to strengthen decision-making support in the health care sector. Machine learning (ML) is becoming a crucial solution to assist patients' diagnoses. ML is an analytical technique for broad and complex tasks such as translating medical history into knowledge, pandemics predictions and studying genomic data [7-9]. ML algorithms [10-15] are considered for different disease predictions, including cervical cancer [13]. This research focuses on finding the most critical attributes associated with cervical cancer.

1.2 Motivation of the Thesis

In developing an increased risk of cervical cancer multiple risk factors are responsible. Those are Human papillomavirus (HPV), smoking, having a weakened immune system, chlamydia infection, long-term use of oral contraceptives, intrauterine devices (IUD), having multiple full-term pregnancies, diethylstilbestrol (DES), having a family history of cervical cancer, etc. [16-18]. Some signs and symptoms of cervical cancer are vaginal bleeding, unusual vaginal discharge, pelvic pain, etc; which are missing in UCI repository dataset. So, the new Bangladeshi dataset considering these related features have been collected. Results also vary for patients of different geographical regions, cultural backgrounds, ethnicity, etc. The accuracy of the results is reduced when the medical data is incomplete. Therefore, research is still needed to find the essential attributes and select the attributes to influence the disease prediction. This research focuses on finding the most critical attributes associated with cervical cancer

.

1.3 Objectives of the Thesis

The goal of this thesis is to predict cervical cancer using different attributes. The specific aims of the work are as follows:

i) To collect data regarding cervical cancer patients and to find the most influential attributes of the cancer.

ii) To introduce a new hybrid ML algorithm suitable for the prediction of cervical cancer using the two datasets.

iii) To find the prediction accuracy of cervical cancer with the use of well-known ML algorithms and the new hybrid algorithm.

1.4 Thesis Outline

The thesis is organized in the following manner

- Introduction: This chapter covers a wide range of issues such as the number of cervical cancer patients worldwide, losses per year due to cervical cancer, tests related to cervical cancer, scarcity of medical instruments in developing countries, etc. Moreover, how ML can assist in the prediction of cervical cancer is discussed. Lastly, the motivation and contribution of the thesis are discussed.
- Literature Review: In this chapter, previous work on the prediction of cervical cancer and comparison between them has been discussed. This chapter also covers the base and ensemble classifier used for this research.
- Dataset and Feature Selection: This chapter covers the overall view of all the features of the UCI and the Bangladeshi dataset, preprocessing of the dataset, how to handle the imbalanced dataset, feature selection method, etc.
- Prediction of Cervical Cancer and Proposed Hybrid Method: In this chapter, the proposed hybrid method and the experimental results of the different

base, ensemble algorithms are represented. The performance of the algorithm (e.g., accuracy, sensitivity, and specificity) was evaluated on both of the datasets.

- Conclusion: Finally, this chapter provides an overview of the thesis work as well as some recommendations for future research.

CHAPTER 2
Literature Review

2.1 Overview of Cervical Cancer

Cancer is a class of diseases involving irregular cell growth that can invade or spread via the development of a subset of neoplasms to other areas of the body. A neoplasm or tumor refers to the uncontrolled growth of cells that may often form a mass or lump (National Cancer Institute, 2015) [19]. Breast cancer, colorectal cancer, lung cancer, and cervical cancer are the most common forms of cancer in females, and the risk increases dramatically with age [20]. A type of cancer that occurs from the cervix is cervical cancer. It may invade or spread to other parts of the body because of the irregular growth of cells. No signs are usually seen in the earlier period. Subsequently, this form of cancer may pose a significant threat to cervical cancer-affected women. Cervical cancer is both the fourth most common cause of cancer worldwide and the fourth most common cause of cancer death in women (National Cancer Institute, 2015) [19]. In developed nations, about 70 percent of cervical cancers occur. Stewart et al. reported that 266,000 deaths were recorded in 528,000 cases of cervical cancer patients in 2012. They have also shown that it accounts for nearly 8% of the total cases and total cancer deaths [20]. Cervical cancer occurs from a ring of mucosa called the cervical transformation zone which is demonstrated in Figure 2.1 [21]. Persistent HPV infections cause cancers primarily in transformation zones between various types of the epithelium (e.g, cervix, anus, and oropharynx) [22]. In the development of cervical cancer, there are four main steps: infection of the metaplastic epithelium in the cervical transformation zone, viral persistence, progression to cervical precancer of the persistently infected epithelium, and invasion through the epithelial basement membrane.

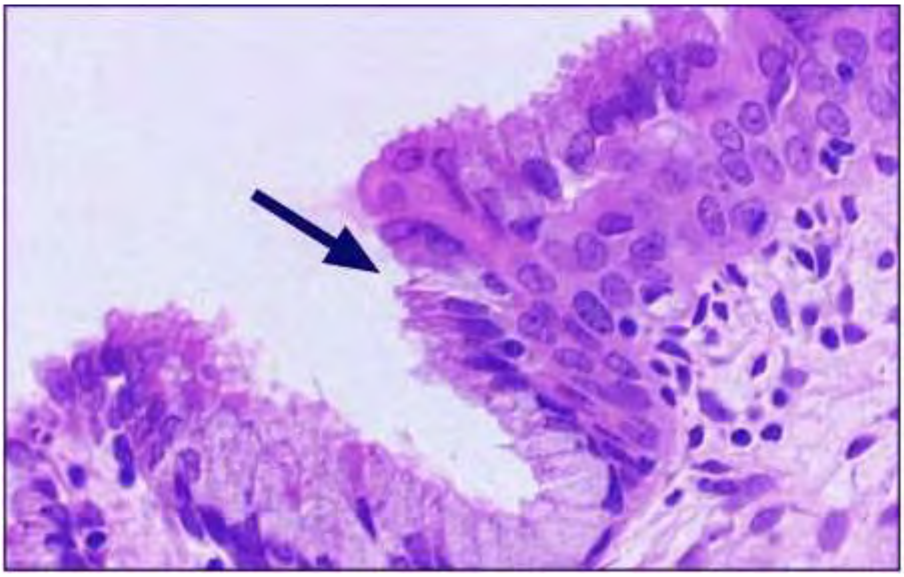


Figure 2.1: Cervical transformation zone [23].

2.1.1 Cervical Cancer and Human Papillomavirus

Zur Hausen for the first time reported the association between human papillomavirus (HPV) and genital cancer [24]. But prior to 1988, it gained little attention from researchers and epidemiologists. Several studies later established HPV infection as a sexually transmitted disease (STD) and a significant risk factor for the development of cervical intraepithelial neoplasia(CIN) and invasive cancer [25-30]. HPV is a group of viruses known to target the reproductive tract, which is highly prevalent worldwide. There are more than 150 forms of HPV, including at least 13 cancer-causing types, which are also known as high-risk types [31]. Cervical cancer is primarily acquired through the transmission of some forms of HPV.

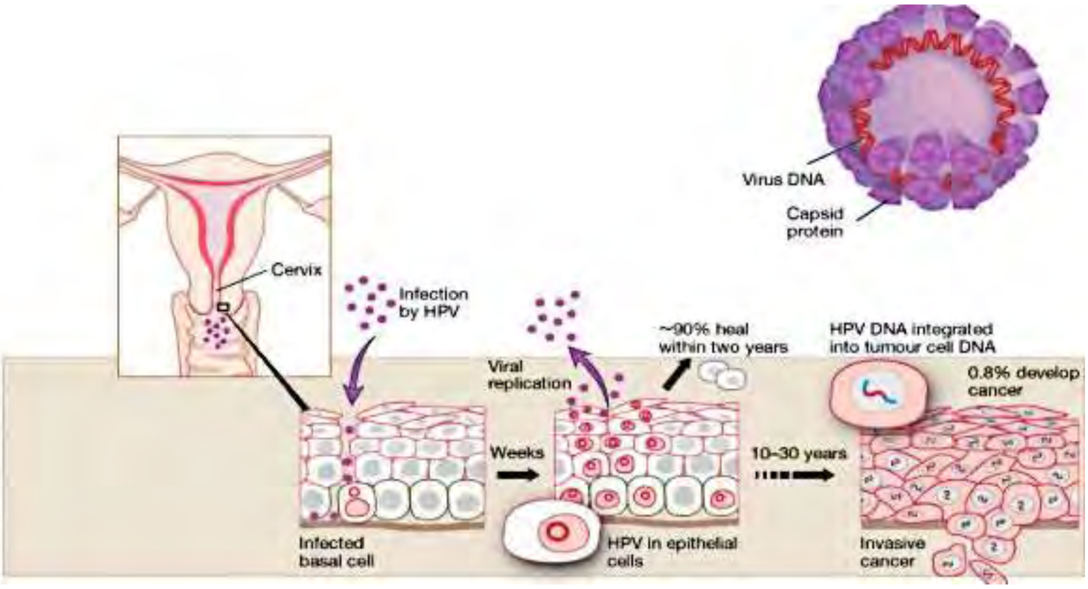


Figure 2.2: Mode of action of human papillomavirus in cervical cancer patients [32].

Women's immune systems normally prevent the HPV from causing significant harm when they are exposed to genital HPV, but in some women, the virus will live for years. As a consequence, the virus can transform ordinary cervical cells into cancerous cells. While cells may only display signs and symptoms of viral infection, there is later development of a precancerous stage called cervical CIN. While this stage typically spontaneously goes away, invasive cervical cancer may progress [33]. The mode of action of HPV in cervical cancer patients has been shown in Figure 2.2.

2.2 Previous Work on Prediction of Cervical Cancer

A lot of work has been done for the prediction of cervical cancer with different types of the dataset available for cervical cancer. A lot of researches have worked on the prediction of cervical cancer; Tseng et al. used three ML approaches, including C5.0, support vector machine (SVM), and extreme learning machine (ELM), to determine the most significant attributes that may influence cervical cancer [14]. William et al. reported that the images obtained from the screening procedure known as the Pap test, which consists of five stages, play a vital role in the early detection of cervical cancer [15]. Sharma et al. performed the k-nearest neighbors (KNN) method to classify the cervical cancer stage from pap-smear images [34]. However, Nithya and Ilango considered recursive feature elimination (RFE), Boruta algorithm, simulated annealing (SA), etc., approaches for choosing the optimal subset of features on a dataset of 858 patients with 36 attributes for the prophecy of cervical cancer [35]. Moreover, several ML algorithms, including C5.0, random forest (RF), recursive partitioning and regression trees (RPART), SVM, and KNN, were applied [35]. Lu et al. proposed a novel ensemble method, a voting strategy to classify cervical cancer more accurately. Their proposed method was reported to be more scalable and practical [36]. A cervical cancer prediction system was proposed using the concept of a convolutional neural network (CNN) [37]. Fatlawi used cost-sensitive classifiers for the classification of cervical cancer patients [38]. In one study, the authors found the five most influential risk factors, and three classification algorithms were performed to classify cancer patients [39]. Al-Wesabi et al. applied three techniques undersampling, oversampling, and combination of both methods on the cervical cancer UCI dataset [40]. As the data are imbalanced, it was found that oversampling gives a better result than the other two methods as higher accuracy is obtained by over-sampling. Gauss-newton representation based algorithm (GNRBA) was proposed by Rustam et al. [41], providing a classification accuracy of 93%. A. H. Gharekhan et al. showed that principal component analysis (PCA) is an excellent approach for highlighting correlations between spectral aspects of various human cervical spectra which is useful to distinguish between diseased and non-diseased tissue types in the preliminary experiment [42]. S. Devi et al. detected cervical cancer advancement through a polarization-based technique [43]. Mukhopadhyay, S. et al. [44] classified precancerous stages of cancer in the cervix with an initial dataset of 35 samples. The study in [44] applied multifractal tissue optical properties with a hidden markov model (HMM). The work by Wei, L. [45] reported that for the case of nonlinear and complex datasets of cervical cancer, ML was useful in predicting cancer. A probabilistic neural network (PNN) was reported by Obrzut, B. [46] to have a ROC value of 0.809 in predicting cervical cancer, which was greater than that of logistic regression and decision tree algorithms. Rahaman, M et al. provided a comprehensive review on deep learning for the processing of cervical cancer cytology images [47]. Furthermore, Li, Y et al. proposed a deep learning framework for the determination of cervical cancer using time-lapsed colposcopic images on 7668 samples [48]. A comparison of relevant research works on cervical cancer is shown in Table 2.1. In the literature, base classifiers were mostly reported for predicting cervical cancer. A diverse set of models have the potential to provide better prediction results compared to individual models. Hence, this thesis work proposes a hybrid algorithm based on two individual models. It is later shown in the paper that the hybrid model has better reliability than individual base classifiers in predicting cervical cancer. Results also vary for patients of different geographical regions, cultural backgrounds, ethnicity, etc.

Table 2.1: Comparison of relevant research works.

Datasets	**ML Model**	**Specific Work on Cervical Cancer**	**Accuracy/True Positive Rate**	**Reference**
Chung Shan Medical University Hospital Tumor Registry	C5.0, SVM, and ELM	Predicting the recurrence-proneness for cervical cancer	96.0% using the C5.0 model	[14]
UCI ML Repository	RF, C5.0, RPART, KNN, SVM	Optimized approaches to feature selection and methods of classification for prediction of cervical cancer	100% using C5.0 and RF with selected features.	[35]
UCI dataset and private	Logistic regression(LR),	An ensemble approach	83.16% using a voting method	[36]

Datasets	ML Model	Specific Work on Cervical Cancer	Accuracy/True Positive Rate	Reference
dataset	SVM, MLP, KNN, Decision Tree(DT), A voting method combining these five			
Herlev database	CNN-ELM, AE, MLP	Classification using CNN and ELM	99.5% using CNN-ELM	[37]
UCI ML Repository	DT, cost-sensitive classification	Enhanced classification model based on cost-sensitive classifier	0.429 (TP rate) using cost-sensitive classification	[38]
UCI ML Repository	XGBoost, SVM, and RF	Analysis of risk factors for cervical cancer	99.49% for target variable Hinselmann using RF	[39]
UCI ML Repository	Gaussian naïve bayes(GNB), KNN, DT, LR, and SVM	Classification of cervical cancer dataset	100% (TP rate) using DT	[40]
UCI ML Repository	GNRBA	Optimal cervical cancer classification using GNRBA.	93.0%	[41]

Datasets	**ML Model**	**Specific Work on Cervical Cancer**	**Accuracy/True Positive Rate**	**Reference**
UCI ML Repository	Naïve Bayes (NB), C4.5, KNN, sequential minimal optimization (SMO), RF, multilayer perceptron (MLP) neural network, simple logistic regression (SLR)	RF gives the best accuracy among the used algorithm where Biopsy is used as the target variable.	96.40%	[49]
UCI ML Repository	SVM, SVM-RFE, SVM-PCA	SVM-PCA gave the best performance with 11 features while the biopsy is the target variable.	94.03%	[50]
UCI ML Repository	DT, SVM, RF, KNN, NB, MP, J48 Trees, and LR.	Use of majority voting algorithms to get better results	94%	[51]
UCI ML Repository	RF	Prediction of cervical cancer using RF with	97.6% target variable-Hinselmann	[52]

Datasets	ML Model	Specific Work on Cervical Cancer	Accuracy/True Positive Rate	Reference
		SMOTE and feature reduction techniques	using SMOTE-RF	
UCI ML Repository	RF	Cervical cancer identification with SMOTE and PCA using RF	96.06% using RF-PCA	[53]

2.3 Classification Algorithm

2.3.1 Base Level Classifier

There are various base classifiers in ML that can be utilized with ensemble ML. The next subsections go over some of the basic classifiers.

2.3.1.1 Support Vector Machine

Support Vector Machines [54-55] are based on the idea of decision planes specifying the boundaries of decisions. A decision plane is one that distinguishes a group of objects with distinct memberships in the class. If there is no linear separability of the results, then a kernel trick is used. Functions that measure similarities between observations are kernels. Polynomial kernels, radial basis kernels, and linear kernels are common types of kernels used to distinguish non-linear data [56-58].

2.3.1.2 Logistic Regression

Logistic regression was first introduced as a statistical method by Statistician D.R. Cox in 1958, and then commonly used in many fields. It became routinely available in statistical packages in the early 1980s [59]. The relationship between a dependent variable and one or more independent variables is discussed by logistic regression. It provides a modeling method for a binary response variable that takes values of 1 and 0.

