

**DETECTION OF CERVICAL CANCER IN EARLY STAGE AND SIGNIFICANT RISK  
FACTOR ANALYSIS EMPLOYING MACHINE LEARNING BASED APPROACH**

**BY**

**NOWUSIN TABASSUM  
ID: 181-15-1843**

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

Supervised By

**Md. Mahfujur Rahman**  
Sr. Lecturer  
Department of CSE  
Daffodil International University

Co-Supervised By

**Tasfia Anika Bushra**  
Lecturer  
Department of CSE  
Daffodil International University



**DAFFODIL INTERNATIONAL UNIVERSITY**

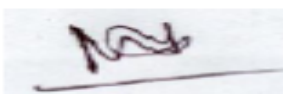
**DHAKA, BANGLADESH**

**JANUARY 2022**

## APPROVAL

This Project titled “Detection of “**Cervical Cancer in Early Stage and Significant Risk Factor Analysis Employing Machine Learning Based Approach**”, submitted by “**Nowusin Tabassum**”, Id No: **181-15-1843** to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on January 2022.

## BOARD OF EXAMINERS



---

**Dr. Md. Ismail Jabiullah**  
**Professor**  
Department of CSE  
Faculty of Science & Information Technology  
Daffodil International University

**Internal Examiner**



---

**Narayan Ranjan Chakraborty**  
**Assistant Professor**  
Department of Computer Science and Engineering  
Faculty of Science & Information Technology  
Daffodil International University

**Internal Examiner**



---

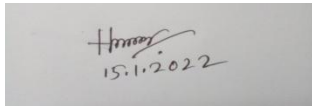
**Dr. Mohammad Shorif Uddin**  
**Professor**  
Department of Computer Science and Engineering  
Jahangirnagar University

**External Examiner**

## DECLARATION

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

### Supervised by:



---

**Md. Mahfujur Rahman**  
Sr. Lecturer  
Department of CSE  
Daffodil International University

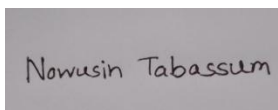
### Co-Supervised by:



---

**Tasfia Anika Bushra**  
Lecturer  
Department of CSE  
Daffodil International University

### Submitted by:



---

**Nowusin Tabassum**  
ID: -181-15-1843  
Department of CSE  
Daffodil International University

## ACKNOWLEDGEMENT

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

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

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

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

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

## ABSTRACT

Cervical cancer (CC) is the most prevalent and second major reason of death of mortality in women in third world nations when compared to certain other vaginal cancers. It is curable if caught in its early stages. From that standpoint, the study aims to develop an appropriate predictor and computer models for detecting CC at a preliminary phase. Cervical cancer detection in the clinic is extremely expensive. No one wants to go for a clinical test when they have cervical cancer in its early stages. As a result, Machine Learning detection is extremely beneficial. The technology we suggest will detect cervical cancer at an early stage and at a reasonable cost. A CC dataset is compiled with four class attributes such as biopsy, cytology, hinselmann, and schiller, and the dataset is divided into four groups based on target attributes. The dataset was prepared in the data preprocessing phase for better analytical result in the further analysis. Then we applied statistical and EDA approach to discover hidden knowledge from the dataset. To develop machine learning model, different supervised machine learning algorithm like Decision Tree Classifier, Logistic Regression, XG Boost, Multilayer Perception and Random Forest are applied to the dataset to find an efficient classifier. Then, the performances of all the applied classifiers are compared based on accuracy, precision, recall, sensitivity, f-measure, AUROC, and kappa statistics. We found that RF provided the best performance for biopsy with 94.57% accuracy. MLP and LR generated 98.06% accuracy as the best performing classifier for cytology. Besides, MLP and XGB generated the best performance for hinselmann with 96.51% accuracy, where MLP produced the best result with 94.53% accuracy. Then, we applied four FST methods to rank and show the feature importance for the target feature. Overall, the findings of the study specifies that the proposed model is highly potential to detect CC in early stage.

Keywords: *Random Forest, FST, Cervical Cancer, Multi-layer perceptron, XGBoost.*

## TABLE CONTENTS

<b>CONTENTS</b>	<b>PAGE</b>
Approval	ii
Declaration	iii
Acknowledgement	iv
Abstract	v
<b>CHAPTER</b>	
<b>CHAPTER 1: INTRODUCTION</b>	<b>1-2</b>
1. Introduction	1
1.1 Research’s Motivation	1-2
1.2 Research’s Objective	2
1.3 Expected Outcome	2
<b>CHAPTER 1: LITERATURE REVIEW</b>	<b>3-4</b>
2.1 Related Works	3-4
<b>CHAPTER 3: RESEARCH METHODOLOGY</b>	<b>5-10</b>
3. Proposed Method	5
3.1 Data Collection	5
3.2 Data Pre-Processing	6
3.3 Machine Learning Algorithms	7-9
3.4 Performance Measurement Criteria	9
3.5 Apply FST Methods	10

<b>CHAPTER 4: RESULTS &amp; DISCUSSION</b>	<b>11-16</b>
4.1 Result of Exploratory Data Analysis (EDA)	11-14
4.2 Result of Machine Learning Analysis	14-15
4.3 Feature Importance Score	15-16
<b>CHAPTER 5: CONCLUSIONS AND FUTURE WORK</b>	<b>17</b>
<b>REFERANCES</b>	<b>18-20</b>

## LIST OF FIGURES

<b>FIGURES</b>	<b>PAGE NO</b>
Figure 1: Research Methodology	5
Figure 2: Box plot to show data distribution and outliers	11
Figure3: Heatmap to show correlation among all the numeric features and significance of them	12
Figure 4: ROC curve, Precision and Recall for biopsy	13
Figure 5: ROC curve, Precision and Recall for cytology	13
Figure 4: ROC curve, Precision and Recall for hinselmann	13
Figure 4: ROC curve, Precision and Recall for schiller	14
Figure 8: Feature importance score of datasets	16



**LIST OF TABLE**

<b>TABLE</b>	<b>PAGE NO</b>
Table 1: Attribute Information	6
Table 2: Performance Measurement Criteria	9
Table 3: Apply FST Methods	10
Table 4: Performance comparison of applied classifiers	14-15

# CHAPTER 1

## INTRODUCTION

Cancer is becoming the second leading cause of death in the world, due to the velocity of social rhythms, people's irregular daily lives and rest, imbalanced dietary habits, and extreme physically and mentally stress [1]. For the past 30 years or above, cervical cancer has been one of the most prevalent cancers found in women, and practically every woman is at danger [2]. According to the American Cancer Society, approximately 14,480 new instances of cancer will be diagnosed with the Disease in 2021, with around 4,290 women suffering from cervical cancer [3]. Cervical cancer starts in a woman's cervix and progresses. Human cells take many years to eons to evolve from a precancerous lesion (cervical intraepithelial neoplasia; CIN) to invasive malignant cervical cancer, with a lengthy and recoverable precancerous lesion phase [4].

