Predicting Cervical Cancer using Machine Learning Methods

Riham Alsmariy[1], Graham Healy[2], Hoda Abdelhafez[3]
College of Computer and Information Sciences, Princess Nourah University, Riyadh, KSA[1, 3]
School of Computing, Dublin City University, Dublin, Ireland[2]
Faculty of Computer and Informatics, Suez Canal University, Ismailia, Egypt[3]

Abstract*—In almost all countries, precautionary measures are less expensive than medical treatment. The early detection of any disease gives a patient better chances of successful treatment than disease discovery at an advanced stage of its development. If we do not know how to treat patients, any treatment we can provide would be useful and would provide a more comfortable life. Cervical cancer is one such disease, considered to be fourth among the most common types of cancer in women around the world. There are many factors that increase the risk of cervical cancer, such as age and use of hormonal contraceptives. Early detection of cervical cancer helps to raise recovery rates and reduce death rates. This paper aims to use machine learning algorithms to find a model capable of diagnosing cervical cancer with high accuracy and sensitivity. The cervical cancer risk factor dataset from the University of California at Irvine (UCI) was used to construct the classification model through a voting method that combines three classifiers: Decision tree, logistic regression and random forest. The synthetic minority oversampling technique (SMOTE) was used to solve the problem of imbalance dataset and, together with the principal component analysis (PCA) technique, to reduce dimensions that do not affect model accuracy. Then, stratified 10-fold cross-validation technique was used to prevent the overfitting problem. This dataset contains four target variables–Hinselmann, Schiller, Cytology, and Biopsy–with 32 risk factors. We found that using the voting classifier, SMOTE and PCA techniques helped raise the accuracy, sensitivity, and area under the Receiver Operating Characteristic curve (ROC_AUC) of the predictive models created for each of the four target variables to higher rates. In the SMOTE-voting model, accuracy, sensitivity and PPA ratios improved by 0.93 % to 5.13 %, 39.26 % to 46.97 % and 2 % to 29 %, respectively for all target variables. Moreover, using PCA technology reduced computational processing time and increasing model efficiency. Finally, after comparing our results with several previous studies, it was found that our models were able to diagnose cervical cancer more efficiently according to certain evaluation measures.

Keywords*—Cervical cancer; machine learning; voting method; risk factors; SMOTE; PCA

I. Introduction

Cancer is a significant health problem, especially as it is one of the most common causes of death in many countries around the world. Breast, cervical and thyroid cancer are the most common types of cancer among women [1]. In the Kingdom of Saudi Arabia (KSA), cancer statistics are significantly increasing. The total number of cancer cases among women registered in the Saudi Cancer Registry (SCR) is 8,565 and cancer in females accounts for 52.8% of all cancer cases in the KSA. Cervical cancer was the fourth most common cancer among Saudi females in 2015, with 403 cases, representing 6.1 % of all cancer cases diagnosed among Saudi women [2]. In 2010, there were 220 cervical cancer cases among Saudi women, representing 4.1 % of all cancer cases, which indicates an annual increase of 9 % in the number of cervical cancer cases [3]. Since then, the number of cases increased even further, to 1073 by the end of 2018, according to a report by the World Health Organization [4].

Cervical cancer occurs and develops in a woman's cervix and is the leading cause of death from cancer among females. All women, at any age, are at risk of cervical cancer; however, it occurs most often in women who are 30 years of age and over. Human papillomavirus (HPV) is a virus that is transmitted from one human to another during sex and is the leading cause of cervical cancer. This virus infects at least half of the sexually active people at some point in their lives. Nevertheless, cervical cancer can be prevented by using a highly effective vaccine intended to prevent HPV infections [5] and the remaining number of cases can be reduced through early cancer detection using screening tests. If it is diagnosed early, cervical cancer is one of the most responsive to treatment forms of cancer and, thus, recovery can be very high [6]. The increasing of cervical cancer cases and deaths resulting from late diagnosis is the motivation behind this paper.

Cervical cancer is the second most prevalent type of cancer in the world. It arises in the mucous membrane ring that is called the cervical transformation zone, where cancer formed through four possible causes: persistent HPV infections in that zone, viral persistence, the persistence progression of a clones of infected cells that leads to cervical precancer and invasion. The risk of cervical cancer is mainly from infection with HPV and the lack of an effective examination [7].