2.3.1.3 K-Nearest Neighbor

K nearest neighbors (KNN) [60-61] is a simple algorithm that stores all available cases and categorizes new ones using a similarity metric (e.g., distance functions). KNN is a lazy learning algorithm that is non-parametric. It aims to predict the classification of a new sample point using a database of data points divided into several groups.

2.3.1.4 Random Forest Classifier

Random Forest (RF) is a well-known supervised classification technique that has been applied to a number of classification problems [62–64]. Breman proposed RF, also known as bagged decision trees, in 2001 [65-66]. It's an ensemble technique [67] that operates on the idea of forming a strong learner from a group of weak learners. In RF, each tree creates an independent decision tree by randomly selecting a subset of the dataset. RF repeatedly splits the selected random subset from the root node to a child node until each tree reaches a leaf node without being pruned. Each tree independently classifies the features and the target variable before voting on the final tree class. The final overall classification is determined by RF based on the majority of votes received from the trees.

2.3.1.5 Decision Tree Classifier

A dataset is used to train a decision tree, which is one of the most widely used ML approaches for classification and regression analysis. Based on a series of questions, this model divides the samples into several classes. The classification process is similar to that of a tree. All samples are grouped together at the tree's root. The process of creating a tree is recursive. The decision tree's main challenge is determining the best partition attributes, which can be explained using information entropy, gain ratio, or Gini index.

2.4 Ensemble Classifier

An ensemble classifier is a supervised learning method because it can be trained and then tested to make predictions. Ensemble classifiers aim to improve the accuracy of predictive models to improve their performance. Ensemble Learning is a method of solving a problem by systematically constructing distinct ML models (such as classifiers). The art of mixing a diverse range of learners (individual models) to improve the model's reliability and predictive capability is known as ensemble learning.

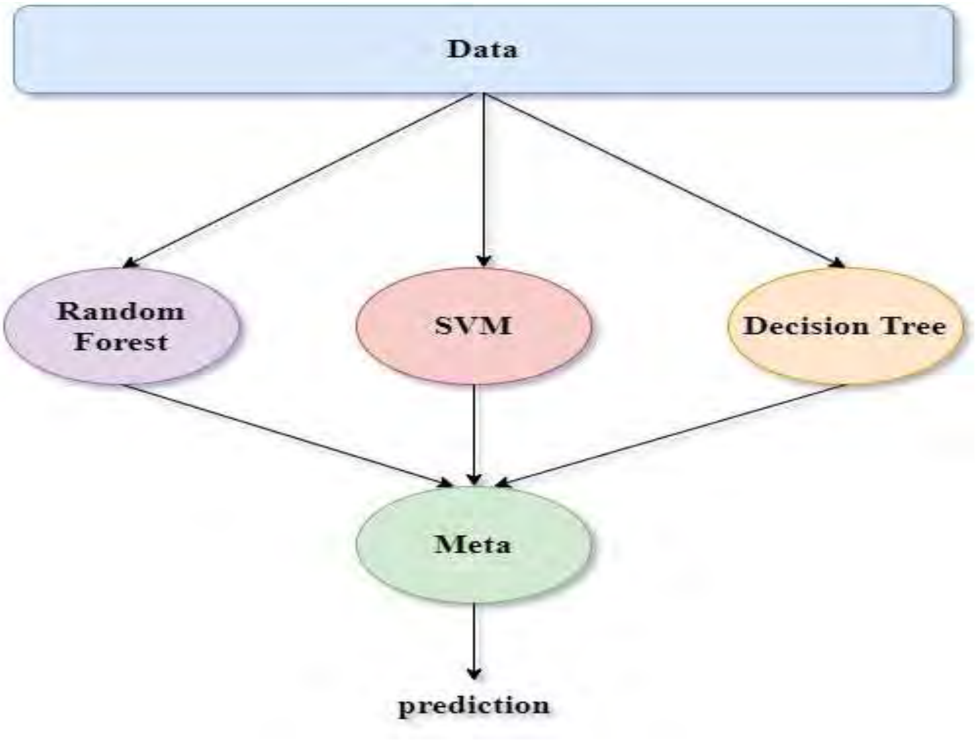


Figure 2.3: Typical prediction by ensemble classifier.

Ensemble learning refers to the process of combining all of the predictions as shown in Figure 2.3

2.4.1 Types of Ensemble Classifier

Although there are a variety of Ensemble learning methods, the four listed below are the most commonly utilized in the field.

2.4.1.1 Bagging

Bagging, commonly known as Bootstrap Aggregation, is one of the Ensemble creation strategies. The basis of the Bagging approach is established by Bootstrap. Bootstrap sampling is a method of selecting "n" observations from a population of "n" observations. However, the selection is completely random, in the sense that each observation can be chosen at random from the initial population, with each observation having an equal chance of being chosen in each iteration of the bootstrapping process. Following the formation of the bootstrapped samples, distinct models are trained using the bootstrapped samples. The bootstrapped samples are selected from the training set in real trials, and the sub-models are tested using the testing set. All of the sub-models projections are merged to provide the final output prediction. Figure 2.4 gives a brief idea of Bagging:

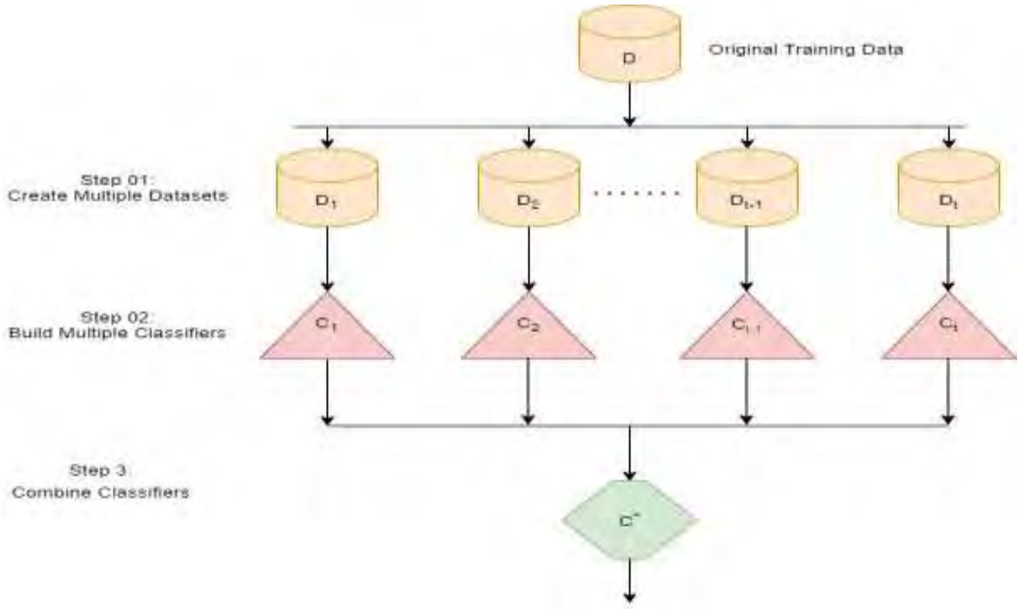


Figure 2.4: Prediction by bagging.

2.4.1.2 Boosting

Boosting is a strategy for sequential learning. The approach works by training a model with the whole training set and then fitting the residual error values of the initial model to create subsequent models. Boosting aims to provide more weight to data that were under-estimated by the preceding model in this way. After the models' sequences have been constructed, the models' predictions are weighted by their accuracy ratings, and the results are combined to form a final estimate. XGBoost (Extreme Gradient Boosting), GBM (Gradient Boosting Machine), AdaBoost (Adaptive Boosting), and other models are commonly employed in the Boosting methodology. Among them, XGBoost is widely used in algorithm competitions, where it outperforms other algorithms. On a single computer, the device is more than ten times faster than existing common solutions, and it scales to billions of examples in distributed or memory-constrained environments [68]. Boosting is an iterative strategy for adjusting an observation's weight based on the previous classification. It seeks to raise the weight of observation if it was classified erroneously, and vice versa. Boosting is a sequential strategy in which the first algorithm is trained on the complete data set, and the succeeding algorithms are generated by fitting the first algorithm's residuals, thus assigning more weight to those observations that were poorly predicted by the prior model. It is based on the creation of a sequence of weak learners, each of which is good for a portion of the data set but not the complete data set. As a result, each model improves the ensemble's performance. Figure 2.5 gives a brief idea of boosting.

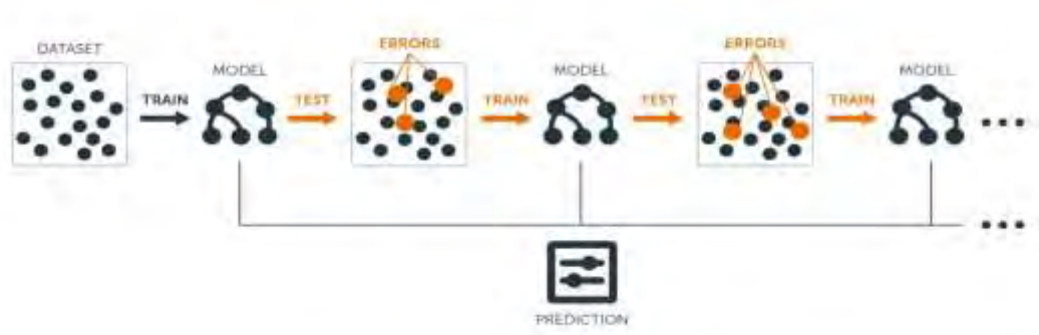


Figure 2.5: Prediction by boosting [69].

2.4.1.3 Voting

Voting is one of the most basic Ensemble learning approaches for combining predictions from several models. The procedure begins with the creation of two or more independent models using the same dataset. Then, to wrap the prior models and aggregate their predictions, a Voting-based Ensemble model might be employed. The Voting based Ensemble model can then be used to create predictions based on new data. Weights can be assigned to the sub-models' predictions. Stacked aggregation is a technique that can be used to figure out how to best weight these predictions. Figure 2.6 gives an idea of Voting-based Ensembles:

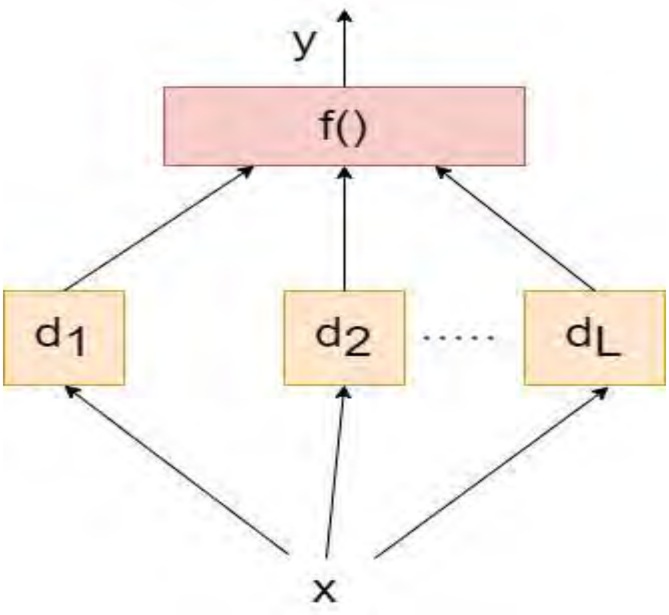


Figure 2.6: Prediction by voting.

For voting classifiers, a different combination of DT, KNN, LR, and SVM as base level classifiers are used to design 4 different voting classifiers.

2.4.1.4 Stacking

Stacking is an ensemble learning strategy that uses a meta-classifier or a meta-regressor to integrate numerous classification or regression models. The meta-model is trained on the outputs of the base level model-like features, after which the base level models are trained on a complete training set. Because multiple learning techniques are frequently used at the base level, stacking ensembles are frequently diverse. Figure 2.7 gives an idea of Stacking Ensembles:

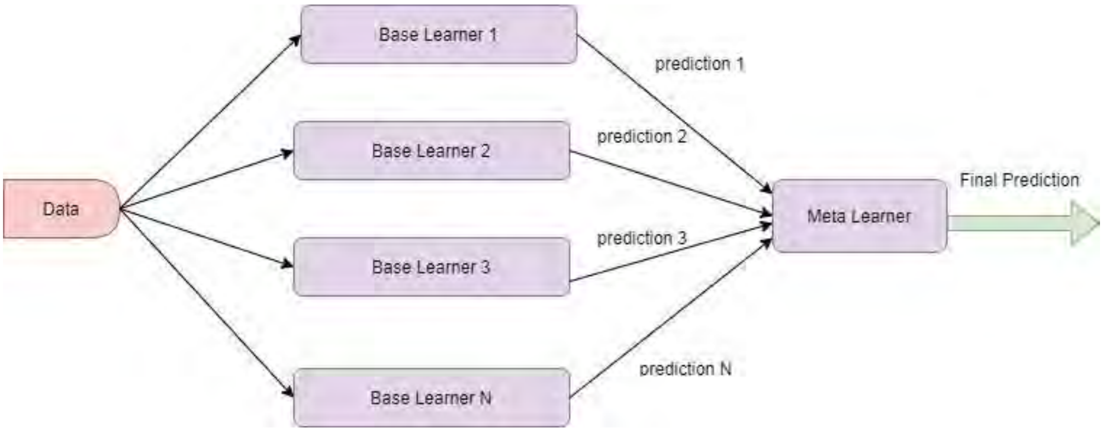


Figure 2.7: Prediction by stacking.

In the vast majority of circumstances, the ensemble of models will outperform the individual models on test case scenarios (unseen data). The combined result of several models is always less noisy than the results of the individual models. As a result, model stability and robustness are achieved.

2.5 Summary

In this chapter, the basic knowledge about cervical cancer, human papillomavirus, base, and ensemble classifiers such as bagging, boosting, voting, and stacking have been discussed. Also, previous works on cervical cancer and comparison have been done for better visualization.

CHAPTER 3

Dataset and Feature Selection

3.1 Introduction

In this chapter, the attributes and number of instances of both the dataset will be discussed. The preprocessing of the dataset will be done by removing the rows and columns having maximum missing values. Then BorderlineSmote technique will be applied to handle the imbalanced dataset. After doing all the preprocessing recursive feature elimination (RFE) method will be used to select the best number of features. In this chapter, some evaluation matrices will be discussed also.

3.2 Description of the UCI Machine Learning Repository Dataset

This research is based on two datasets. The first dataset is from the UCI Machine Learning Repository [70], and the second one is the collected Bangladeshi dataset. The description of the first dataset is shown in Table 3.1. The first dataset consists of (0-35) 36 attributes and 858 instances. As a target variable, a new column named 'Final target' by computing the OR operation among four test result labels Hinselmann, Biopsy, Schiller, and Cytology is used. Hinselmann is a way of studying cervical cancer by analyzing cells on a microscope called a colposcope. A biopsy is a medical procedure to determine the presence or extent of disease, extraction to collect sample cells or tissues. Schiller is an experiment that uses iodine added to the cervix. When the iodine is brown, it suggests a good hue cell; otherwise, when iodine is white or yellow, it remains unstained, indicating irregular cells. Cytology is a condition of the cervix, a cell taken from a microscopic pap-smear test.