Cervical cancer can strike any woman at any age, although it strikes women between the ages of 30 the most frequently. The human papillomavirus (HPV) is a virus that spreads from one person to another through sexual activity and is the main reason of cervical cancer. At some point in life, this virus spreads at up to half of all sexual intercourse persons [5]. As a result, the time of the first intimate intercourse, the rate of sexual partners, the amount of deliveries, and contraception use have all been linked with an increased risk of cervical cancer. Cervical cancerization can be avoided, and the normal outcome of cervical cancer for people with precancerous lesions can be changed, if certain contributing factors are adequately controlled [4,8].

### 1.1 Research's Motivation

Women were questioned in order to assess motivational factors. The majority of the 70percentage points who had the checkup were prompted by doctor recommendation instead of direct general publicity, according to the findings. While routine cervical Smear test and cytological screening can diagnose CC in its early stages, this needs a significant amount of resources and patient commitment, therefore mortality remains high in many countries. Surgery and/or radiotherapy are efficient treatment methods, but they have a number of negative side effects on patients, which are more severe if the CC is not detected early. Many factors contribute to low involvement, also with a low level of awareness and the high expense of treatment, but the result is that many cases of CC are diagnosed late, when treatment is more difficult and survival is less likely. Ladies who had not undergone the test, most of whom were elderly and poor, were not hostile, but generally unmotivated. Lay teaching to encourage women to receive medical help, as well as doctor instruction to encourage and carry out their part in routine annual cytological analysis of cervical exudates, might virtually eliminate cervical cancer death. However, because ML has been shown to have strong predictive value for many diseases, this problem could be tackled utilizing ML methodologies. In that light, the study's goal is to

create a machine learning model with high efficiency and accuracy that can detect early-stage cancer at a low cost.

## **1.2 Research's Objective**

Every woman in the globe has the ability to recognize CC at an early stage. The budget of treatment will be decreased. Early identification of cervical cancer will raise awareness with all women across the country. Because early-stage CC treatment can cure the condition, the mortality rate will be minimized. Using the suggested ML model, doctors and clinicians would be able to decide on treatment immediately after diagnosing CC.

## **1.3 Expected Outcome**

The report's aim is to create a machine learning-based model for detecting cervical symptoms early on and detecting the most important risk factors.

## CHAPTER 2

### LITERATURE REVIEW

#### 2.1 Related Works

Cervical cancer categorization seems to be the subject of numerous studies. To discover and evaluate the existence of cervical cancer, researchers used a medical aspects method, a biological aspects method, and picture categorization and fragmentation. To improve accuracy and reduce categorization mistakes of false positive and false negative reports, and to discover the most associated factors of cervical cancer, all of the above approaches employ various categorization and segmentation algorithms. The studies that follow focus on medical content techniques, with similar works included in next section.

Muhammed F. and et al. in 2017 proposed a Machine Learning method which can detect cervical cancer [6]. They used different approach like Multilayer Perceptron, BayesNet and k-Nearest Neighbor. Kemal Akyol in 2018 used Test Variable Selection method and Balanced Data to investigate the cervical cancer [7]. On their dataset, they use the Random Under-Sampling (RUS) and Random Over-Sampling (ROS) approaches. In their data collection, there are 190 missing values. If all missing data is available, it may be more effective. The results revealed that the ROS-based SS approach outperformed the RUS-based SS method. Abdoh et al. in the time of 2018 proposed a machine learning approach where they suggested Random Forest Classifier with SMOTE and Feature Reduction Techniques which can detect cervical cancer [8]. Random Forest was employed to determine the risk factors. Sequential backward removal and classification techniques were the two strategies employed in this study to choose the features. Whichever the case may be, they refused to explain why features extraction approaches are used. Because increasing the precision was ineffective.

A. Ghoneim and et al. in the year of 2019 proposed a cervical cancer cell detection and classification system based on convolutional neural networks (CNNs) [9]. To extract deep-learning features, the cell pictures are loaded into a CNNs model. Transfer learning and fine tuning are utilized to implement the CNN model. Their investigation was carried out with the help of the Herlev database. The proposed method of CNN-ELM-based system provided 91.2% accuracy in the classification problem. As we can see the CNN-ELM-based system which we think does not give a good accuracy to classification problem. Xiaoyu Deng et al. in 2018 consummated that risk factor can analysis by using different algorithm for cervical cancer [10]. They used different methodology like SVM, Decision tree, Random Forests and XGBoost. The top five risk factors which affect the diagnosis most were found but, in their methodology, XGBoost and Random Forest. SVM were less effective than two others. Accuracy of SVM is only 90.34%, which is not satisfactory to detect cervical cancer.

In addition to these, in 2006, J. Jantzen and et al. proposed the pap-smear benchmark database provides data for comparing classification methods. They employed a subset of the Herlev database for their methodology [11]. They detected 108 normal cells and 41 malignant cells in their subset of 149 cell pictures. We can't draw any conclusions regarding their system methods based on such a short sample. Başaran et al. in 2015 looked at traditional procedures as well as contemporary imaging technologies such magnetic resonance imaging, positron emission tomography, and computed tomography in cases of cervical cancer [12]. Wang et al. suggested a cervical cell segmentation and classification method [13]. They used shape and texture features, as well as Gabor features and the SVM, to classify the data. The categorization accuracy for normal and cancerous cells was over 89 percent. A private database was created by the authors. In 1992, Benedet et al. used a scoring method to perform a vaginal ultrasound elastography on 113 women (13 patients with cervical cancer) [14]. There were considerable changes between the normal cervical and cervical cancer elastic pictures, however the difference did not have statistical significance.

A dataset obtained from Kaggle was categorized in our research. An specialist might perform the same categorization, but to improve impartiality, the outputs must be evaluated using machine learning algorithms. As a result, machine learning algorithms were used to classify the data.

## CHAPTER 3

### RESEARCH METHODOLOGY

The processes or strategies used to find, collect, organize, and managing data from varied sources are referred to as research methods. The methodological portion of a scientific report lets the reader examine the research main accuracy and dependability. The existing approach for obtaining an accurate diagnosis of cervical cancer is shown in Figure 1.