The massive increase in data over the past years has led to the need to organize, analyze, and extract hidden knowledge from it. During this period, experiments demonstrated the effectiveness of machine learning in assisting experts to analyze data and predict results. Machine learning (ML) is a specific artificial intelligence (AI) branch that collects data from training data. ML technologies allow the computer to obtain knowledge from previous samples and use it to understand patterns from complex datasets. In the medical field, physicians have been able to improve the accuracy of detection, either of the presence or absence of diseases, to predict the disorders and to classify them. Therefore,

researchers are seeking to build better ML models to analyzing medical data to obtain results that would assist the doctors in making correct decisions in diagnosing diseases [8].

This research provides an effective model for improving the performance of using machine learning methods and classification techniques for diagnosing cervical cancer to reduce mortality rates. It focuses also on the sensitivity and overall accuracy of the model to be certain whether patients really have cervical cancer disease or not. The results of this research can assist cancer researchers and physicians in diagnosing of cervical cancer. Then, they can begin treating the disease, thus increasing the patient's chances of recovery.

This paper is organized as follows. Section II discusses literature review and previous work. Section III describes research methodology including imbalance problem, feature selection and classification algorithms. Section IV describes Cervical cancer dataset, data pre-processing and missing data. Section V focuses on implementation and discussing the results. Section VI is the conclusion of the paper and finally Section VII discusses the future work.

II. Literature Review

A. Cervical Cancer and Risk Factors

The National Comprehensive Cancer Network (NCCN) has warned of the necessity of early detection of cervical cancer because the delay in its diagnosis is the main cause of an increase in the number of female deaths in the world [9]. Consequently, numerous medical and scientific research studies have been conducted that examine cervical cancer—its causes, symptoms and methods of detection and prevention. Scientists have also tried to identify the risk factors that cause the occurrence and development of this type of cancer.

Abdoh et al. [10] concluded in their research that the following factors pose the highest risk for the development of this disease: sexually transmitted disease (STDs), intra-uterine device (IUD), hormonal contraceptives and the age at which first sexual intercourse happens. Wu and Zhou [11] claimed that the number of sexual partners, the age when first sexual intercourse happens, the number of smoke packs smoked per year and the number of years that the patient uses hormonal contraceptives increase the possibility of developing cervical cancer. Nithya and Ilango [12] identified ten core features as being most important for predicting cancer.

Age plays a major role in increasing the risk of developing this disease. In Teame et al. [13], the researchers claimed that women 40–49 years of age are twice more likely to have persistent HPV infections than women under 40 and that women with a history of sexually transmitted diseases (STDs) are thrice more likely to have cervical cancer than others. Furthermore, the number of pregnancies, age, number of sexual partners, use of hormonal contraceptives and age at which first sexual intercourse occurred were the five risk factors identified by Deng et al. [14].

All the factors given above are used in this paper to perform analysis and generate results.

B. Related Work

ML algorithms are used in this research to efficiently detect cervical cancer by developing a model inspired by previous research models utilized in the same field.

Abdoh et al. [10] showed in their research that performance could be increased with traditional classification technique when using the synthetic minority oversampling technique (SMOTE) [10]. This study built a classification model using random forest (RF) that was based on cervical cancer risk factors. The results showed that the RF model, after applying SMOTE with all features of cervical cancer risk factors, outperforms the same model after applying two feature selection techniques in term of specificity and positive predictive accuracy (PPA). The two methods for selecting the features used in this study were recursive feature elimination (RFE) and principal component analysis (PCA). However, the researchers did not explain why the use of feature selection techniques was not effective in increasing the accuracy result. The dataset they used was gathered from the Universitario de Caracas Hospital in Caracas and is available at the repository of the University of California at Irvine (UCI) [10]. Wen Wu et al. [11] used the same dataset with three approaches to diagnosing cervical cancer: (1) support vector machine (SVM), (2) support vector machine principal component analysis (SVM-PCA) and (3) support vector machine recursive feature elimination (SVM-RFE). They concluded that SVM works well and gives results in specificity, positive predictive accuracy, and accuracy higher than the other two classifiers.