Table 3.1**:** Feature information of the UCI repository cervical cancer dataset

Feature No.	**Attribute Name**	**Type**	**Range of Values**
0	Age	Int	13-84
1	Number of sexual partners	Int	1-28
2	First sexual intercourse (age)	Int	10-32
3	Num of pregnancies	Int	1-11
4	Smokes	Bool	0,1
5	Smokes (years)	Bool	0,1
6	Smokes (packs/year)	Bool	0,1
7	Hormonal Contraceptives	Bool	0,1
8	Hormonal Contraceptives (years)	Int	0-30
9	Intrauterine Device (IUD)	Bool	0,1
10	IUD (years)	Int	0-19
11	Sexually Transmitted Diseases (STDs)	Bool	0,1
12	STDs (number)	Int	0
13	STDs: condylomatosis	Bool	0,1
14	STDs: cervical condylomatosis	Bool	0,1
15	STDs: vaginal condylomatosis	Bool	0,1
16	STDs: vulvo-perineal condylomatosis	Bool	0,1
17	STDs: syphilis	Bool	0,1
18	STDs: a pelvic inflammatory disease	Bool	0,1
19	STDs: genital herpes	Bool	0,1
20	STDs: molluscum contagiosum	Bool	0,1
21	STDs:AIDS	Bool	0,1
22	STDs: HIV	Bool	0,1
23	STDs: Hepatitis B	Bool	0,1
24	STDs:HPV	Bool	0,1
25	STDs: Number of diagnoses	Int	0-3
26	STDs: Time since the first diagnosis	Int	1-22

Feature No.	Attribute Name	Type	Range of Values
27	STDs: Time since the last diagnosis	Int	1-22
28	Dx: Cancer	Bool	0,1
29	Dx: Cervical Intraepithelial Neoplasia (CIN)	Bool	0,1
30	Dx: Human Papillomavirus (HPV)	Bool	0,1
31	Diagnosis: DX	Bool	0,1
32	Hinselmann: target variable	Bool	0,1
33	Schiller: target variable	Bool	0,1
34	Cytology: target variable	Bool	0,1
35	Biopsy: target variable	Bool	0,1

3.3 Description of the Bangladeshi Dataset

There are many government and non-government hospitals for cancer patients, such as the National Institute of Cancer Research and Hospital, Ahsania Mission Cancer Hospital, etc. However, due to the hospitals' privacy policy, it is not possible to collect data from everywhere. So, data regarding cervical cancer has been collected with the help of Department of Biochemistry and Molecular Biology at the University of Dhaka. The data collection has been done from the patients of Dhaka Medical College (DMC) Hospital, Shaheed Suhrawardy Medical College and Hospital (SSMC), Bangladesh Cancer Society Hospital & Welfare Home in Bangladesh with patients’ consents. This is because patients of this institution are more accessible than other hospitals. At the time of collecting data, many difficulties have been faced, so the patients are informed about the aim and advantage of the research. They were informed about their rights to withdraw from the study at any time. They were informed that their contribution would maximize the benefits of human welfare. A structured questionnaire for the collection of data is used. The first and second page of the sample questionnaire (written in Bangla language) is shown in Appendix. The results obtained from a part of the questionnaire formed the Bangladeshi or the second dataset. This is given in

Table 3.2. The second dataset consists of (0-20) 21 attributes and 228 instances. Visual inspections with acetic acid (VIA) test are considered as the target variable for this dataset.

Table 3.2: Feature information of the Bangladeshi dataset

Feature No.	**Attribute Name**	**Type**	**Range of Values**
0	Age	Int	20-77
1	Marital Status	Bool	0,1
2	Children	Int	0-10
3	Previous exposure to cancer	Bool	0,1
4	Other chronic diseases	Bool	0,1
5	Family history of cancer	Bool	0,1
6	H/O operation	Bool	0,1
7	Chemotherapy	Bool	0,1
8	Radiotherapy	Bool	0,1
9	Menstrual period Regular	Bool	0,1
10	Menstrual period Irregular	Bool	0,1
11	Postmenopausal	Bool	0,1
12	Whitish secretion	Bool	0,1
13	Bad odor	Bool	0,1
14	Oral Contraceptive Pill (OCP)	Bool	0,1
15	Contraceptive Injection	Bool	0,1
16	Intrauterine contraceptive device (IUCD)	Bool	0,1
17	Use of Condom	Bool	0,1
18	Cancer's Grade	Int	1,2,3
19	Cancer Stage-	---------	---------
20	Visual inspections with acetic acid (VIA Test)	Bool	0,1

3.4 Preprocessing of the Dataset

In the UCI repository dataset, two columns named ‘time since first diagnosis’ and ‘time since the last diagnosis’ had excessive missing values. So, those were deleted to reduce the impact of missing values. Besides, the rows which have missing values are also deleted to ensure the accuracy of the results. The lack of four target attributes, in the beginning, complicates the classification task. The construction of attributes provides a solution to the problem by creating a new attribute from other attributes' information. By using logical OR operation among the four target variables (Hinselmann, Schiller, Cytology, and Biopsy), a single column named 'Final target' has been created that represents whether the person has cancer or not. Finally, after performing all this preprocessing, the first dataset contains 668 rows and 35 columns.

In the Bangladeshi dataset, all the ‘Yes’ and ‘No’ values were converted into 1 and 0, respectively. The mean value of the columns replaced the missing values of the dataset. Columns, named ‘Cancer Stage’, ‘GRADE’, ‘chemotherapy’,’ radiotherapy’, and ‘H/O operation’, were dropped as those were not relevant to our analysis. Finally, the Bangladeshi dataset contains 228 rows and 16 columns.

3.5 Handling the Imbalanced Dataset

The UCI dataset [70] is imbalanced as the numbers of cancer and non-cancer patients are 86 and 582, respectively. Before analyzing the dataset, the impact of the imbalanced data has to be reduced. For addressing the imbalanced data, SMOTE is one of the over-sampling techniques [71]. It uses the minority class sample to effectively handle the manual sample production and distribution, which can effectively manage the overfitting induced by poor decision intervals. But, if outlying observations in the minority class occur in the majority class, it is a problem for SMOTE because it creates a line bridge with the majority class. So, to solve this problem Borderline-SMOTE has been introduced [72]. In this research, Borderline-SMOTE is applied to handle the imbalanced UCI dataset. The minority class observations are first classified using this approach. If all of the neighbors are in the majority class, it identifies any minority observation as noise and ignores it while generating synthetic data. Furthermore, it resamples entirely from a few border places that include both majority and minority classes as neighborhoods. There is no need to use the oversampling method in the Bangladeshi dataset as the data is not imbalanced

3.6 Feature Selection

The selection of features is the technique of selecting some insightful and appropriate features from a larger set of features that produce a much better characterization of multi-class patterns. There are a variety of feature selection techniques, including filtering-based methods, wrapper methods, and embedded methods. The filtering based method is further divided into information gain, chi-square test, fisher's score, correlation coefficient, variance threshold, mean absolute difference, dispersion ratio. From all these methods, information gain, correlation coefficient are used in this research work. The wrapper method is further divided into three groups-Forward Selection, Backward Elimination, and Recursive Feature Elimination (RFE). Abdoh et al. used RFE and PCA method for feature selection in cervical cancer diagnosis using RF classifier with SMOTE and feature reduction techniques [52,73]. The RFE proposed by Guyon et al. is also used to identify cancer patients' important features. It was also used in the microarray of genes where thousands of features were counted. For the identification of variable importance, RFE is also used with RF [74]. It is a greedy algorithm that seeks to find out the simplest subset of performing functions. It repeatedly creates models and, at each iteration, keeps aside the simplest or worst performing feature. Until all the features are depleted, it builds a subsequent model with the left features. It then lists the characteristics based on the order of their removal.

Firstly, the relations between the features are analyzed using a correlation matrix from the filtering-based method. This matrix depicts the relationship between all properties in a color palette. The correlation matrix has values ranging from -1 to +1. Negative correlations are viewed as values close to -1, whereas positive correlations are interpreted as values close to +1. If the value is near 0, it means that these two variables have no relationship. The variables which have a close value around +1 in aspect to target variable have higher feature importance. For the UCI machine learning repository dataset, the correlation between all the features is shown in Figure 3.1. From the figure, it is shown that the features IUD, IUD(years), Number of pregnancies, Smoke, etc have 0.09,0.08, 0.49, and 0.46 correlation respectively in the aspect of the target variable named “final target” which means these features are important for training the model. Similarly, for the Bangladeshi dataset, Figure 3.2 depicts the correlation between all the features. From this figure, it is shown that the features whitish secretion, bad odor, OCP, use of condom, children, etc have 0.71,0.67, 0.27, 0.44, and 0.32 correlation respectively in the aspect of the target variable named “VIA TEST” which means these features are important for training the model.

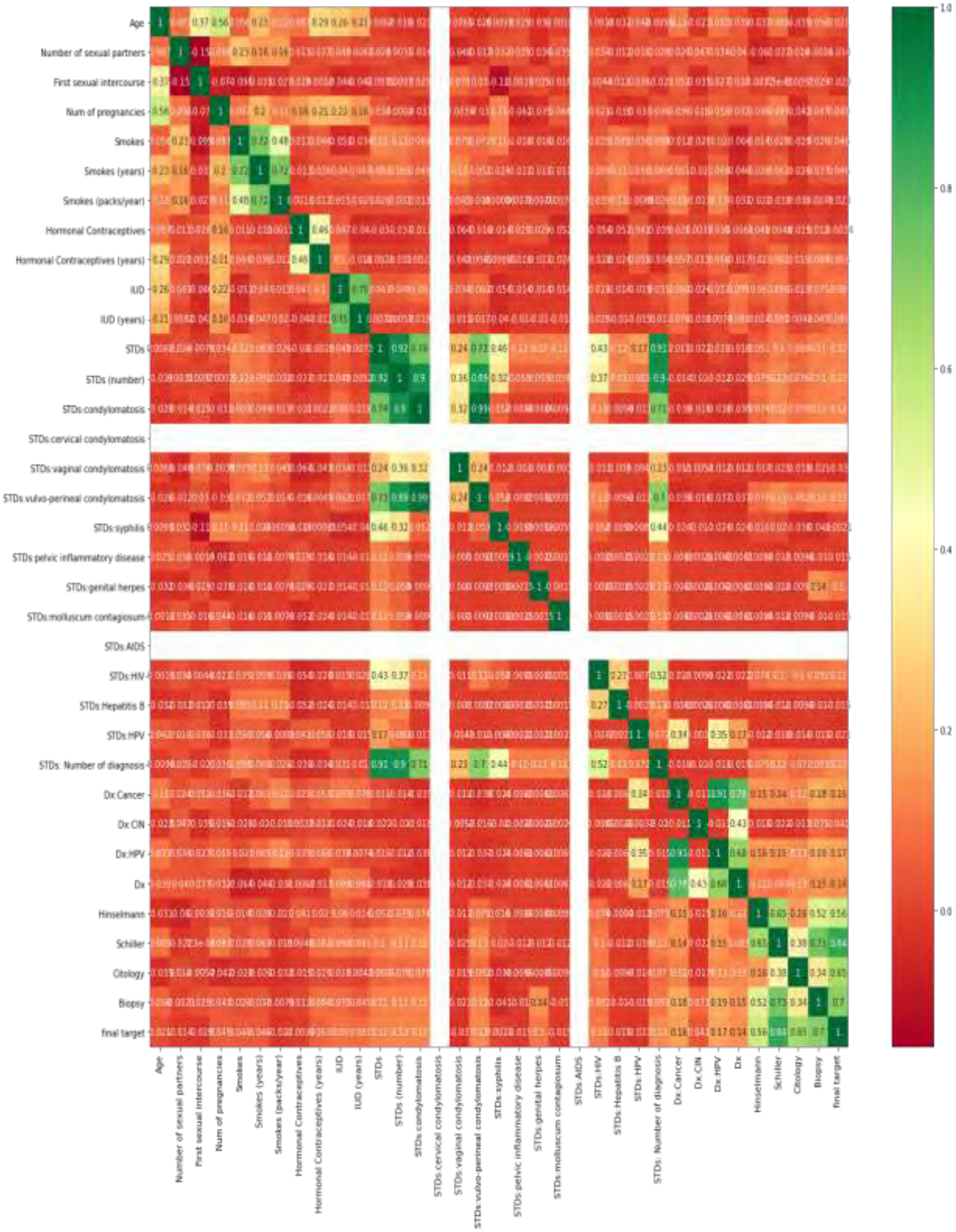


Figure 3.1: Correlation between all features of UCI dataset.

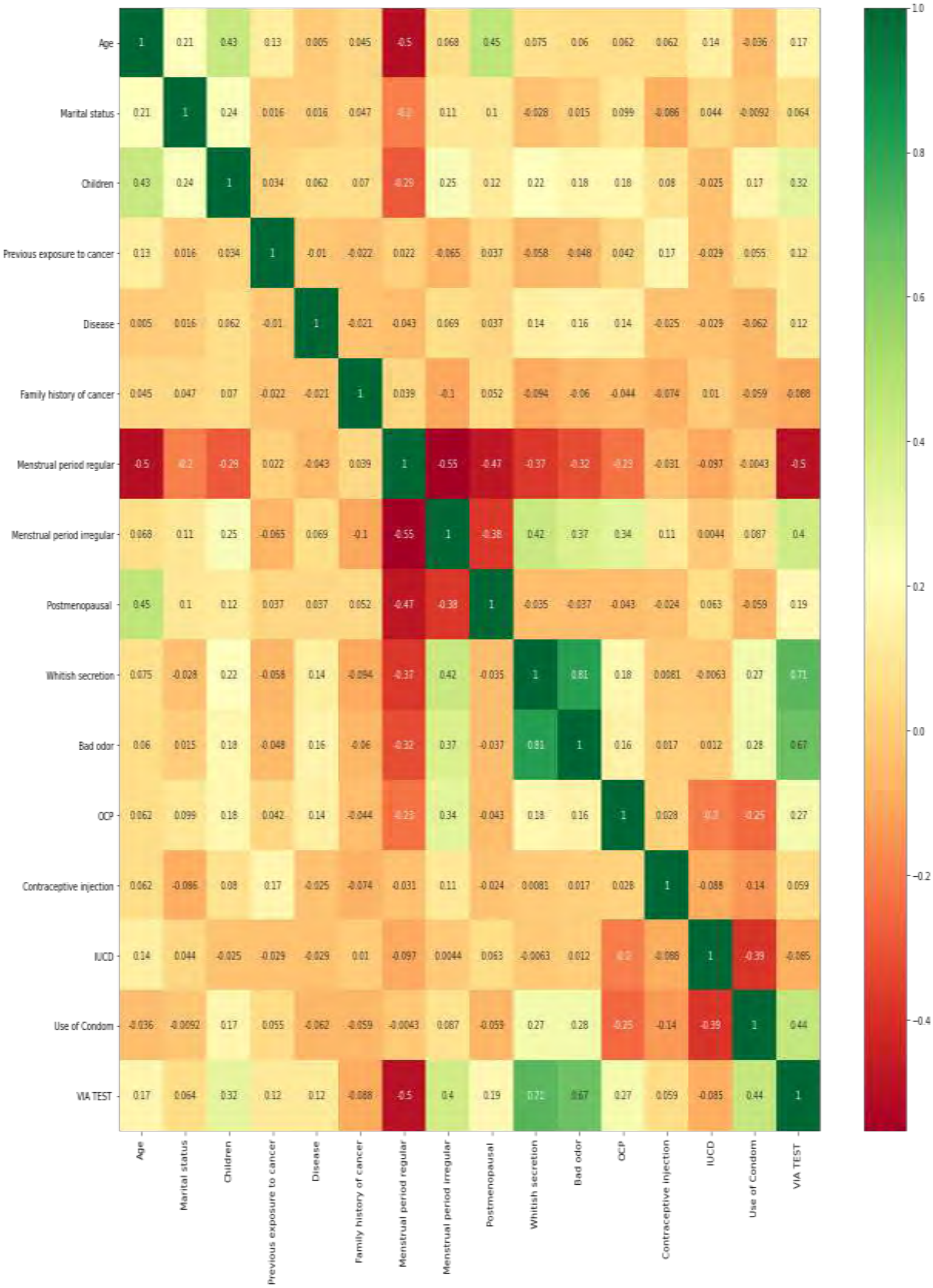


Figure 3.2: Correlation between all features of Bangladeshi datasets.

Feature ranking of all UCI features is demonstrated in Table 3.3. From Table 3.3 it is depicted that feature no 0 has a higher ranking than feature no 1 based on their feature importance value. For example, Feature no 8 has feature importance 0.112172 which is greater than the feature importance 0.1111933 of feature no 3. So Feature no 8 is ranked as 5, and feature no 3 is ranked as 6. The rest of the table is ranked following the above procedure.