### 3. Proposed Method

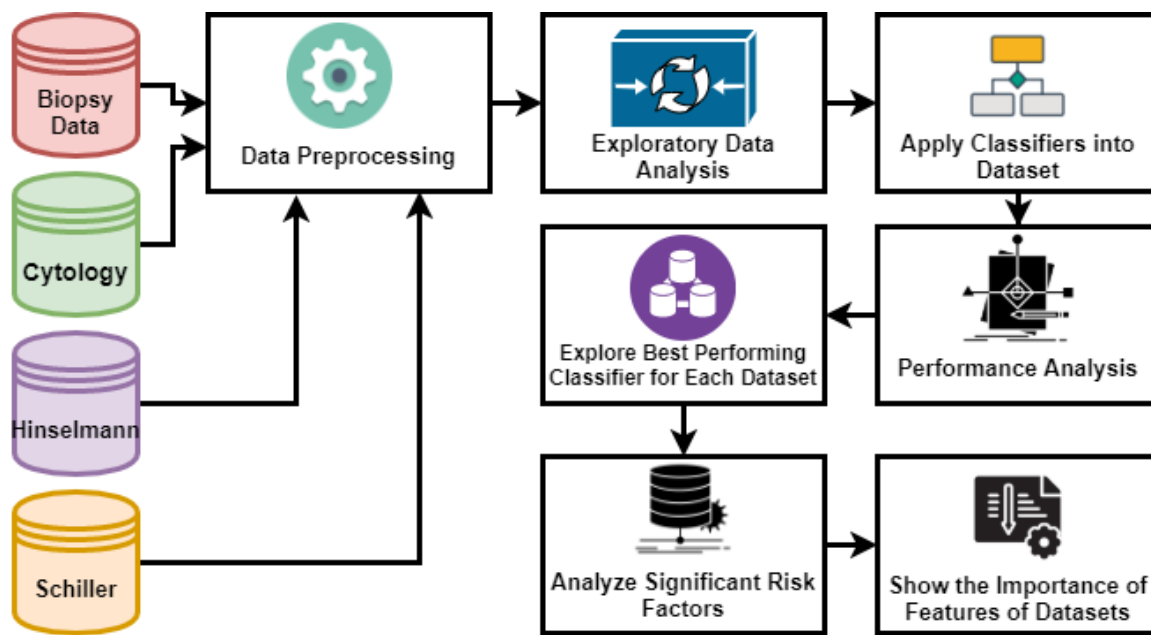


Figure 1: Works prototype of proposed model

Here we're going to preprocess four different types of datasets. The data was then analyzed. After that, we applied various classifiers before looking at the performance analysis. The dataset's best performance classifier is investigated, followed by an analysis of major risk variables and a presentation of the dataset's most important features.

#### 3.1 Data Collection

The dataset was collected from the UCIML repository's dataset archive. Patients' demographic data, behaviors, and prior health records are all included in the dataset. There are 36 attributes and four target attributes which are Biopsy, Cytology test, Hinselmann test result and Schiller test. In the table, there are 858 instances from the dataset [19].

**Table 1**  
**Attribute Information**

<b>Feature</b>	<b>Type</b>	<b>Feature</b>	<b>Type</b>
Age	Numeric	STDs:pelvic inflammatory disease	Binary
Number of sexual partners	Numeric	STDs:genital herpes	Binary
First sexual intercourse	Numeric	STDs:molluscum contagiosum	Binary
Num of pregnancies	Numeric	STDs:AIDS	Binary
Smokes	Binary	STDs:HIV	Binary
Smokes (years)	Binary	STDs:Hepatitis B	Binary
Smokes (packs/year)	Binary	STDs:HPV	Binary
Hormonal Contraceptives	Binary	STDs: Number of diagnoses	Numeric
Hormonal Contraceptives (years)	Numeric	STDs: Time since first diagnosis	Numeric
IUD	Binary	STDs: Time since last diagnosis	Numeric
IUD (years)	Numeric	Dx:Cancer	Binary
STDs	Binary	Dx:CIN	Binary
STDs (number)	Numeric	Dx:HPV	Binary
STDs:condylomatosis	Binary	Dx	Binary
STDs:cervicalcondylomatosis	Binary	Biopsy: target variable	Binary
STDs:vaginalcondylomatosis	Binary	Cytology: target variable	Binary
STDs:vulvo-perinealcondylomatosis	Binary	Hinselmann: target variable	Binary
STDs:syphilis	Binary	Schiller: target variable	Binary

### 3.2 Data Pre-Processing

Because the efficiency of a machine learning strategy is dependent on how effectively the dataset is organized and formatted, data preprocessing is required for any machine learning or data mining strategy. After handling missing values with a Replace Missing Values filter, another detector known as the Inter quartile Range (IQR) was used to identify outliers and excessive values during the pre-processing stage. The IQR is a way of calculating the variability around a dataset's median. Outliers is a piece of data that falls beyond the anticipated range of observations and could be presumed to be attributable to reporting mistakes or other unimportant occurrences for the objectives of the study. The performance will be determined by how we organize and analyze data. We can improve machine learning accuracy by preprocessing data.

For numerical data, the mean is utilized and for binary data, the mode approach is employed. The Biopsy, Cytology, Hinselmann, and Schiller dataset preprocessing systems are all similar. As a result, several exploratory data analysis (EDA) were carried out (like box plots) to ensure that the dataset was clear of misfits, and the information was displayed as IQR and heatmap to discover correlation between the attributes of infected and non-diseased persons as per age demographics.

### 3.3 Machine Learning Algorithms

The dataset was subjected to five (5) classification algorithms in determining the most effective performing method prediction accuracy and some other statistical properties. Decision Tree (DT), Logistic Regression (LR), XGBoost (XGB), Multilayer Perceptron (MP), and Random Forest (RF) were always the techniques used.

#### I. Decision Tree Classifier

One of the most widely used machine learning algorithms is the Decision Tree (DT) [5]. It's applied to a set of data in order to do classification or regression analysis. Based on a series of questions, this program separates the data into several categories. The procedure starts with the fundamental node, also known as the tree's root, which contains all samples. In this tree-structured predictor, the edges represent dataset features, the branching represent equations, but every leaf offers the decision. It's called a decision tree because it's structured like a tree, starting with the base and growing into a tree-like structure with more nodes. Decision Tree has the advantage of being able to deal with a wide range of input data types, analyze missing values, and attain acceptable initial accuracy while being simple to construct. [15]. Decision trees are expansive and easier to decode [16].

$$E(S) = \sum_{j=1}^c -p_j \log_2 P_j$$

#### II. Logistic Regression

Logistic regression (LR) is a statistical technique that has become increasingly popular in medical research in recent decades [17]. When the dependent variables are binary, it is utilized to examine a dataset. The link between one dependent binary variable and one or more independent variables can be determined using LR as a prediction model. The logistic curve produced by logistic regression ranges from 0 to 1 [5]. This regression is similar to linear regression, only it includes the natural logarithm of the odds for the target variables in the curved construction process rather than probabilities. Moreover, each group's indicators do not need to have equal probability or a normal probability plot.