The voting and deep neural networks (DNN) classifiers were used with the same UCI dataset in [15] to build a model to predicting cervical growth. The voting classifier achieved the highest accuracy (97% to 99%) when compared to a DNN classifier. The author suggested using feature extracting in future works because it could help improve the predictor model.

A study by [14] used three types of ML algorithms to classify the UCI cervical cancer dataset after the Borderline-SMOTE application to handle dataset imbalance. After analyzing the results of the classifiers, XGBoost and random forest were found to better classify malignant and benign cancer than SVM. Because this dataset has a lot of missing values, F. Ashraf et al. [16] used four specific techniques to treat the null values. These techniques are the next observed carried backward (NOCB), last observed carried forward (LOCF), fill with median value and Fill with mode values. They used six ML algorithms: logistic regression (LR), random forest (RF), decision tree, naïve Bayes, neural network (NN) and SVM—to predict the Biopsy target variable. They concluded that the SVM and LR, when used with NOCB pre-processing, achieved the highest Precision, f1-score and accuracy. In another research study [17], Fernandes et al. presented a model that helped reduce learning dimensions and classified the UCI dataset using an artificial neural network (ANN). However, they did not fully explain how they dealt with the null values. In the end, they made a comparison between their model and the baseline model, which contained a deep-fed neural network and acquired a better accuracy than the baseline. This proposed model, through deep learning

techniques accomplished accurate prediction results (the upper area under curve [AUC] = 0.6875).

On the other hand, A. Ghoneim et al. [18] proposed a new and effective model for predicting cervical cancer using the gene sequence module, but it will not be applied in our paper. The data they used consisted of private and public datasets. The private dataset was created from 472 questionnaires obtained from a Chinese hospital, where each patient who filled out her data in the survey had a corresponding gene sequence dataset. The public dataset was obtained from Universitario de Caracas Hospital in Venezuela and it includes 32 risk factors and 858 records. This study also addressed the challenges associated with previous studies on cervical cancer through adopting a voting strategy. This method helped predict disease because it is more practical and scalable effectively.

Unlike the dataset to be used in our paper, the Herlev database was used in the experiment by G. Muhammad et al. [19]. It contained 917 cells and 7 classes, with 3 classes representing normal cells and 4 classes representing abnormal ones. The study gave 242 normal and 675 non-normal images. A cervical cancer prediction and classification model that uses convolutional neural networks (CNNs) was proposed. The deep-learned features were extracted by feeding the cells images into the CNNs model. Subsequently, the extreme learning machine (ELM)-based, multi-layer perceptron (MLP) or autoencoder (AE)-based classifiers classified the input images. This proposed system with the ELM-based classifier accomplished a 99.7 % accuracy in the 2-class detection problem and a 91.2 % accuracy in the 7-class problem.

III. Research Methodology

Choosing appropriate methods and algorithms for a dataset is an essential step in building an efficient and accurate model. This section reviews possibilities for dealing with the imbalance problem in a dataset and appropriate ways of feature selection. Moreover, it discusses the classification methods and algorithms that were applied to the dataset:

A. Classification Algorithms

1) Logistic Regression: Logistic Regression (LR) is a statistical process, which has been increasingly used in medical research, especially in the past two decades. It is used to analyze a dataset when dependent variables are binary. LR as a predictive model helps obtain the relationship between one dependent binary variable and one or more independent variables. LR is distinguished by not assuming a linear relationship between the dependent and independent variables but by displaying a relationship between the output and predictive values [20].

The logistic curve that results from the logistic regression is between 0 and 1. This regression is similar to linear regression, but it uses the natural logarithm of the odds for the target variables in the curve creation process, instead of the probabilities. Furthermore, predictors are not required to have equal variance in each group or normal distribution.

2) Decision Tree: Decision Tree (DT) is one of the most frequently used machine learning algorithms. It is implemented to a dataset with the aim of classification or regression analysis. This algorithm divides the data into various subgroups based on a sequence of questions. The process begins with the primary node, which is called the root of the tree and containes all samples. Each node is split into secondary nodes in either a multi-split or binary form. For the construction of the tree, the "divide and conquer" approach is followed. This approach checks whether all the training samples have the same label or not. Subsequently, the training samples that have different labels are represented in a separate subtree [9], [18].