Table 3.3: Feature Ranking of UCI dataset

Features Ranking	Feature No.	Feature Importance
1	0	0.134082
2	1	0.131963
3	2	0.123380
4	7	0.115748
5	8	0.112172
6	3	0.111933
7	9	0.040135
8	4	0.036862
9	5	0.028545
10	10	0.027839
11	6	0.025855
12	11	0.020444
13	12	0.016275
14	16	0.015874
15	13	0.012883
16	25	0.010088
17	22	0.007895
18	17	0.007031
19	28	0.005422
20	26	0.004672

Features Ranking	Feature No.	Feature Importance
21	19	0.003823
22	29	0.002576
23	15	0.002426
24	27	0.000764
25	24	0.000576
26	20	0.000307
27	23	0.000281
28	18	0.000147
29	21	0.000000
30	14	0.000000

The feature ranking of all Bangladeshi datasets is demonstrated in Table 3.4. From Table 3.4, it is depicted that feature no 9 has a higher ranking than feature no 10 based on their feature importance value. For example, Feature no 14 has feature importance 0.105441 which is greater than the feature importance 0.103388 of feature no 6. So Feature no 14 is ranked as 3, and feature no 6 is ranked as 4. The rest of the table is ranked following the above, method.

Table 3.4: Feature Ranking of Bangladeshi datasets

Feature Ranking	Feature No.	Feature Importance
1	9	0.235784
2	10	0.170992
3	14	0.105441
4	6	0.103388
5	16	0.079096
6	11	0.042203
7	2	0.038654
8	4	0.038491
9	15	0.035303

Feature Ranking	Feature No.	Feature Importance
10	0	0.034276
11	7	0.033149
12	8	0.030606
13	3	0.025150
14	1	0.011416
15	12	0.009406
16	13	0.004869
17	5	0.001775

The feature importance of all the UCI attributes is shown in Figure 3.3. Where it is shown that the importance of feature 0 is higher than the importance of feature 1. Similarly, feature 2 is of higher importance than feature 7.

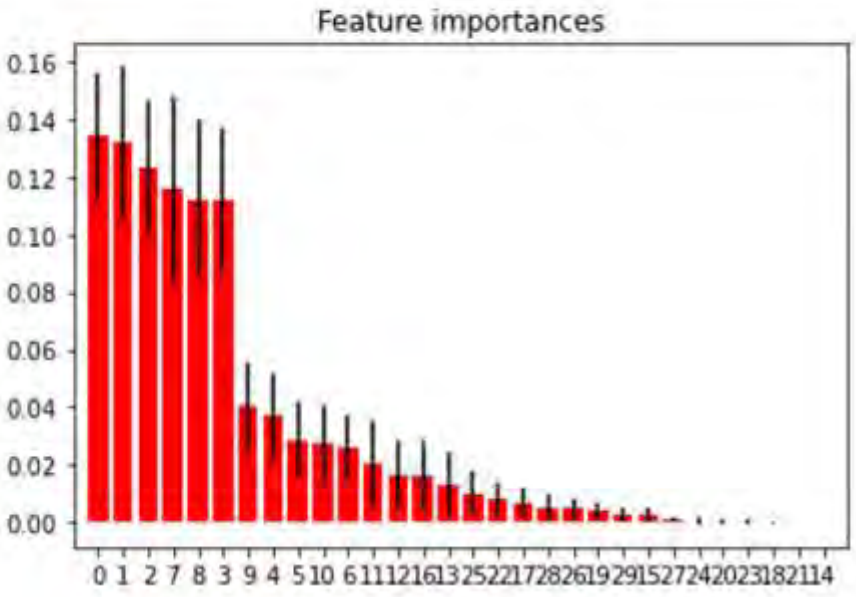


Figure 3.3: Features importance of all attributes of UCI dataset.

The feature importance of all the Bangladeshi attributes is shown in Figure 3.4. Where it is shown that the importance of feature 9 is higher than the importance of feature 10. Similarly, feature 14 is of higher importance than feature 6.

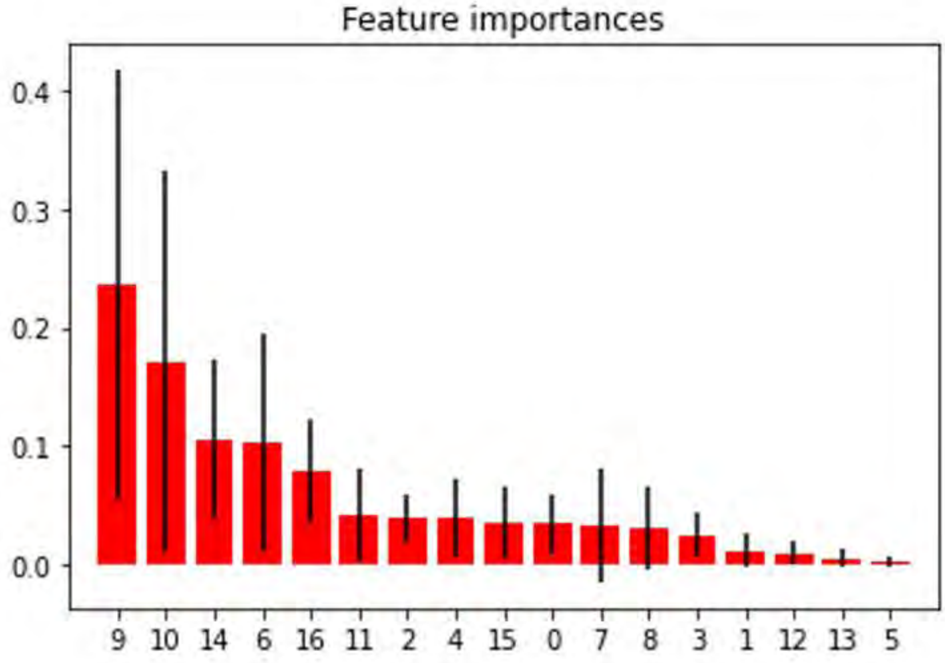


Figure 3.4: Features importance of all attributes of Bangladeshi datasets.

Finally, the RFE method is used to determine the best number of features to get the highest accuracy. In Table 3.5, the highest accuracy 93.5% is obtained when the number of features is 10. So for both of the datasets, the number of features is chosen 10.

Table 3.5: Accuracy using the RFE method based on different numbers of features.

Number of Features	Accuracy using RFE method
3	0.862
4	0.892
5	0.913
6	0.920
7	0.924
8	0.924
9	0.930
10	0.935
11	0.933
12	0.931
13	0.926
14	0.931
15	0.930
16	0.930
17	0.930
18	0.925
19	0.925
20	0.931
21	0.932
22	0.933
23	0.932
24	0.929
25	0.933
26	0.930
27	0.930
28	0.930
29	0.930
30	0.932

The selected features include: 'Age', 'Number of sexual partners', 'First sexual intercourse', 'Num of pregnancies', 'Smokes (years)', 'Smokes (packs/year)', 'Hormonal Contraceptives', 'Hormonal Contraceptives (years)', 'IUD (years)', 'STDs (number)'. For different types of algorithm evaluation, these ten features are used. For UCI dataset the ten best-selected features are shown in Figure 3.5.

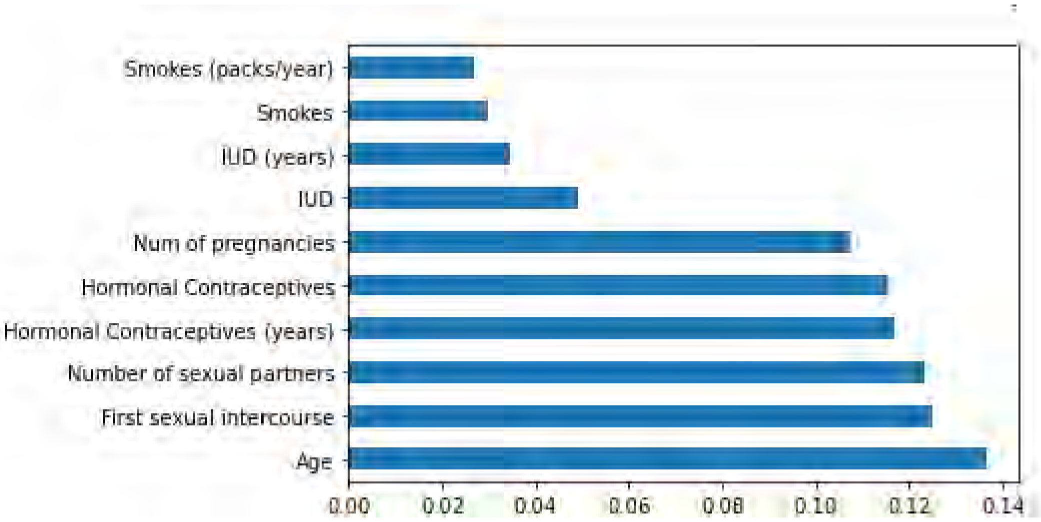


Figure 3.5: Feature importance of ten best-selected attributes of UCI dataset.

For the Bangladeshi dataset, the number of best features is also 10, which are ‘Age’, ‘Children’, ‘Previous exposure to cancer’, ‘Disease’, ‘Menstrual period regular’,’ Menstrual period irregular’, ‘Whitish secretion’, ‘Bad odor’, ‘OCP’, ‘Use of Condom’. These ten best-selected features are shown in Figure 3.6.

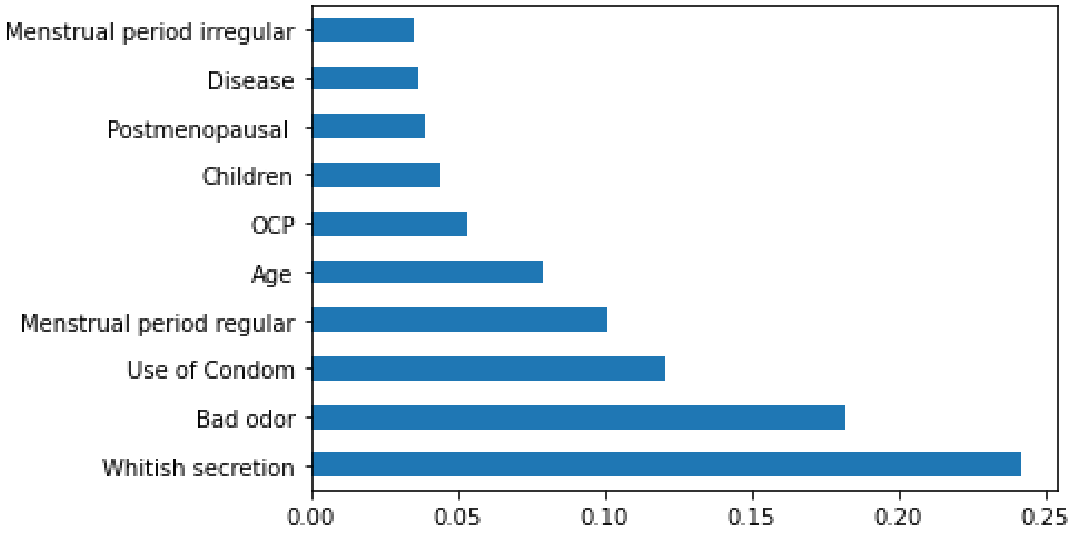


Figure 3.6: Feature importance of ten best-selected attributes of Bangladeshi dataset.

Further, in Table 3.5, the second-highest accuracy is 93.0% when the number of features is 15. So for both the datasets, all the experimental results have been carried out for 15 number o features also for doing the comparison.

For UCI dataset the selected 15 features include: 'Age', 'Number of sexual partners', 'First sexual intercourse', 'Num of pregnancies', 'Smokes (years)', 'Smokes (packs/year)', 'Hormonal Contraceptives', 'Hormonal Contraceptives (years)', 'IUD (years)', 'IUD', 'STDs (number)' ,'Smokes', 'STDs','STDs:condylomatosis','STDs:vulvo-perineal condylomatosis'. For UCI dataset the fifteen best-selected features are shown in Figure 3.7.

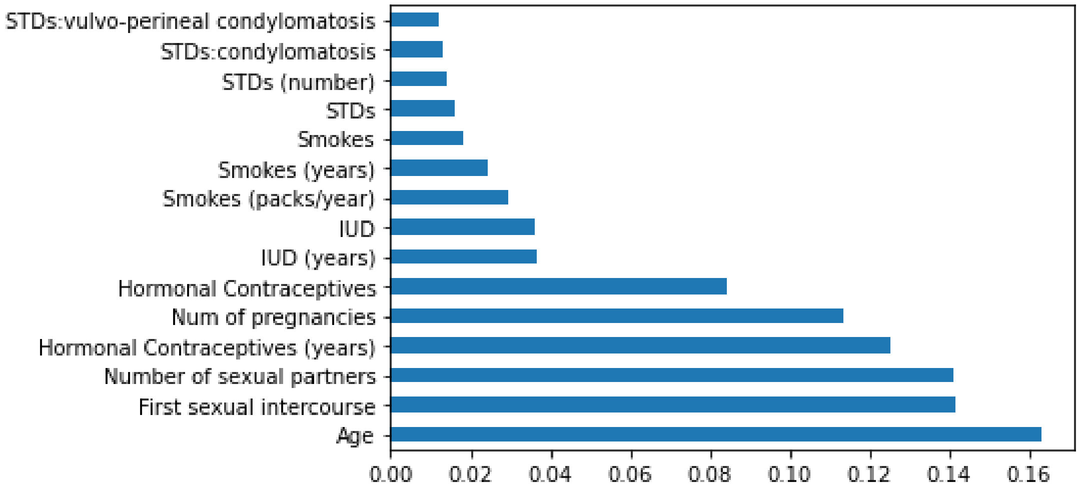


Figure 3.7: Feature importance of fifteen best-selected attributes of UCI dataset.

For the Bangladesi dataset, the selected 15 features include: 'Age', 'Children', 'Previous exposure to cancer', 'Disease', 'Menstrual period regular',' Menstrual period irregular', 'Whitish secretion', 'Bad odor', 'OCP', 'Use of Condom', 'IUCD', 'Marital status', 'Contraceptive injection', 'Family history of cancer', 'Postmenopausal'. For the Bangladeshi dataset, the fifteen best-selected features are shown in Figure 3.8.

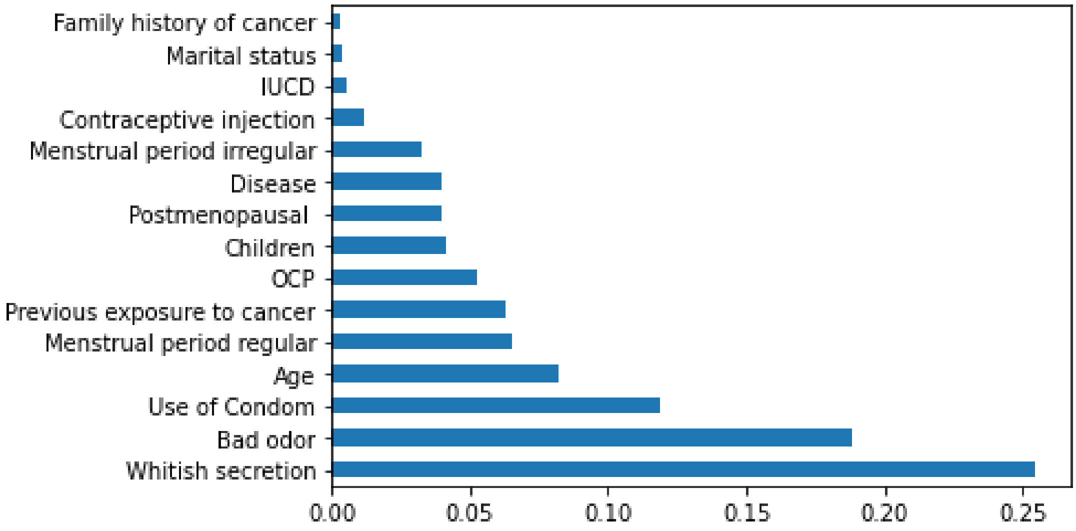


Figure 3.8: Feature importance of fifteen best-selected attributes of Bangladeshi dataset.