$$y = \alpha_o + \alpha_1 Z_1 + \alpha_2 Z_2 + \dots + \alpha_n Z_n$$



The dependent variables is  $y$ , and the predictor variables are  $Z_1, Z_2, Z_3, \dots, Z_n$ . The logistic function can be obtained by applying the sigmoid function to the equation.

$$l = 1/[1 + e^{-(\alpha_0 + \alpha_1 Z_1 + \alpha_2 Z_2 + \dots + \alpha_n Z_n)}]$$

### III. XGBoost

XGBoost is also known as a scalable machine learning system for tree boosting, and it is frequently used in algorithm competitions, where it gives better solution than other algorithms [18]. On a single machine, the system is more than ten times faster than previous methods, and it scales to billions of samples in distributed or memory-limited environments. When dealing with unstructured data prediction difficulties we use XGBoost algorithm. It is a great blend of software and hardware combinatorial optimization that produce successful outcomes in the smallest period of time with the lowest number of computational power.

$$Obj(\theta) = \sum_i^n l(y_i - \hat{y}_i) + \sum_{k=1}^n \Omega(f_k)$$

### IV. Multilayer Perceptron (MLP)

MLP is such a well neural network-based categorization with three or more layers, including an input layer, a convolution layers, and first or perhaps more processing elements seen between output neuron [20]. The input layer receives the data that needs to be processed. Predicting and classification are two tasks that fall under the outputting layer's purview. The MLP's linked core is a random arrangement of hidden units added between both the input and output layers. Each layer has a set of 'neurons' that connect the layers together. MLP is a numerous non-linear translation methodology that uses supervised machine learning methods to find and expand from training phase. The usage of proper input variables and design and analysis specifications is required by MLP learner [21].

$$e_k(n) = d_k(n) - c_k(n)$$

### V. Random Forest (RF)

RF is a data classification approach achieved by the proposed learning and DT [22]. While in the training stage, it generates a vast number of trees as well as a forest of decision trees [23]. Every tree in the forest predicts the class label for each and every event during the testing period. When each tree predicts a class label, the final selection for each test data is made via majority vote [24]. The class label that receives the most votes is considered the most appropriate label for the test data. This cycle is repeated for each piece of data in the collection. The best appropriate randomized responsible for a considerable for this experiment was 123, that offered the best effectiveness for the presented collection.

$$\int_n^i(x) = \frac{1}{N^e(A_n(x))} \sum_{y_i \in A_n(x) | I_i=e} Y_i$$

### 3.4 Performance Measurement Criteria

**Table 2**

**Performance Measurement Criteria**

Metrics	Explanation	Formula
Accuracy	The proportion of observations that are correctly categorized [25].	$A = \frac{TP + TN}{TP + TN + FP + FN}$
Sensitivity	It is a breakdown of the real positive vs. all the expected positives [26].	$S_n = \frac{TP}{TP + FN}$
Specificity	This calculates the percentage of real negatives compared all expected negatives [27].	$S_p = 1 - \left(\frac{FP}{FP + TN}\right)$
Precision	Precision is defined as the ratio of accurately predicted positives to all predicted positives. [28].	$P = \frac{TP}{TP + FP}$
Recall	It is a machine learning model's prediction of the system is characterized of True Positives. [29].	$R = \frac{TP}{TP + FN}$
F-Measure	It is the balanced harmonic mean of the algorithm's precision and recall, and it is used to calculate a model's accuracy [30].	$F = \frac{2 * TP}{2 * TP + FN + FP}$
AUROC	An ROC is a simple medical test screening test that is created by comparing actual performance against the false positive rate at estimation circumstances [31].	$TR = \frac{TP}{TP + FN}$ $FR = \frac{FP}{FP + TN}$
Kappa Statistics	It assesses the performance of qualitative features based on inter-rater interaction that is expected and witnessed [32].	$K_p = 1 - \frac{1 - p_o}{1 - p_e}$

True positive and true negative are represented by TP and TN, while false positive and false negative are represented by FP and FN, correspondingly [37].

### 3.5 Apply FST Methods

**Table 3**  
**Apply FST Methods**

CFST	Abbreviation	Details	Formula
Correlation based Feature Subset Selection	CFSSE	It determines the relevance of a subset of qualities by assessing the unique predictive capacity of each functionality as well as the level of variance among them [33].	$F_s = \frac{N * ra}{N + N(N - 1)r_n}$
Gain Ratio based Attribute Evaluation	GRAE	It calculates the gain ratio in terms of the class to highlight the worth of a feature. [34].	$GR(C, A) = \frac{(H(C)_H(C A))}{H(A)}$
Info Gain based Attribute Evaluation	IGAE	It calculates the gain of knowledge in terms of the class to assess the significance of a performance [34].	$IG(C, A) = (H(C)_H(C A))$
ReliefF based Attribute Evaluation	RFAE	It evaluates the impact of an attribute by testing a case periodically and assessing the values for a particular attribute for the closest case of the same kind of independent class [35].	$R_x = P(\text{diff } X   \text{diff class}) - P(\text{diff } X   \text{same class})$

## CHAPTER 4

### RESULTS & DISCUSSION

Various approaches were examined in this work, and the ones that performed the best were reported. The data was separated into two categories: training and testing. So, same train data was used to train all of the algorithms inside this work. Accordingly, the same test set of data was used to evaluate all of the approaches. However, neither of the test sample values appears in the train dataset.

#### 4.1 Result of Exploratory Data Analysis (EDA)

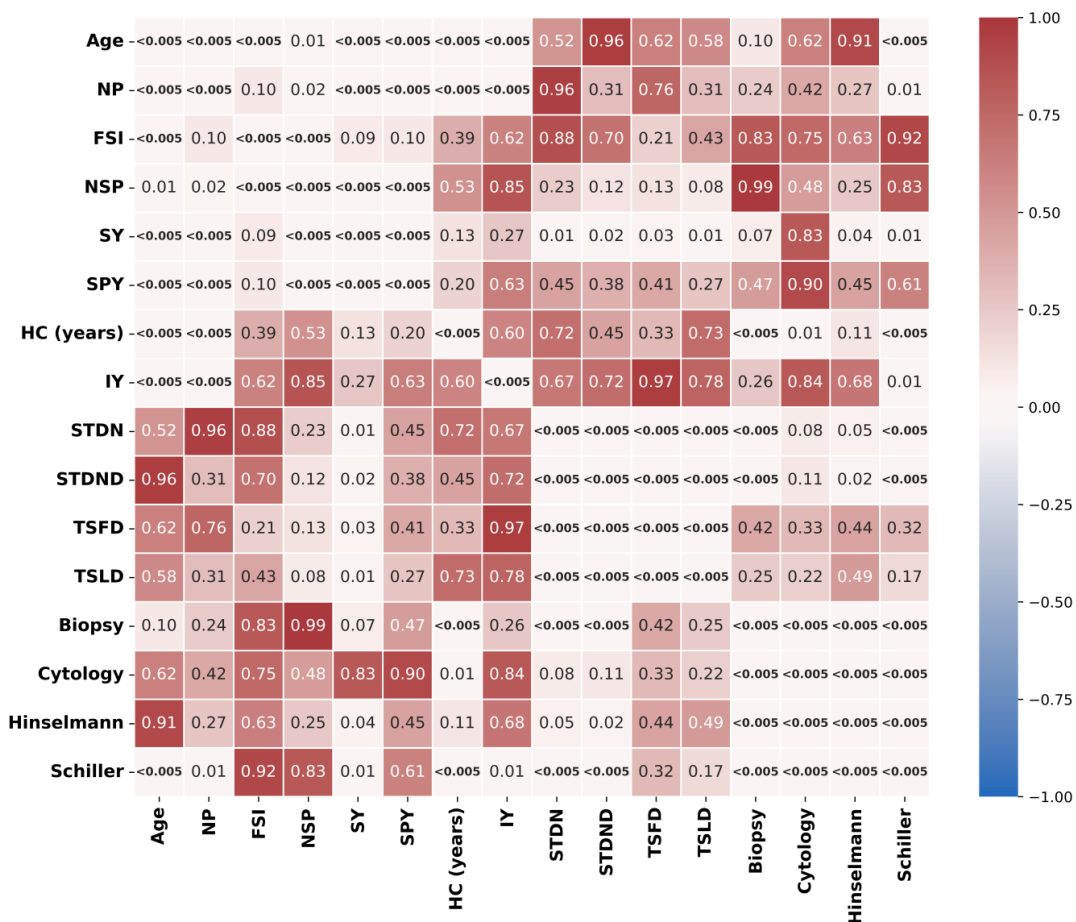


Figure 3: Heatmap to show correlation among all the numeric features and significance of them

The associated values and correlations between attributes are represented by a heatmap in Figure 3. the hue of each colorful cell represents the strength of the relationships among two attributes

and their corresponding values. Negative association is shown by a reliability coefficient less than zero, whereas no association is indicated by a correlate mean of 0.

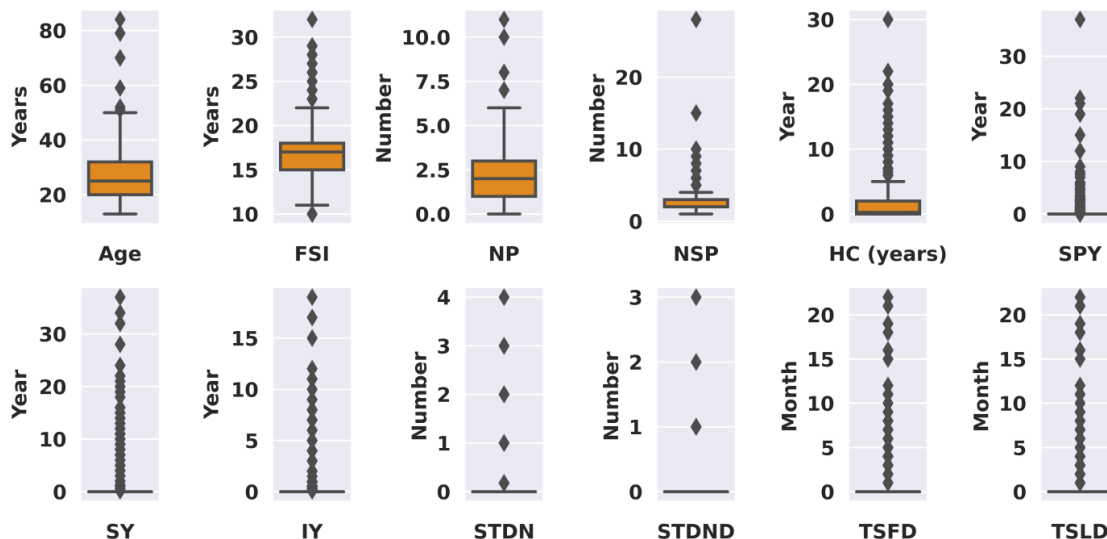


Figure 2: Box plot to show data distribution and outliers

Boxplots are a way to see whether a data set's data is spread. This graphs show the data set's lowest, highest, median, first quartile, and third quartile. Creating boxplots for every data set allows you to compare the spread of data among data sets. The distribution of the dataset's quantitative properties is depicted in Figure 2. Outliers are values or points outside of the boxes and whiskers. This diagram shows all of the observed outliers at the start of the process. The inter-quartile range was used to identify outliers, which were then eliminated from the dataset. After this filtering, there are no outliers in this dataset, as seen in the picture. Following the removal of all outliers, the remaining instances of the dataset are used for further analysis.

## 4.2 AUC with Precision and Recall

The AUC is a description of the ROC curve and measures a classifier's right to differentiate across classifications. The AUC measures how well the researchers have distinguished among positive and negative classifications. The greater the AUC, the stronger the model's accuracy.