3.7 Model Evaluation

Total accuracy alone is not enough to evaluate a ML algorithm for the case of biomedical data including cervical cancer samples. It is more important to diagnose patients correctly. Furthermore, any incorrect prediction of a non-cancer patient as a cancer patient can be a serious issue. For the appropriate diagnosis of patients with cancer, this work considers several metrics. For the performance evaluation, true positive (TP) refers to correctly identified as cancer patients. The number of normal people who correctly get negative predictions that is they are classified as non-cancer patients is called true negative (TN). False-negative (FN) is the number of unrecognized patients who have cancer. False-positive (FP) refers to the number of samples without cancer but wrongly classified as cancer patients. With this consideration, true positive rate (TPR), true negative rate (TNR), false-positive rate (FPR), false-negative rate (FNR) are mathematically defined as follows.

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

$$\text{TNR}= \frac{TN}{TN+FP}$$

$$\text{FPR}= \frac{FP}{TN+FP}$$

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

For performance evaluation metrics such as training accuracy, testing accuracy, precision, recall, F1-measure are used.

3.7.1 Accuracy

Accuracy, ac, can be expressed as follows [75].

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

Training accuracy and testing accuracy are defined as the accuracy obtained for training and testing samples respectively

3.7.2 Precision

The number of predicted cancer patients among positive results is called precision, also known as positive predictive rate. Precision, pr, can be mathematically written as follows [75].

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

3.7.3 Recall

The term recall refers to the ratio of the number of correctly classified cancer patients to the total number of patients. Recall indicates the accuracy of a model in predicting the positive class for the case where the actual class is positive [75]. Recall is additionally referred to as sensitivity, TPR, and detection rate (DR). The term recall, can be given by

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

3.7.4 F1-Score

The F1-Measure is the weighted mean of the precision and recall [75] and represents the general performance given by

$$\text{f}_1= \frac{2\times pr\times re}{pr+re}$$

3.7.5 ROC-AUC

In addition to the above assessment parameters, the receiver operating characteristic (ROC) curve and the area under curve (AUC) can evaluate the advantages and disadvantages of the classier. The ROC curve shows the trade-off between the true positive rate (TPR) and the false positive rate (FPR) [75]. If the ROC curve is closer to the upper left corner of the graph, the model is better. The AUC is the area under the curve. When the area is closer to 1, the model is better.

3.8 Implementation Environment

Scikit Learn [76] was used for the implementation of the hybrid and other ML algorithms. To execute the python [76] file, the Jupiter notebook IDE was used.

3.9 Summary

In this chapter, the attributes and instances of the UCI and Bangladeshi datasets are discussed. Preprocessing of the dataset, handling the imbalanced dataset, the best number of features selection using the RFE method have been done. Then some evaluation matrices have been discussed.

.

CHAPTER 4
Prediction of Cervical Cancer with Proposed Hybrid Method

4.1 Introduction

This is the final and most important part of the thesis study, in which all of the data is analyzed. A two-staged hybrid method is proposed where in the first stage ExtraTreeClassifier and in the second stage Stacking ensemble method is used. Different types of classifier are applied on both the dataset and their performance are listed in Table 4.2 and 4.3. For analyzing the performance, ROC curves for both datasets have been shown.

4.2 Proposed Hybrid Method

In this section, a hybrid algorithm for predicting cervical cancer disease is introduced. This algorithm is adapted from Jogarah et al., who proposed a hybrid 2-staged ML method [77]. In our proposed hybrid method, a classification algorithm is applied in the first stage, and another classification algorithm is used in the second stage. The hybrid ML model is presented in Figure 4.1 The processed dataset (PD1) contains the ten best-selected features and the label column. The label column indicates cancerous(1) or noncancerous(0) state. PD1 is fed into the classification method used in stage 1. After the prediction of the classification algorithm using a cross-validation approach, PD1 is filtered to obtain cases where the predicted label and actual label are equal. The resultant filtered dataset defined as PD2 is then fed into the classification algorithm used in stage 2. Next, the accuracy of the classification algorithm in stage 2 is calculated by counting the accurately predicted values using a cross-validation approach. In both stages, 30 fold cross-validation approach has been considered. The pseudocode of the above technique can be written as:

Pseudocode 4.1: The Pseudocode for 2 staged hybrid algorithm

```
1.IMPORT all the library file
2.READ the dataset
3.DROP the columns which have huge number of missing values.
4. SET X to features column
   SET Y to target column
5.APPLY Borderline-SMOTE technique on whole dataset to handle the imbalance classification of the dataset.
6. RUN recursive feature elimination (RFE) method to select 10 best features.
7.  SET $X'$ to 10 best selected features from X
    SET $Y'$ to Y
8. Model-1: APPLY and FIT the ExtraTreeClassifier  on $X'$ and $Y'$.
9. EVALUATE  Model-1 using RepeatedStratifiedKFold cross-validation where divide the dataset( $X'$, $Y'$) randomly into k equal sized subsamples. For testing the model reserve a single subsample as the validation data of the k subsamples and use the remaining k-1 subsample as training data. Repeat the process k times, with each of the k subsamples used exactly once as the validation data. Average the k results to produce a single estimation to find the accuracy, precision, recall, and F1 score.
10. INITIALIZE Y_predict to Model-1 predicted values on $X'$
11. WHILE Y_predict  equal to $Y'$
        SET  $X''$ to $X'$
        SET   $Y''$ to $Y'$
    ENDWHILE
$X''$ and $Y''$ consists of the values which have been predicted correctly by the Model-1.
12. Model-2: APPLY and FIT the Stacking classifier which is a combination of five classifiers (Random forest, XGBoost, SVM, ExtraTreeClassifier, BaggingClassifier) on $X''$ and $Y''$
13. EVALUATE the Model-2 using RepeatedStratifiedKFold cross-validation where divide the dataset( $X''$ and $Y''$) randomly into k equal sized subsamples. For testing
```

the model reserve a single subsample as the validation data of the k subsamples and use the remaining k-1 subsample as training data. Repeat the process k times, with each of the k subsamples used exactly once as the validation data. Average the k results to produce a single estimation to find the accuracy, precision, recall, and F1 score.

14. END

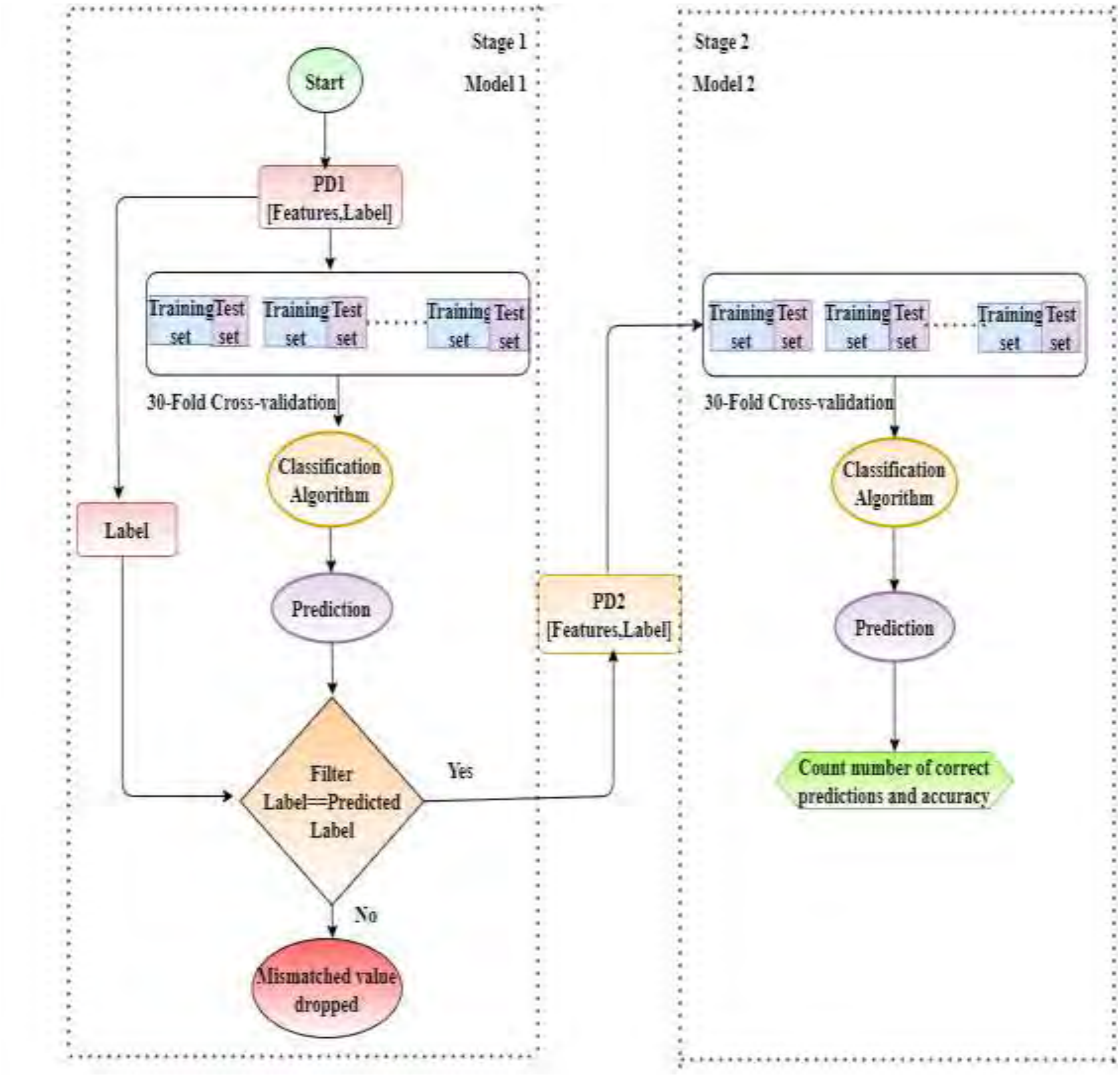


Figure 4.1: The training process of a hybrid model.

4.3 Result Analysis

This study used the scikit-learn library for programming in Python. Several evaluation metrics, including accuracy, precision, recall, and F1 score [78], the receiver operating characteristic (ROC) curve, and the area under the curve (AUC) [79-80] are used to evaluate the effectiveness of different classifiers. Different types of ML algorithms such as RF, SVM, KNN, DT, Extra Tree (ET), etc. are used for classification. Besides, four types of ensemble methods such as bagging, boosting, voting, and stacking are applied. Moreover, the three different types of boosting algorithms, AdaBoost, XGBoost, and Gradient Boosting, are applied. The hyperparameters of the different algorithms are shown in Table 4.1. In this work, the hyperparameters used for SVM are kernel=rbf, C= 2, and gamma=1 for the case of the UCI repository dataset. On the other hand, the hyperparameters for the Bangladeshi dataset are kernel=linear, C=default, and gamma=default. The hyperparameter number of neighbor (K) of KNN algorithm is selected 4 for UCI dataset. The hyperparameter max depth of DT is selected 20 for both datasets.

Table 4.1 Hyperparameter of different algorithms

Algorithm	**Hyper Parameter**
SVM	Kernel, C, and gamma
KNN	The number of neighbor,K
DT	Max depth
LR	No hyper parameter
NB	No hyperparameter

Tables 4.2, 4.3, 4.4 show the performances of different algorithms on the first dataset for the 10, 20, and 30 fold cross-validation approaches. Among these

different numbers of cross-validation approaches 30 fold cross-validation approach gives the best result. So, for 30 fold cross-validation in the case of the first dataset, RF [81-82] classifier provides 94.0% accuracy, which means the number of correctly predicted cancer patients out of all the samples is 94.0%. The confusion matrix using the RF classifier is shown in Figure 4.2. Among all these methods, stacking with a combination of RF, ExtraTreeClassifier, XGBoost, and Bagging classifier with base estimator ExtraTreeClassifier gives us the highest performance with 95.3% accuracy.

Table 4.2: Different algorithms performance on UCI Machine Learning Repository dataset when the value of k=10

Algorithm	**Accuracy**	**Precision**	**Recall**	**F1-score**
Random Forest	0.924	0.954	0.896	0.924
XGBoost	0.926	0.952	0.894	0.921
K Nearest Neighbors (k=4)	0.926	0.951	0.893	0.921
Support Vector Machine (kernel='rbf', C=2, gamma=1)	0.928	0.952	0.896	0.922
Decision Tree (max_ depth=20)	0.924	0.950	0.896	0.926
AdaBoost	0.926	0.950	0.899	0.922
Bagging	0.931	0.949	0.912	0.929
ExtraTreeClassifier	0.933	0.944	0.916	0.931
Gradient Boosting	0.914	0.962	0.862	0.908
Stacking	0.942	0.942	0.930	0.933
Hybrid Model	0.949	0.946	0.943	0.944

Table 4.3: Different algorithms performance on UCI Machine Learning Repository dataset when the value of k=20

Algorithm	Accuracy	Precision	Recall	F1-score
Random Forest	0.931	0.956	0.903	0.927
XGBoost	0.914	0.932	0.894	0.911
K Nearest Neighbors (k=4)	0.850	0.825	0.892	0.856
Support Vector Machine (kernel='rbf', C=2, gamma=1)	0.882	0.943	0.814	0.873
Decision Tree (max_ depth=20)	0.859	0.852	0.873	0.860
AdaBoost	0.848	0.871	0.821	0.843
Bagging	0.936	0.950	0.922	0.935
ExtraTreeClassifier	0.937	0.945	0.924	0.935
Gradient Boosting	0.897	0.949	0.841	0.890
Stacking	0.950	0.942	0.937	0.941
Hybrid Model	0.953	0.951	0.949	0.952

Table 4.4: Different algorithms performance on UCI Machine Learning Repository dataset when the value of k=30

Algorithm	Accuracy	Precision	Recall	F1-score
Random Forest	0.940	0.955	0.915	0.937
XGBoost	0.914	0.934	0.894	0.912
K Nearest Neighbors (k=4)	0.855	0.924	0.908	0.862
Support Vector Machine (kernel='rbf', C=2, gamma=1)	0.888	0.951	0.821	0.878
Decision Tree (max_ depth=20)	0.865	0.858	0.883	0.868
AdaBoost	0.848	0.867	0.828	0.844
Bagging	0.943	0.948	0.939	0.942

Algorithm	**Accuracy**	**Precision**	**Recall**	**F1-score**
ExtraTreeClassifier	0.941	0.944	0.931	0.936
Gradient Boosting	0.907	0.960	0.851	0.899
Voting Classifier (Bagging, ExtraTreeClassifier) Hard voting	0.943	0.962	0.927	0.941
Voting Classifier (Bagging, ExtraTreeClassifier) Soft voting	0.939	0.945	0.935	0.939
Stacking	0.953	0.949	0.940	0.943
Hybrid Model	0.959	0.956	0.953	0.954

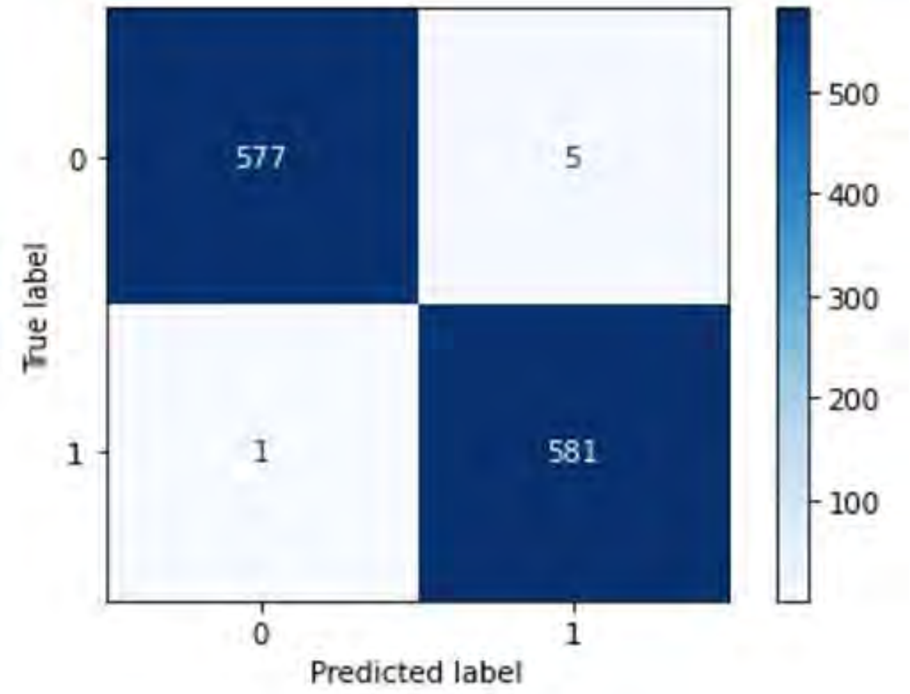


Figure 4.2: Confusion matrix for UCI dataset using RF classifier.