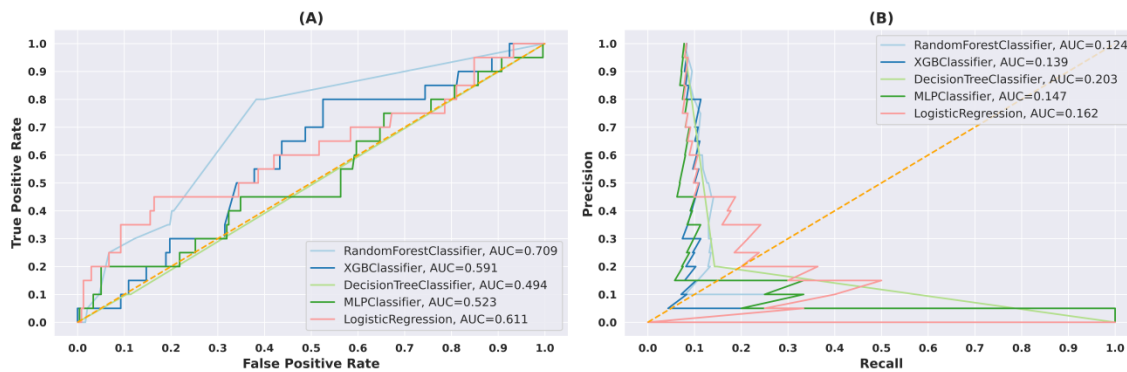


Figure 4: ROC curve, Precision and Recall for biopsy

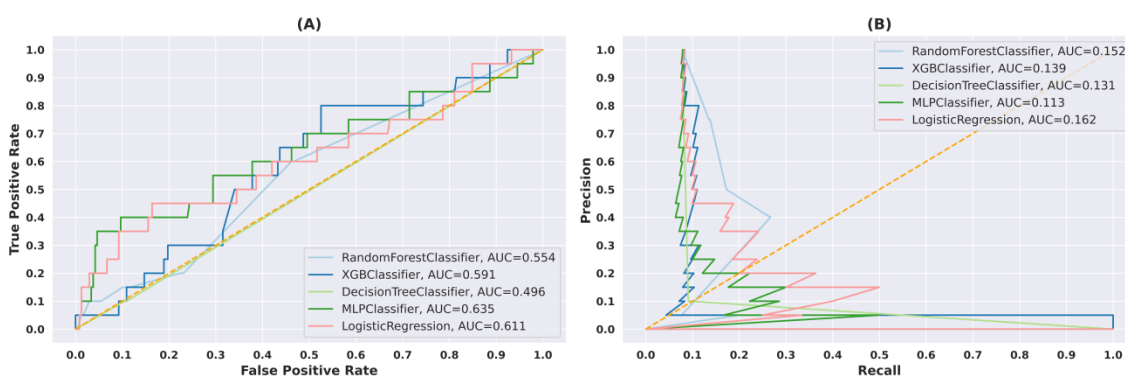


Figure 5: ROC curve, Precision and Recall for cytology

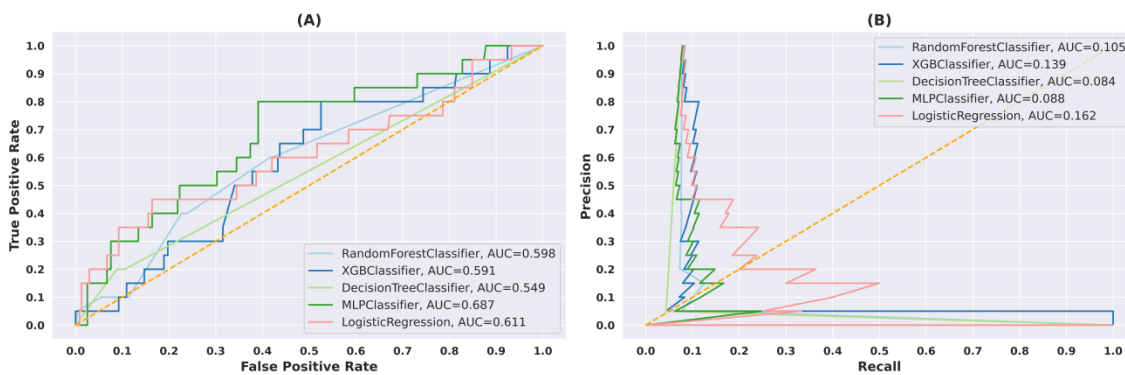


Figure 6: ROC curve, Precision and Recall for hinselmann

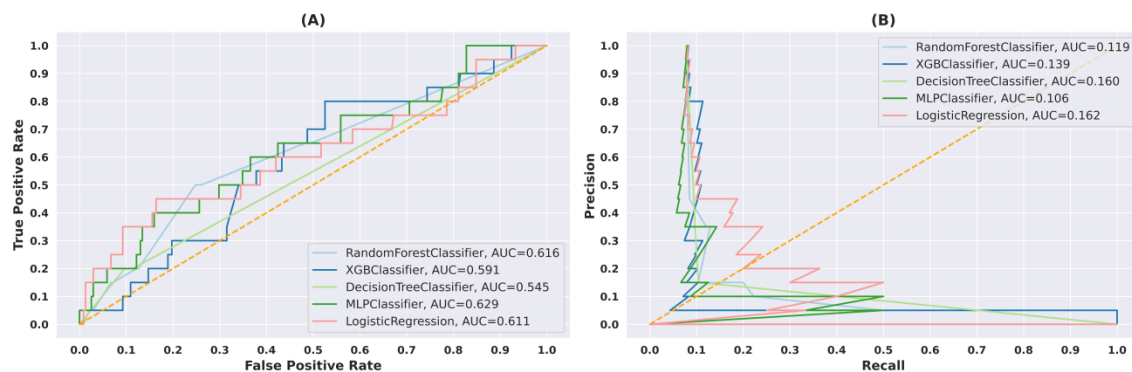


Figure 7: ROC curve, Precision and Recall for schiller

Figure (A) represent the ROC curve and figure (B) represent the precision and recall curve. In figure (A) we can see ROC curve for four different dataset, where it has true positive rate and false positive rate. In figure (B) we can see precision and recall curve for four different dataset.

### 4.3 Performance Analysis

The cervical cancer illness information was analyzed for this research, factors were identified and deleted, and a variety of classifiers, including MLP, DT, RF, LR, and XGB, were used.

**Table 4**

**Performance comparison**

Dataset	Algorithms	Accuracy	Precision	Recall	F1 Score	Log loss	AUC
Biopsy	DT	91.86	0.92	0.92	0.92	2.81	0.494
	MLP	93.02	0.93	0.93	0.93	2.41	0.523
	LR	93.02	0.93	0.93	0.93	2.27	0.611
	XGB	93.02	0.93	0.93	0.93	2.27	0.591
	RF	94.57	0.95	0.95	0.95	1.87	0.709
Cytology	DT	93.41	0.93	0.93	0.93	2.27	0.496
	MLP	98.06	0.98	0.98	0.98	0.66	0.635
	LR	98.06	0.98	0.98	0.98	0.66	0.611
	XGB	97.67	0.97	0.97	0.97	0.80	0.591
	RF	94.18	0.94	0.94	0.94	2.00	0.554
Hinselmann	DT	93.02	0.93	0.93	0.93	2.40	0.549
	MLP	96.51	0.96	0.96	0.96	1.20	0.687
	LR	96.12	0.96	0.96	0.96	1.33	0.611
	XGB	96.51	0.96	0.96	0.96	1.20	0.591
	RF	94.18	0.94	0.94	0.94	2.00	0.598
Schiller	DT	82.17	0.82	0.82	0.82	6.15	0.545
	MLP	94.57	0.94	0.94	0.94	1.87	0.629

	LR	92.63	0.92	0.92	0.92	2.54	0.611
	XGB	93.79	0.93	0.93	0.93	2.14	0.591
	RF	92.63	0.92	0.92	0.92	2.54	0.616

The measurements of all of these different classifiers were evaluated on the datasets to discover the best performing algorithm for predicting cervical cancer. The performance outcome metrics of the classification algorithms used, specifically sensitivity, specificity, and accuracy, are shown in Table 4. The factors RF, MLP, LR, and XGBoost all produce positive results. We discovered that RF had the best performance for biopsy, with 94.57 percent accuracy, while MLP and LR had the best classifier performance, with 98.06 percent accuracy. Furthermore, MLP and XGB produced the best results for hinselmann, with 96.51 percent accuracy for MLP and 94.53 percent accuracy for XGB.

### 4.3 Feature Importance Score

Feature Importance refers to methods for calculating a value for one of a figure's input data; the values actually describe the "importance" of every feature. A higher value indicates that a particular feature would have a greater influence on the system which used forecast a particular classification.



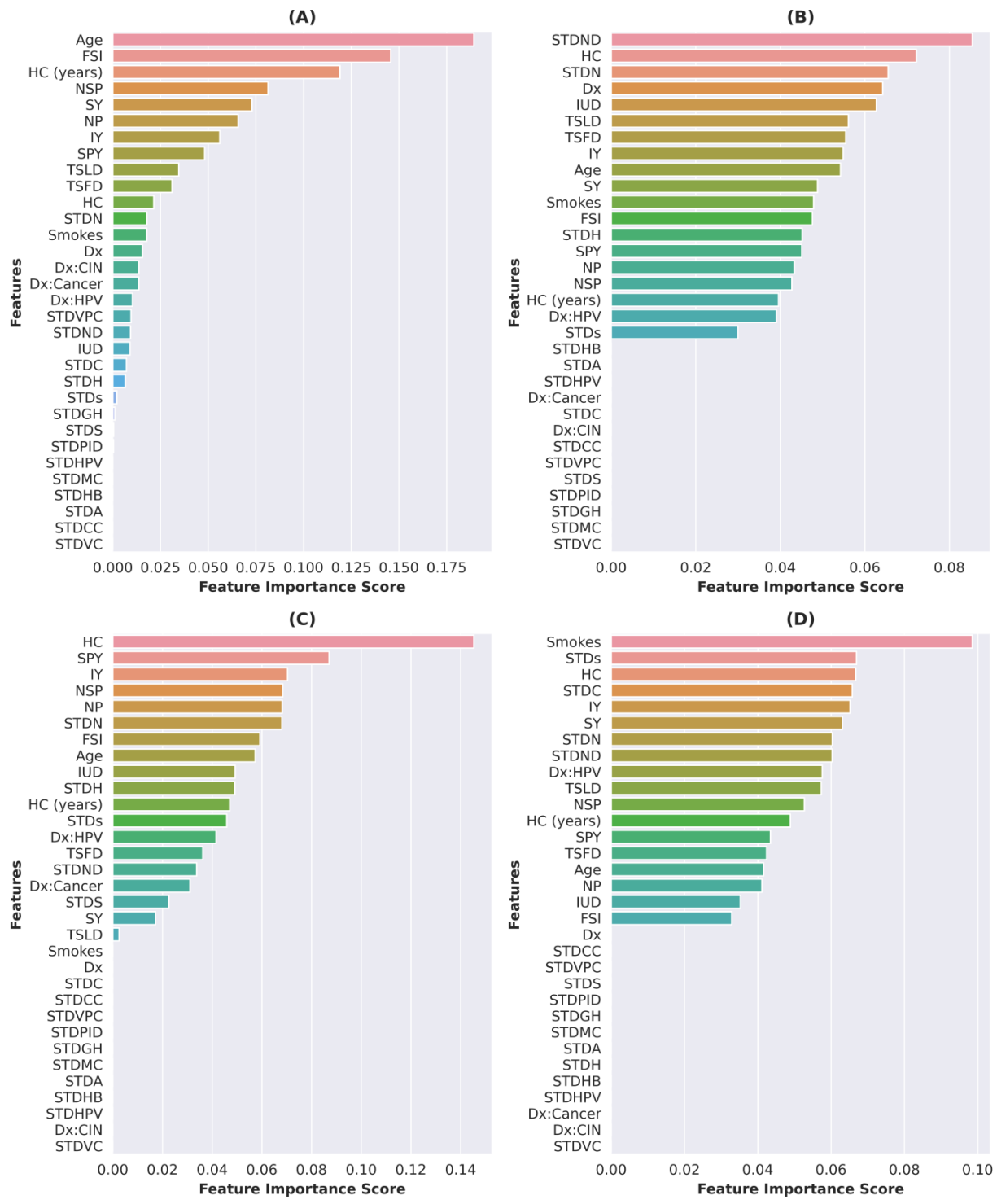


Figure 8: Feature importance score of datasets

## **CHAPTER 5**

### **CONCLUSIONS AND FUTURE WORK**

According to the American Cancer Society, 14,480 new instances of CC are identified each year, and 4290 persons die as a result of the disease. However, if correct therapy is received, it is feasible to recover. It is only achievable if it is discovered at an early stage of the CC. However, the major reason for not diagnosing and detecting CC at an early stage is a lack of education, awareness, and the high expense of diagnostics. The study's goal was to build an effective machine learning model to help physicians and individuals identify and detect CC. We suggested a methodology, as well as suitable predictor, to help clinicians predict and identify CC with greater accuracy without the need of clinical tests. In this study, the characteristics were also rated using several features ranking algorithms, allowing clinicians to examine the risk variables. Overall, the system will be a valuable tool for clinicians and patients in detecting CC at an early stage. In the future, we will use increasingly advanced and up-to-date methodologies to create a more efficient method for identifying CC at an early stage.