On the other hand, for the Bangladeshi dataset, all the base and ensemble classifiers have been applied and among them the AdaBoost algorithm performs the best by providing 95.6% accuracy. Figure 4.3 demonstrated a confusion matrix for the Bangladeshi dataset by using RF classifier.

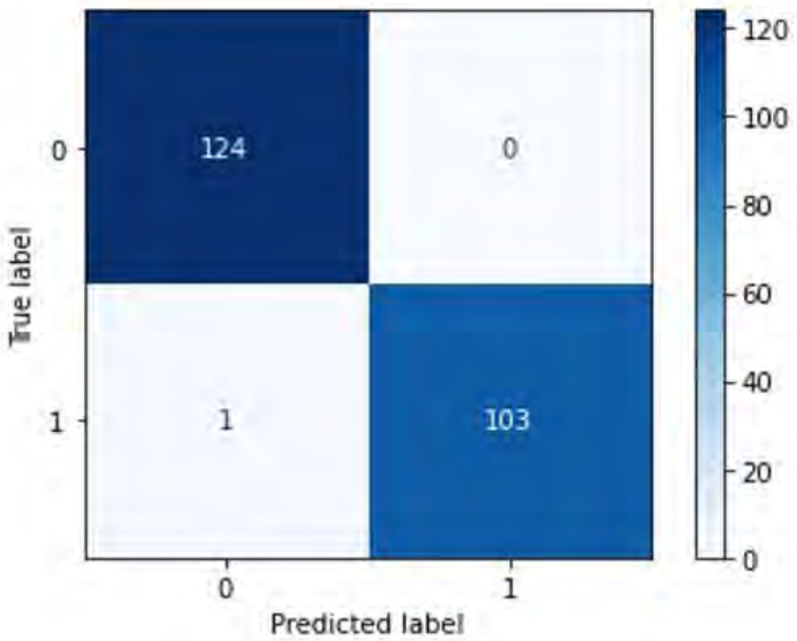


Figure 4.3: Confusion matrix for Bangladeshi dataset using RF classifier.

Tables 4.5, 4.6, 4.7 show the performances of different algorithms on the Bangladeshi dataset for the 10, 20, and 30 fold cross-validation approaches. Among these different numbers of cross-validation approaches 30 fold cross-validation approach gives the best result.

Table 4.5: Different algorithms performance on Bangladeshi dataset when the value of k=10

Algorithm	**Accuracy**	**Precision**	**Recall**	**F1-score**
Random Forest	0.940	0.958	0.904	0.934
Logistic Regression	0.938	0.973	0.889	0.919
XGBoost	0.930	0.965	0.882	0.918
Support Vector Machine (kernel='linear')	0.940	0.970	0.898	0.930
GaussianNB	0.886	0.770	0.770	0.855
Decision Tree (max_ depth=20)	0.927	0.929	0.925	0.920
AdaBoost	0.946	0.956	0.930	0.939
Bagging	0.944	0.970	0.908	0.935
ExtraTreeClassifier	0.927	0.941	0.903	0.918
Gradient Boosting	0.935	0.961	0.897	0.924
Stacking	0.940	0.945	0.928	0.938
Hybrid Model	0.950	0.973	0.937	0.942

Table 4.6: Different algorithms performance on Bangladeshi dataset when the value of k=20

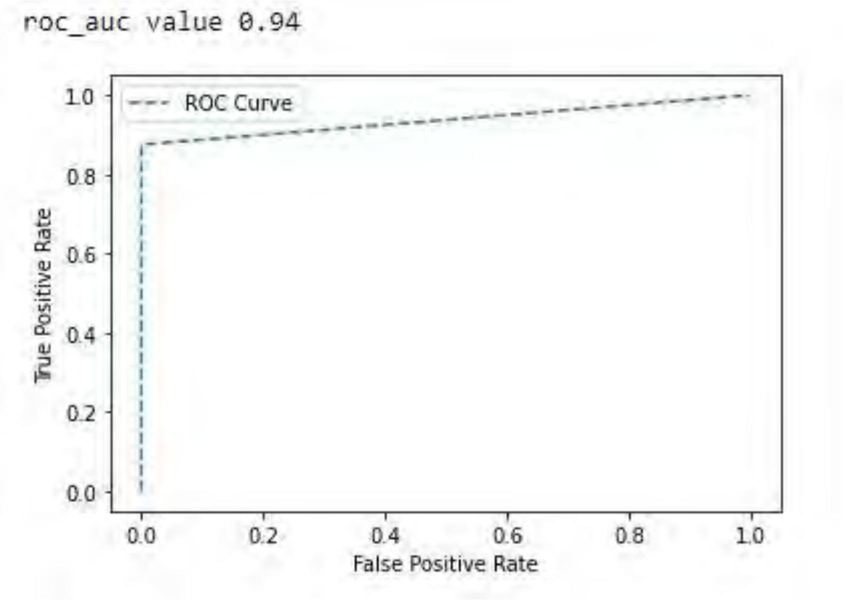
Algorithm	**Accuracy**	**Precision**	**Recall**	**F1-score**
Random Forest	0.939	0.969	0.920	0.938
Logistic Regression	0.939	0.975	0.896	0.928
XGBoost	0.934	0.971	0.887	0.921
Support Vector Machine (kernel='linear')	0.939	0.975	0.895	0.927
GaussianNB	0.888	0.984	0.770	0.853
Decision Tree (max_ depth=20)	0.925	0.937	0.916	0.919

Algorithm	Accuracy	Precision	Recall	F1-score
AdaBoost	0.955	0.977	0.929	0.947
Bagging	0.952	0.975	0.923	0.943
ExtraTreeClassifier	0.940	0.968	0.906	0.929
Gradient Boosting	0.942	0.973	0.904	0.931
Stacking	0.942	0.949	0.930	0.940
Hybrid Model	0.955	0.979	0.940	0.945

Table 4.7: Different algorithms performance on Bangladeshi dataset when the value of k=30

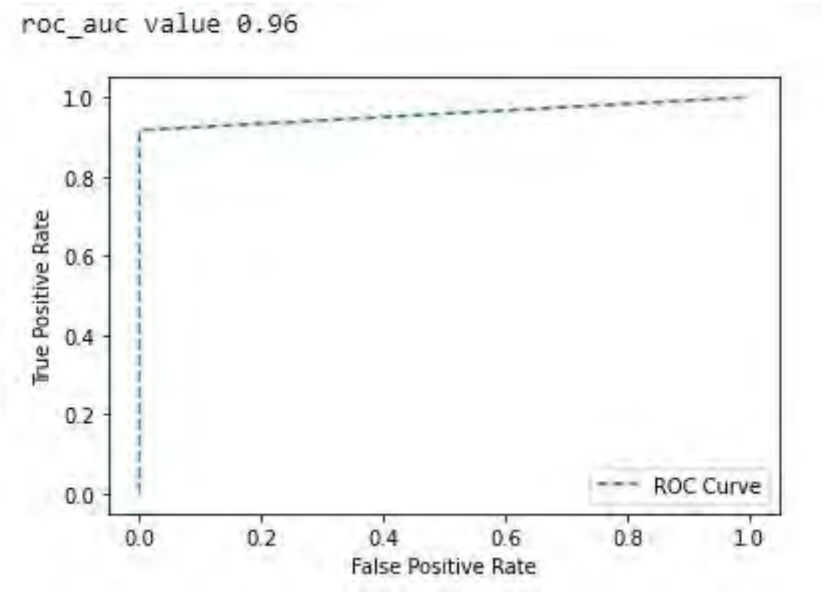
Algorithm	Accuracy	Precision	Recall	F1-score
Random Forest	0.940	0.968	0.916	0.925
Logistic Regression	0.914	0.964	0.853	0.889
XGBoost	0.943	0.981	0.898	0.926
Support Vector Machine (kernel='linear')	0.921	0.975	0.855	0.894
GaussianNB	0.907	0.983	0.813	0.872
Decision Tree (max_ depth=20)	0.942	0.956	0.925	0.931
AdaBoost	0.956	0.976	0.933	0.946
Bagging	0.943	0.975	0.903	0.926
ExtraTreeClassifier	0.926	0.961	0.885	0.907
Gradient Boosting	0.934	0.967	0.893	0.917
Voting Classifier (RF, Bagging, ExtraTreeClassifier) Hard voting	0.938	0.970	0.903	0.920
Voting Classifier (Bagging, Extra Tree Classifier) Soft voting	0.934	0.964	0.889	0.912
Stacking	0.944	0.958	0.932	0.943
Hybrid Model	0.962	0.980	0.941	0.948

Figure 4.4 and Figure 4.5 represent the ROC curve of UCI and the Bangladeshi datasets, respectively. For the UCI dataset and the Bangladeshi dataset, AUC values of 0.94 and 0.96 represent 94.0% and 96.0% chance that the physician reading the data will correctly distinguish between cancer and non-cancer patients, respectively.



```
roc_auc value 0.94
```

Figure 4.4: ROC curve for UCI dataset.



```
roc_auc value 0.96
```

Figure 4.5: ROC curve for the Bangladeshi dataset.

For the first dataset, ExtraTreeClassifier performed the best among all the single classifiers. So, in the proposed hybrid method ExtraTreeClassifier was used in the first stage, and stacking was applied in the second. Using ExtraTreeClassifier, 94.1% accuracy in the first stage is achieved. It was found that the ExtraTreeClassifier algorithm misclassified six samples. In the second stage, stacking ensemble classifier as a combination of RF, ExtraTreeClassifier, XGBoost, bagging classifier with base estimator ExtraTreeClassifier was applied and 95.9% accuracy was achieved. For the Bangladeshi dataset, 93.7% accuracy using ExtraTreeClassifier at the first stage was obtained. In the second stage, 96.2% accuracy was achieved using the stacking ensemble method which showed improvement compared to the first stage. Here, staking is a combination of basic DT, XGBoost, AdaBoost, RF algorithms. The analysis shows that the performance of a single classifier can be improved by using the hybrid method.

Further, the performance of all the algorithms is measured considering the fifteen best number of features. It is observed that in both cases the performance varies very slightly with the performance of algorithms when the number of features is chosen 10. Tables 4.8 and 4.9 demonstrate the performance of all algorithms for both datasets when the best number of features is fifteen.

Table 4.8: Different algorithms performance on UCI dataset when the best number of features is 15

Algorithm	**Accuracy**	**Precision**	**Recall**	**F1-score**
Random Forest	0.929	0.959	0.897	0.927
XGBoost	0.912	0.938	0.885	0.909
K Nearest Neighbors (k=4)	0.848	0.825	0.890	0.854
Support Vector Machine (kernel=’rbf’, C=2, gamma=1)	0.883	0.951	0.811	0.872
Decision Tree (max_ depth=20)	0.873	0.867	0.884	0.877

Algorithm	Accuracy	Precision	Recall	F1-score
AdaBoost	0.855	0.876	0.835	0.852
Bagging	0.930	0.945	0.913	0.928
ExtraTreeClassifier	0.932	0.948	0.913	0.928
Gradient Boosting	0.907	0.956	0.856	0.900
Voting Classifier (Bagging, ExtraTreeClassifier) Hard voting	0.931	0.952	0.914	0.929
Voting Classifier (Bagging, ExtraTreeClassifier) Soft voting	0.931	0.947	0.916	0.930
Stacking	0.952	0.945	0.938	0.940
Hybrid Model	0.957	0.954	0.950	0.951

Table 4.9: Different algorithms performance on Bangladeshi dataset when the best number of features is 15

Algorithm	Accuracy	Precision	Recall	F1-score
Random Forest	0.942	0.968	0.925	0.932
XGBoost	0.938	0.979	0.894	0.923
K Nearest Neighbors (k=4)	0.801	0.924	0.620	0.715
Support Vector Machine (kernel=’rbf’, C=2, gamma=1)	0.772	0.909	0.576	0.576
Decision Tree (max_ depth=20)	0.931	0.943	0.918	0.921
AdaBoost	0.961	0.980	0.943	0.955
Bagging	0.952	0.977	0.927	0.943
ExtraTreeClassifier	0.938	0.969	0.905	0.925
Gradient Boosting	0.944	0.980	0.905	0.932
Voting Classifier (Bagging, ExtraTreeClassifier) Hard voting	0.947	0.974	0.924	0.939

Algorithm	Accuracy	Precision	Recall	F1-score
Voting Classifier (Bagging, ExtraTreeClassifier) Soft voting	0.941	0.964	0.918	0.931
Stacking	0.952	0.956	0.930	0.942
Hybrid Model	0.961	0.981	0.941	0.947

Next, the cross-dataset case was considered. The cross-dataset concept was developed using the UCI ML repository and the Bangladeshi dataset. It was found that among the two datasets, four features were in common. Those are 'Age,' 'the number of pregnancies,' 'Hormonal Contraceptive,' and 'IUD'. The model was trained by using the UCI ML repository dataset and tested by using the Bangladeshi dataset. For splitting the training and testing portion of the datasets, the cross-validation method was considered. Using the AdaBoostClassifier, a testing accuracy of 67.8% and a training accuracy of 56.57% are achieved. This indicated that for the case of cross-dataset, the proposed algorithms did not perform well.

4.4 Discussion

In the following, the results of this work are compared with previous studies. It can be noted that the performance of different classifiers varies when different datasets are taken into consideration. Previous studies applied different algorithms to different datasets to predict cervical cancer [41, 49, 51, 83, 84, 85, 86]. Z. Rustam et al. applied the GNRBA approach on the UCI ML repository resulting in 93% accuracy with all number of features [41]. Razali, N et al. [49] reported that SVM-PCA provided an accuracy of 94.03% when UCI dataset is considered and the target variable is Biopsy. An enhanced ensemble method by Ilyas and Ahmad obtained 94% accuracy in predicting cervical cancer for UCI dataset [51]. F.Asadi [83] applied SVM, QUEST, C&R tree, multi-layer perceptron artificial neural network (MLP-ANN) and radial basis function (RBF-ANNs) to analyze a dataset consisting of 23 attributes of 145 patients from Shohada Hospital in Tehran, Iran, and acquired 95.55% accuracy. Kurniawati et al. [84] used SVM method obtaining 79% accuracy and 85% AUC value for UCI dataset. By analyzing hierarchical representations from Swedish electronic health data, Weegar, R et al. [85] achieved the best result with an AUC of 70% for RF algorithm. Kusy et al. [86] reported that the RBF neural network algorithm obtained 67 percent specificity in the case of UCI dataset. Compared to the previous studies, our work obtains acceptable performance results. The proposed hybrid method offers classification accuracy of 95.9% for UCI dataset and 96.2% for the Bangladeshi dataset.

Though the UCI repository dataset is available online, the Bangladeshi dataset will turn over a new leaf in the field of oncology. In the precancerous stage, most women do not have any symptoms. With time, symptoms start to appear in the early cancerous stage. Symptoms may be more severe depending on the stage and grade of cervical cancer. The results obtained from one dataset may not be fully applicable to different scenarios. The factors contributing to cancer may vary for patients of different geographical regions, cultural backgrounds, ethnicity, etc. So, compared to the UCI dataset, a Bangladeshi dataset may be more useful in predicting cervical cancer in the context of Bangladesh. The new Bangladeshi dataset considers important attributes related to the menstrual period missed in the existing UCI dataset. When ML is applied to the Bangladeshi dataset, it was found that some of the important features in predicting cervical cancer are blood spots, menstrual bleeding longer and heavier than usual, bleeding after menopause [87], abnormal vaginal discharge with bad odor, problems in the menstrual cycle [88]. The UCI repository dataset does not contain the attributes mentioned above. These new features of the Bangladeshi dataset will be helpful in the prediction of cervical cancer.