## REFERENCES

1. Shi, W., Wang, Y., Hou, L., Ma, C., Yang, L., Dong, C., Wang, Z., Wang, H., Guo, J., Xu, S. and Li, J., 2021. Detection of living cervical cancer cells by transient terahertz spectroscopy. *Journal of Biophotonics*, 14(1), p.e202000237.
2. Singh, S.K. and Goyal, A., 2020. Performance Analysis of Machine Learning Algorithm for Cervical Cancer Detection. *International Journal of Healthcare Information Systems and Informatics (IJHISI)*, 15(2), pp.1-21.
3. The American Cancer Society <https://www.cancer.org/cancer/cervical-cancer/about/key-statistics.html> [Accessed on 22-06-2021]
4. Lu, J., Song, E., Ghoneim, A. and Alrashoud, M., 2020. Machine learning for assisting cervical cancer diagnosis: An ensemble approach. *Future Generation Computer Systems*, 106, pp.199-205.
5. Alsmariy, R., Healy, G. and Abdelhafez, H., 2020. Predicting Cervical Cancer using Machine Learning Methods. *IJACSA thesis. org*.
6. Unlarsen, M.F., Sabanci, K. and Özcan, M., 2017. Determining cervical cancer possibility by using machine learning methods. *International Journal of Latest Research in Engineering and Technology*, 3(12), pp.65-71.
7. Akyol, K., 2018. A Study on Test Variable Selection and Balanced Data for Cervical Cancer Disease. *International Journal of Information Engineering & Electronic Business*, 10(5).
8. Abdoh, S.F., Rizka, M.A. and Maghraby, F.A., 2018. Cervical cancer diagnosis using random forest classifier with SMOTE and feature reduction techniques. *IEEE Access*, 6, pp.59475-59485.
9. Ghoneim, A., Muhammad, G. and Hossain, M.S., 2020. Cervical cancer classification using convolutional neural networks and extreme learning machines. *Future Generation Computer Systems*, 102, pp.643-649.
10. Deng, X., Luo, Y. and Wang, C., 2018, November. 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)* (pp. 631-635). IEEE.
11. Jantzen, J. and Dounias, G., 2006, November. Analysis of pap-smear image data. In *Proceedings of the Nature-Inspired Smart Information Systems 2nd Annual Symposium* (Vol. 10).
12. Başaran, M., Başaran, A. and Küçükaydın, Z., 2015. Restaging in cervical cancer. *TurkiyeKlinikleri J GynecolObst-Special Topics*, 8(1), pp.117-127.

13. Wang, P., Wang, L., Li, Y., Song, Q., Lv, S. and Hu, X., 2019. Automatic cell nuclei segmentation and classification of cervical Pap smear images. *Biomedical Signal Processing and Control*, 48, pp.93-103.
14. Benedet, J.L., Anderson, G.H. and Maticic, J.P., 1992. A comprehensive program for cervical cancer detection and management. *American journal of obstetrics and gynecology*, 166(4), pp.1254-1259.
15. Ashraf, F.B. and Momo, N.S., 2019, July. Comparative analysis on Prediction Models with various Data Preprocessings in the Prognosis of Cervical Cancer. In *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)* (pp. 1-6). IEEE.
16. Akter, L., Islam, M.M., Al-Rakhami, M.S. and Haque, M.R., 2021. Prediction of Cervical Cancer from Behavior Risk Using Machine Learning Techniques. *SN Computer Science*, 2(3), pp.1-10.
17. Boateng, E.Y. and Abaye, D.A., 2019. A review of the logistic regression model with emphasis on medical research. *Journal of Data Analysis and Information Processing*, 7(4), pp.190-207.
18. Chen, T. and Guestrin, C., 2016, August. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acmsigkdd international conference on knowledge discovery and data mining* (pp. 785-794).
19. <https://archive.ics.uci.edu/ml/datasets/Cervical+cancer+%28Risk+Factors%29>
20. Kwon, K., Kim, D. and Park, H., 2017. A parallel MR imaging method using multilayer perceptron. *Medical physics*, 44(12), pp.6209-6224. Kwon, K., Kim, D. and Park, H., 2017. A parallel MR imaging method using multilayer perceptron. *Medical physics*, 44(12), pp.6209-6224.
21. Tajmiri, S., Azimi, E., Hosseini, M.R. and Azimi, Y., 2020. Evolving multilayer perceptron, and factorial design for modelling and optimization of dye decomposition by bio-synthesized nano CdS-diatomite composite. *Environmental research*, 182, p.108997.
22. L. Breiman, Random forests, *Mach. Learn.* 45 (1) (2001) 5–32. Breiman, L., 1996. Bagging predictors. *Machine learning*, 24(2), pp.123-140.
23. S.M.M. Hasan, M.A. Mamun, M.P. Uddin, M.A. Hossain, February. Comparative analysis of classification approaches for heart disease prediction, in: 2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2), IEEE, 2018, pp. 1–4.
24. J.R. Quinlan, Induction of decision trees, *Mach. Learn.* (1986) 81–106.
25. Singh, P., Singh, S. and Pandi-Jain, G.S., 2018. Effective heart disease prediction system using data mining techniques. *International journal of nanomedicine*, 13(T-NANO 2014 Abstracts), p.121.
26. Mohan, S., Thirumalai, C. and Srivastava, G., 2019. Effective heart disease prediction using hybrid machine learning techniques. *IEEE Access*, 7, pp.81542-81554.

27. Thabtah, F. and Peebles, D., 2020. A new machine learning model based on induction of rules for autism detection. *Health informatics journal*, 26(1), pp.264-286.
28. Saito, T. and Rehmsmeier, M., 2015. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PloS one*, 10(3), p.e0118432.
29. Keilwagen, J., Grosse, I. and Grau, J., 2014. Area under precision-recall curves for weighted and unweighted data. *PloS one*, 9(3), p.e92209.
30. Thabtah, F., 2019. Machine learning in autistic spectrum disorder behavioral research: A review and ways forward. *Informatics for Health and Social Care*, 44(3), pp.278-297.
31. Li, D. and Tian, Y., 2018. Survey and experimental study on metric learning methods. *Neural networks*, 105, pp.447-462.
32. Satu, M.S., Akter, T., Arifen, M.S. and Mia, M.R., 2017. Predicting accidental locations of dhaka-aricha highway in bangladesh using different data mining techniques. *International Journal of Computer Applications*, 165(12).
33. Satu, M.S., Ahamed, S., Hossain, F., Akter, T. and Farid, D.M., 2017, December. Mining traffic accident data of N5 national highway in Bangladesh employing decision trees. In *2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC)* (pp. 722-725). IEEE.
34. Satu, M.S., Tasnim, F., Akter, T. and Halder, S., 2018, February. Exploring significant heart disease factors based on semi supervised learning algorithms. In *2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2)* (pp. 1-4). IEEE.
35. Urbanowicz, R.J., Meeker, M., La Cava, W., Olson, R.S. and Moore, J.H., 2018. Relief-based feature selection: Introduction and review. *Journal of biomedical informatics*, 85, pp.189-203.
36. G. Casalicchio, C. Molnar, B. Bischl, Visualizing the feature importance for black box models, in: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, Springer, Cham, 2018, pp. 655–670.
37. S. Asaduzzaman, M.R. Ahmed, H. Rehana, S. Chakraborty, M.S. Islam, T. Bhuiyan, Machine learning to reveal an astute risk predictive framework for Gynecologic Cancer and its impact on women psychology: Bangladeshi perspective, *BMC Bioinf.* 22 (1) (2021) 1–17.