4.5 Summary

In this chapter, a new hybrid method is proposed for both dataset and it shows that the hybrid method works better than the previous algorithm. Chapter 4 also demonstrate the results of different basic and ensemble algorithm on both datasets and finds out which algorithm performs best on both dataset. Lastly a discussion session has been added where the importance of the thesis work and comparioson between previous work has been discussed.

.

CHAPTER 5
Conclusion and Future Works

5.1 Summary of Findings

The thesis work has been done on cervical cancer based on the existing UCI dataset and the Bangladeshi dataset. The preprocessing of both the dataset have been done as there are some missing values in the datasets. The imbalanced UCI dataset has been balanced by using the Borderline-SMOTE technique. As the Bangladeshi dataset is balanced, the Borderline-SMOTE algorithm is not applied to this dataset. To improve the performance of the model and reduce the computational cost of modeling, it is desirable to minimize the number of input variables. Feature selection is the procedure of minimizing the number of input variables when developing a predictive model. From the filtering-based method information gain and correlation matrix have been used. The correlation matrix illustrates the relationships between all the features of the Bangladeshi and UCI dataset whereas information gain shows the importance of all the features serially. From the wrapper method, RFE has been used. By using a group of several features accuracy has been calculated and it is found that when the number of features is 10 the highest accuracy can be obtained by using the RFE method. So the number of features is selected 10 for both of the datasets. The best ten selected features for UCI dataset include: 'Age', 'Number of sexual partners, 'First sexual intercourse, 'Num of pregnancies', 'Smokes (years)', 'Smokes (packs/year)', 'Hormonal Contraceptives', 'Hormonal Contraceptives (years)', 'IUD (years)', 'STDs (number)' . The best ten selected features for the Bangladeshi dataset include: ‘Age’, ‘Children’, ‘Previous exposure to cancer’, ‘Disease’, ‘Menstrual period regular’,’ Menstrual period irregular’, ‘Whitish secretion’, ‘Bad odor’, ‘OCP’, ‘Use of Condom’. With these ten best-selected features, different types of classification algorithms such as RF, KNN, SVM, DT, XGBoost, AdaBoost, Gradient Boosting, Bagging, Voting, Stacking are run. From all of these classifiers, the ensemble classifier stacking gives the best performance for both the dataset. Here stacking consists of base classifiers that perform well individually. It implies that By using ensemble classifier stacking the performance of the base classifier can be improved. The main contribution of the research work is the proposed hybrid method. In the proposed hybrid method, there are two stages. In the first stage, a classification algorithm is used and the output of the first stage is fed into the second stage as input where a different classification algorithm is used. By using the hybrid method, it is shown that the hybrid method outperformed the stacking classifier which is giving the best performance previously for both datasets. For the UCI dataset, 95.9% accuracy and Bangladeshi dataset 96.2% accuracy is achieved. A block diagram and pseudocode are also introduced for better visualization and analysis. Next, a cross-dataset concept is introduced using the UCI and Bangladeshi datasets. Finally, there is a discussion section where the performance of different papers is compared to this research work.

5.2 Conclusion

This thesis work applies ML algorithms to classify cervical cancer patients in the case of an existing and a novel dataset. The current dataset has 36 attributes and is freely available in the UCI ML repository. The new dataset has 21 attributes. For the existing dataset, the target attribute is formed by the OR operation of four attributes, namely Hinselmann, Biopsy, Schiller, and Cytology. For the case of the new dataset, the VIA test is the target attribute. The new dataset considers important attributes related to the menstrual period that is missed in the existing dataset. The RFE algorithm is applied to find the best important attributes of cervical cancer. Next, a two-stage hybrid classifier is proposed where the first stage consists of ExtraTreeClassifier and the second stage has stacking. This new hybrid classifier obtains accuracy values of 95.9% and 96.2% when applied to the existing and the new datasets, respectively. Finally, classification accuracy is determined for the case of cross-dataset with the four common attributes. When the existing dataset is used for training, and the new dataset is used for testing, the accuracy value drops to 67.0%.

5.3 Future Work

Despite outperforming the base algorithm, the suggested hybrid method has a few limitations that were discovered during the research. They will assist in determining the algorithm's potential for further development. The following is a list of these issues:

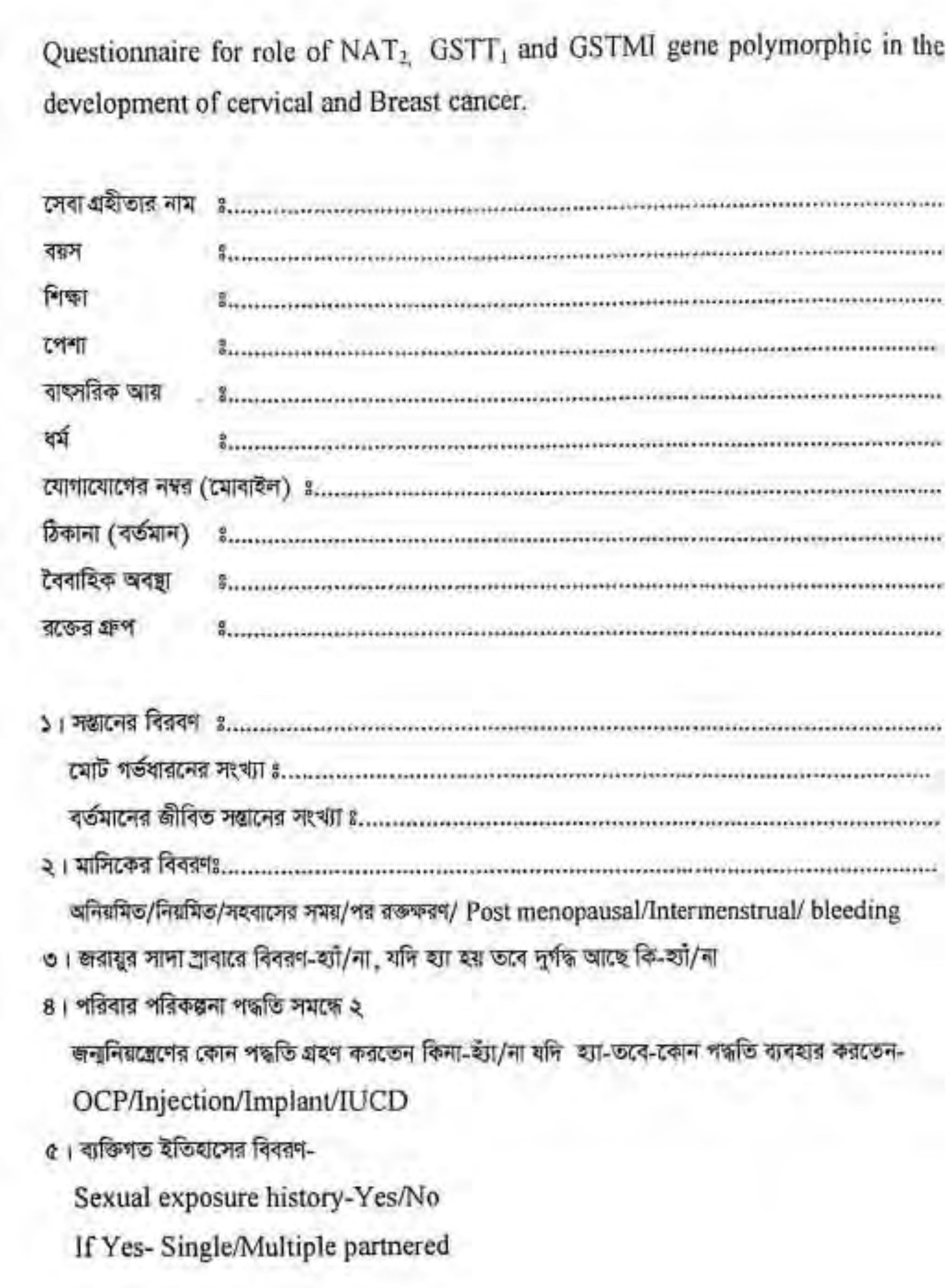
- Only the base and ensemble classifier algorithms of ML have been applied. No deep learning method has been used.
- The results obtained from one dataset may not be fully applicable to different scenarios.
- There are many government and non-government hospitals for cancer patients in Bangladesh, such as the National Institute of Cancer Research and Hospital, Ahsania Mission Cancer Hospital, etc. Data collection can be done from these hospitals in the future.
- As many hospitals have policies for the assurance of privacy concerns, it is difficult to collect a large number of data samples of cervical cancer.

There are a number of areas where future work might be done to improve the algorithm's functionality and give extra performance evaluation. They are as follows:

.

- Other prominent classifiers, such as deep learning with artificial neural networks and base level methods, may be taken into consideration
- This study provides results on UCI dataset and the Bangladeshi dataset. Since these two datasets have only a few columns in common, the cross dataset results only considered 4 attributes. More research should be done on cross datasets where two different datasets have many attributes in common
- The results of the proposed hybrid classifier must be validated for larger datasets so that efficiency can be increased in the prediction of cancerous and non-cancerous patients
- New hybrid algorithms can be developed from multiple base classifiers to make the prediction more reliable.

Appendix



Questionnaire for role of NAT$_2$, GSTT$_1$ and GSTMI gene polymorphic in the development of cervical and Breast cancer.

সেবা গ্রহীতার নাম ঃ..

বয়স ঃ..

শিক্ষা ঃ..

পেশা ঃ..

বাৎসরিক আয় ঃ..

ধর্ম ঃ..

যোগাযোগের নম্বর (মোবাইল) ঃ..

ঠিকানা (বর্তমান) ঃ..

বৈবাহিক অবস্থা ঃ..

রক্তের গ্রুপ ঃ..

১। সন্তানের বিরবণ ঃ..

মোট গর্ভধারনের সংখ্যা ঃ..

বর্তমানের জীবিত সন্তানের সংখ্যা ঃ..

২। মাসিকের বিবরণঃ..

অনিয়মিত/নিয়মিত/সহবাসের সময়/পর রক্তক্ষরণ/ Post menopausal/Intermenstrual/ bleeding

৩। জরায়ুর সাদা স্রাবারে বিবরণ-হ্যাঁ/না, যদি হ্যা হয় তবে দুর্গন্ধ আছে কি-হ্যাঁ/না

৪। পরিবার পরিকল্পনা পদ্ধতি সমন্ধে ২

জন্মনিয়ন্ত্রেণের কোন পদ্ধতি গ্রহণ করতেন কিনা-হ্যাঁ/না যদি হ্যা-তবে-কোন পদ্ধতি ব্যবহার করতেন-

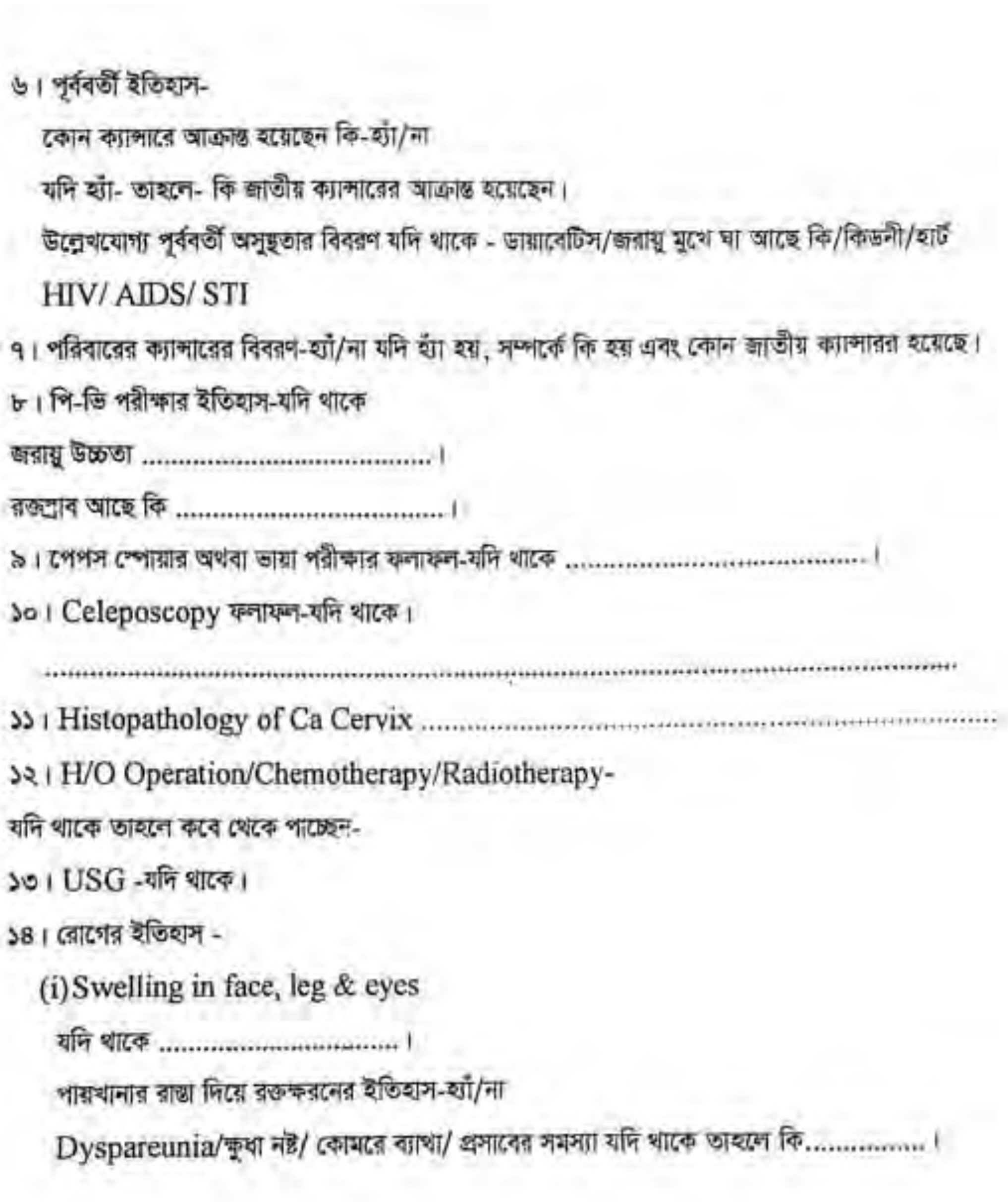
OCP/Injection/Implant/IUCD

৫। ব্যক্তিগত ইতিহাসের বিবরণ-

Sexual exposure history-Yes/No

If Yes- Single/Multiple partnered

Figure A.1: First page of the sample questionnaire



৬। পূর্ববর্তী ইতিহাস-

কোন ক্যান্সারে আক্রান্ত হয়েছেন কি-হ্যাঁ/না

যদি হ্যাঁ- তাহলে- কি জাতীয় ক্যান্সারের আক্রান্ত হয়েছেন।

উল্লেখযোগ্য পূর্ববর্তী অসুস্থতার বিবরণ যদি থাকে - ডায়াবেটিস/জরায়ু মুখে ঘা আছে কি/কিডনী/হার্ট

HIV/ AIDS/ STI

৭। পরিবারের ক্যান্সারের বিবরণ-হ্যাঁ/না যদি হ্যাঁ হয়, সম্পর্কে কি হয় এবং কোন জাতীয় ক্যান্সারর হয়েছে।

৮। পি-ভি পরীক্ষার ইতিহাস-যদি থাকে

জরায়ু উচ্চতা ....।

রক্তস্রাব আছে কি ....।

৯। পেপস স্পোয়ার অথবা ভায়া পরীক্ষার ফলাফল-যদি থাকে ....।

১০। Celeposcopy ফলাফল-যদি থাকে।

....

১১। Histopathology of Ca Cervix ....

১২। H/O Operation/Chemotherapy/Radiotherapy-

যদি থাকে তাহলে কবে থেকে পাচ্ছেন-

১৩। USG -যদি থাকে।

১৪। রোগের ইতিহাস -

(i) Swelling in face, leg & eyes

যদি থাকে ....।

পায়খানার রাস্তা দিয়ে রক্তক্ষরনের ইতিহাস-হ্যাঁ/না

Dyspareunia/ক্ষুধা নষ্ট/ কোমরে ব্যাথা/ প্রসাবের সমস্যা যদি থাকে তাহলে কি...............।

Figure A.2: Second page of the sample questionnaire

References

[1] MHFW, 2017. National Strategy for Cervical Cancer Prevention and Control Bangladesh (2017-2022). Directorate General of Health Services (DGHS) Ministry of Health and Family Welfare Government of the people's Republic of Bangladesh, 2017.

[2] Chandran, V., Sumithra, M. G., Karthick, A., George, T., Deivakani, M., Elakkiya, B., & Manoharan, S. (2021). Diagnosis of Cervical Cancer based on Ensemble Deep Learning Network using Colposcopy Images. BioMed Research International, 2021.

[3] Worldwide cancer statistics. Retrieved from http://www.cancerresearchuk.org/cancerinfo/cancerstats/world/ Retrieved on: 15 July 2021.

[4] Dong, N., et al., Inception v3 based cervical cell classification combined with artificially extracted features. Applied Soft Computing, 2020. 93: p. 106311.

[5] Yusufaly, T.I., et al., A knowledge-based organ dose prediction tool for brachytherapy treatment planning of patients with cervical cancer. Brachytherapy, 2020. 19(5): p. 624-634.

[6] Sompawong, N., et al., Automated pap smear cervical cancer screening using deep learning in 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). 2019. IEEE. p. 7044-7048.

[7] Shalev-Shwartz, S. and S. Ben-David, Understanding machine learning: From theory to algorithms. 2014: Cambridge university press.

[8] Hastie, T., R. Tibshirani, and R.J. Tibshirani, Extended comparisons of best subset selection, forward stepwise selection, and the lasso. arXiv preprint arXiv:1707.08692, 2017.

[9] Marsland, S. and M. Learning, An Algorithmic Perspective. 2020: Boca Raton, FL, USA: CRC Press.

[10] Hearst, M.A., et al., Support vector machines. IEEE Intelligent Systems and their applications, 1998. 13(4): p. 18-28.

[11] Wang, G. A survey on training algorithms for support vector machine classifiers. in 2008 Fourth International Conference on Networked Computing and Advanced Information Management. 2008. IEEE.

[12] Laaksonen, J. and E. Oja. Classification with learning k-nearest neighbors. in Proceedings of International Conference on Neural Networks (ICNN'96). 1996. IEEE.

[13] Win, S.L., et al., Cancer recurrence prediction using machine learning. International Journal of Computational Science and Information Technology (IJCSIT), 2014. 6(1).

[14] Tseng, C.-J., et al., Application of machine learning to predict the recurrence-proneness for cervical cancer. Neural Computing and Applications, 2014. 24(6): p. 1311-1316.

[15] William, W., et al., A review of image analysis and machine learning techniques for automated cervical cancer screening from pap-smear images. Computer methods and programs in biomedicine, 2018. 164: p. 15-22.

[16] Liu, Z.-C., et al., Multiple sexual partners as a potential independent risk factor for cervical cancer: a meta-analysis of epidemiological studies. Asian Pacific Journal of Cancer Prevention, 2015. 16(9): p. 3893-3900.

[17] Gast, K. and T. Snyder, Combination oral contraceptives and cancer risk. Kansas medicine: the journal of the Kansas Medical Society, 1990. 91(7): p. 201-208.

[18] Slattery, M.L., et al., Cigarette smoking and exposure to passive smoke are risk factors for cervical cancer. Jama, 1989. 261(11): p. 1593-1598.

[19] National Cancer Institute (2015) What Is Cancer?, National cancer institute. Available at: https://www.cancer.gov/about-cancer/understanding/what-is-cancer (Accessed: 3 February,2021)

[20] Wild, C.P., B.W. Stewart, and C. Wild, World cancer report 2014. 2014: World Health Organization Geneva, Switzerland.

[21] Burd, E.M., Human papillomavirus and cervical cancer. Clinical microbiology reviews, 2003. 16(1): p. 1-17.

[22] Parkin, D.M. and F. Bray, The burden of HPV-related cancers. Vaccine, 2006. 24: p. S11-S25.

[23] Schiffman, M., et al., Human papillomavirus and cervical cancer. The Lancet, 2007. 370(9590): p. 890-907.

[24] Zur Hausen, H., Human papillomaviruses and their possible role in squamous cell carcinomas. Current topics in microbiology and immunology, 1977: p. 1-30.

[25] Cuzick, J., et al., Type-specific human papillomavirus DNA in abnormal smears as a predictor of high-grade cervical intraepithelial neoplasia. British Journal of Cancer, 1994. 69(1): p. 167-171

[26] Cox, J.T., et al., An evaluation of human papillomavirus testing as part of referral to colposcopy clinics. Obstetrics and Gynecology, 1992. 80(3 Pt 1): p. 389-395.

[27] Koutsky, L.A., et al., A cohort study of the risk of cervical intraepithelial neoplasia grade 2 or 3 in relation to papillomavirus infection. New England journal of medicine, 1992. 327(18): p. 1272-1278.

[28] Lorincz, A.T., et al., Human papillomavirus infection of the cervix: relative risk associations of 15 common anogenital types. Obstetrics and gynecology, 1992. 79(3): p. 328-337.

[29] Zur Hausen, H., Viruses in human cancers. Science, 1991. 254(5035): p. 1167-1173.

[30] Reeves, W.C., et al., Human papillomavirus infection and cervical cancer in Latin America. New England Journal of Medicine, 1989. 320(22): p. 1437-1441d.

[31] Gillison, M.L., A.K. Chaturvedi, and D.R. Lowy, HPV prophylactic vaccines and the potential prevention of noncervical cancers in both men and women. Cancer, 2008. 113(S10): p. 3036-3046.

[32] O. Ayinde, Human Papillomavirus (HPV) : The unseen leading cause of cervical cancer in women, MicroBiotics, October 13,2013. Accessed on: 20 April 2021.[Online]. Available: https://microbiotics.com.ng/human-papillomavirus-hpv-the-unseen-leading-cause-of-cervical-cancer-in-women.

[33] Winer, R.L., et al., Condom use and the risk of genital human papillomavirus infection in young women. New England Journal of Medicine, 2006. 354(25): p. 2645-2654.

[34] Sharma, M., et al., Classification of clinical dataset of cervical cancer using KNN. Indian Journal of Science and Technology, 2016. 9(28): p. 1-5.

[35] Nithya, B. and V. Ilango, Evaluation of machine learning based optimized feature selection approaches and classification methods for cervical cancer prediction. SN Applied Sciences, 2019. 1(6): p. 1-16.

[36] Lu, J., et al., Machine learning for assisting cervical cancer diagnosis: An ensemble approach. Future Generation Computer Systems, 2020. 106: p. 199-205.

[37] Ghoneim, A., G. Muhammad, and M.S. Hossain, Cervical cancer classification using convolutional neural networks and extreme learning machines. Future Generation Computer Systems, 2020. 102: p. 643-649.

[38] Fatlawi, H.K., Enhanced classification model for cervical cancer dataset based on cost sensitive classifier. International Journal of Computer Techniques, 2017. 4(4): p. 115-20.

[39] Deng, X., Y. Luo, and C. Wang. Analysis of risk factors for cervical cancer based on machine learning methods. in 2018 5th IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS). 2018. IEEE.

[40] Al-Wesabi, Y., A. Choudhury, and D. Won. Classification of cervical cancer dataset. in Avishek Choudhury, Wesabi, Classification of Cervical Cancer Dataset, Proceedings of the 2018 IISE Annual Conference, Orlando. 2018.

[41] Rustam, Z., V. Hapsari, and M. Solihin. Optimal cervical cancer classification using Gauss-Newton representation based algorithm. in AIP Conference Proceedings. 2019. AIP Publishing LLC.

[42] Gharekhan, A.H., et al. PCA based polarized fluorescence study for detecting human cervical dysplasia. in Dynamics and Fluctuations in Biomedical Photonics X. 2013. International Society for Optics and Photonics.

[43] Devi, S., P.K. Panigrahi, and A. Pradhan, Detecting cervical cancer progression through extracted intrinsic fluorescence and principal component analysis. Journal of biomedical optics, 2014. 19(12): p. 127003.

[44] Mukhopadhyay, S., et al., Tissue multifractality and hidden Markov model based integrated framework for optimum precancer detection. Journal of Biomedical Optics, 2017. 22(10): p. 105005.

[45] Luo, W., Predicting Cervical Cancer Outcomes: Statistics, Images, and Machine Learning. Frontiers in Artificial Intelligence, 2021. 4: p. 83.

[46] Obrzut, B., et al., Prediction of 10-year overall survival in patients with operable cervical cancer using a probabilistic neural network. Journal of Cancer, 2019. 10(18): p. 4189.

[47] Rahaman, M.M., et al., A survey for cervical cytopathology image analysis using deep learning. IEEE Access, 2020. 8: p. 61687-61710.

[48] Li, Y., et al., Computer-aided cervical cancer diagnosis using time-lapsed colposcopic images. IEEE transactions on medical imaging, 2020. 39(11): p. 3403-3415.

[49] Razali, N., et al. Risk factors of cervical cancer using classification in data mining. in Journal of Physics: Conference Series. 2020. IOP Publishing.

[50] Wu, W. and H. Zhou, Data-driven diagnosis of cervical cancer with support vector machine-based approaches. IEEE Access, 2017. 5: p. 25189-25195.

[51] Ilyas, Q.M. and M. Ahmad, An Enhanced Ensemble Diagnosis of Cervical Cancer: A Pursuit of Machine Intelligence Towards Sustainable Health. IEEE Access, 2021. 9: p. 12374-12388.

[52] Abdoh, S.F., M.A. Rizka, and F.A. Maghraby, Cervical cancer diagnosis using random forest classifier with SMOTE and feature reduction techniques. IEEE Access, 2018. 6: p. 59475-59485.

[53] Geetha, R., et al., Cervical cancer identification with synthetic minority oversampling technique and PCA analysis using random forest classifier. Journal of medical systems, 2019. 43(9): p. 1-19.

[54] Hill, T. and P. Lewicki, "Support Vector Machines," in Electronic Statistics Textbook, StatSoft Inc. 1995.

[55] Bishop, C., Bishop-Pattern Recognition and Machine Learning-Springer 2006. Antimicrob. Agents Chemother, 2014: p. 03728-14.

[56] Yekkehkhany, B., et al., A comparison study of different kernel functions for SVM-based classification of multi-temporal polarimetry SAR data. The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences, 2014. 40(2): p. 281.

[57] Liu, Y. and K.K. Parhi. Computing RBF kernel for SVM classification using stochastic logic. in 2016 IEEE International Workshop on Signal Processing Systems (SiPS). 2016. IEEE. P. 327-332.

[58] Ring, M. and B.M. Eskofier, An approximation of the Gaussian RBF kernel for efficient classification with SVMs. Pattern Recognition Letters, 2016. 84: p. 107-113.

[59] Jin, R., F. Yan, and J. Zhu, Application of logistic regression model in an epidemiological study. Science Journal of Applied Mathematics and Statistics, 2015. 3(5): p. 225-229.

[60] Bronshtein, A., A quick introduction to K-Nearest Neighbors Algorithm. Noteworthy-The Journal Blog, 2017.

[61] Losing, V., B. Hammer, and H. Wersing. KNN classifier with self adjusting memory for heterogeneous concept drift. in 2016 IEEE 16th international conference on data mining (ICDM). 2016. IEEE.

[62] Lin, W.-Z., et al., iDNA-Prot: identification of DNA binding proteins using random forest with grey model. PloS one, 2011. 6(9): p. e24756.

[63] Khalilia, M., S. Chakraborty, and M. Popescu, Predicting disease risks from highly imbalanced data using random forest. BMC medical informatics and decision making, 2011. 11(1): p. 1-13.

[64] Mohan, A., et al., Automatic classification of protein structures using physicochemical parameters. Interdisciplinary Sciences: Computational Life Sciences, 2014. 6(3): p. 176-186.

[65] Seera, M. and C.P. Lim, A hybrid intelligent system for medical data classification. Expert systems with applications, 2014. 41(5): p. 2239-2249.

[66] Breiman, L., Random Forest, vol. 45. Mach Learn, 2001. 1.p. 5–32.

[67] Breiman, L., et al., Classification and regression trees. Belmont, CA: Wadsworth. International Group, 1984. 432: p. 151-166.

[68] Chen, T. and C. Guestrin. Xgboost: A scalable tree boosting system. in Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016.

[69] Mariajesusbigml , Introduction to Boosted Trees, BigML, March 14,2017. Accessed on: 25 May 2021. [Online]. Available: https://blog.bigml.com/2017/03/14/introduction-to-boosted-trees/ .

[70] Kelwin Fernandes, Jaime S. Cardoso, and Jessica Fernandes. 'Transfer Learning with Partial Observability Applied to Cervical Cancer Screening.' Iberian Conference on Pattern Recognition and Image Analysis. Springer International Publishing, 2017. Available: https://archive.ics.uci.edu/ml/datasets/Cervical+cancer+(Risk+Factors).

[71] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res. 2002;16:321-357

[72] Han, H., W.-Y. Wang, and B.-H. Mao. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. in International conference on intelligent computing. 2005. Springer.

[73] Zhang, C., et al. Feature selection of power system transient stability assessment based on random forest and recursive feature elimination. in 2016 IEEE PES Asia-Pacific Power and Energy Engineering Conference (APPEEC). 2016. IEEE.

[74] Guyon, I., et al., Gene selection for cancer classification using support vector machines. Machine learning, 2002. 46(1): p. 389-422..

[75] Sokolova, M., N. Japkowicz, and S. Szpakowicz. Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation. in Australasian joint conference on artificial intelligence. 2006. Springer.

[76] “scikit-learn: machine learning in python — scikit-learn 0.20.0 documentation.” Available: http://scikit-learn.org/stable/. [Accessed 15 january 2021]

[77] Jogarah, K.K., et al., Hybrid machine learning algorithms for fault detection in android smartphones. Transactions on Emerging Telecommunications Technologies, 2018. 29(2): p. e3272.

[78] Rijsbergen CV. Information Retrieval. Butterworth-Heinemann, Newton MA, USA, 1979

[79] Ferraris, V.A., Commentary: Should we rely on receiver operating characteristic curves? From submarines to medical tests, the answer is a definite maybe! The Journal of thoracic and cardiovascular surgery, 2019. 157(6): p. 2354-2355..

[80] Mandrekar, J.N., Receiver operating characteristic curve in diagnostic test assessment. Journal of Thoracic Oncology, 2010. 5(9): p. 1315-1316.

[81] Chen, Q., et al. Prediction of DNA-binding protein using random forest and elastic net. in 2017 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD). 2017. IEEE.

[82] Biau, G., Analysis of a random forests model. The Journal of Machine Learning Research, 2012. 13: p. 1063-1095.

[83] Asadi, F., C. Salehnasab, and L. Ajori, Supervised Algorithms of Machine Learning for the Prediction of Cervical Cancer. Journal of Biomedical Physics & Engineering, 2020. 10(4): p. 513.

[84] Kurniawati, Y.E., A.E. Permanasari, and S. Fauziati. Comparative study on data mining classification methods for cervical cancer prediction using pap smear results. in 2016 1st International Conference on Biomedical Engineering (IBIOMED). 2016. IEEE.

[85] Weegar, R. and K. Sundström, Using machine learning for predicting cervical cancer from Swedish electronic health records by mining hierarchical representations. Plos one, 2020. 15(8): p. e0237911.

[86] Kusy, M., B. Obrzut, and J. Kluska, Application of gene expression programming and neural networks to predict adverse events of radical hysterectomy in cervical cancer patients. Medical & biological engineering & computing, 2013. 51(12): p. 1357-1365.

[87] Harry, V.N., M.E. Cruickshank, and D.E. Parkin, Auditing the use of colposcopy versus general gynecology clinics to investigate women with postcoital or intermenstrual bleeding: a case for a new outpatient service. Journal of lower genital tract disease, 2007. 11(2): p. 108-111.

[88] Stapley, S. and W. Hamilton, Gynaecological symptoms reported by young women: examining the potential for earlier diagnosis of cervical cancer. Family practice, 2011. 28(6): p. 592-598.