A DT has several advantages, including the ability to deal with many types of data, the processing of lost values, the ability to achieve good initial accuracy and the ease of implementation [16].

3) Random Forest: Random forest (RF) is one of ML algorithms. It is a supervised classification and ensemble technique that uses a set of decision trees to form a powerful learner. RF applies classification and regression tree (CART) technology to improve a not correlated combination or various decision trees based on bootstrap aggregation technologies. The aim is to find the correct classification and to know the relationship between the dependent variables (y) and independent variables (x) [10], [14].

Each tree is created randomly from a subset of the training set, using approaches like information gain or GINI index to create an independent decision tree (DT). The more trees, the more robust. Features classification and target variable are created independently from each DT, as if the tree votes for that class. Then, if there is a classification problem, the RF selects the classification that obtains the most votes; or, if there is a regression problem, it calculates the mean of all the trees [9], [21].

4) Ensemble Methods: Ensemble Learning (EL) as an effective technique has been adopted in recent years. It expands the traditional machine learning algorithm to combine multiple and stand-alone machine learning algorithms with improving overall classification accuracy. This technique has the advantage of mitigating the problem of having a small sample size by combining and averaging several classification models to decrease the possibility of overfitting the training data. In this manner, the efficiency of the training dataset can be increased, which is critical for various biological applications that have a small sample size. The purpose of using EL methods is to obtain a more accurate classification of training data and better generalization on unseen data [22], [23]. There are several methods of popular ensemble such as boosting, bagging, and voting.

- Bagging

In the bagging method, each classifier is trained on a group of samples and features to get a little varied classification hypothesis. Then, the classifiers are combined to form the ensemble. This method improves the generalization by decreasing variance [22].

- Boosting

In the boosting method, each classifier is trained and combined from the samples, but with several classification weights and different hypotheses. This method achieves the generalization by decreasing bias [22].

- Voting

The voting method is a good strategy to use if one classifier algorithm's defects can be an advantage for another classifier. The voting classifier incorporates the prediction outputs of the classifiers, selecting extremely predicted classes as class variables of test samples [24], as shown in Fig. 1.

There are two voting method types: hard and soft voting. In hard voting, there is one vote for each stand-alone classifier. Then, the class label selected is the one which has a majority, that is more than half the votes. At the same time, the average class label probabilities are used as a voting score in soft voting. Then, the final class label should has the highest voting score or an average probability from each classifier [25].

B. Imbalanced Data

Any dataset can be considered as having an imbalance problem if the number of cases between the classes is not equal. In practice, when applying the classification algorithm to an unbalanced dataset, an exaggerated predictive accuracy is given because the predictive accuracy of the minority class does not exceed 10%. In comparison, the accuracy approaches for the majority class of 100% [20]. Sampling methods represent one of the best solutions for solving the data imbalance problem, which is based on the idea of modifying the distribution of the unbalanced dataset. Several studies have shown that classifiers have better performance with a balanced dataset when compared to an unbalanced dataset. Sampling methods consist of modifying the original dataset, either by increasing the minority class—which is called the oversampling technique—or by reducing the majority class—which is called the under-sampling technique—until the classes are represented approximately evenly [21].

- SMOTE: Synthetic Minority Over-sampling Technique

The simplest way to increase the size of the minority class is a random increase in sampling, but this method can cause overfitting. Hence, a new technique was proposed by Chawla et al. [20] —the synthetic minority oversampling technique (SMOTE)—for reducing the risk of overfitting that occurs when inserting duplicates of cases in the training set based on k-nearest neighbours [20]. SMOTE uses the following equation (1).

$$x_{syn} = x_i + (x_{knn} - x_i) \times t \quad (1)$$

Which x_i for feature vector, x_{knn} for the K-nearest neighbours and t for a random number between 0 and 1 [11].

C. Feature Selection

Feature selection algorithms help increase model performance. These algorithms have many benefits, such as reducing noise in the dataset, helping to increase understanding of the model's classification algorithms, and helping to simplify application, thus improving the model [22].

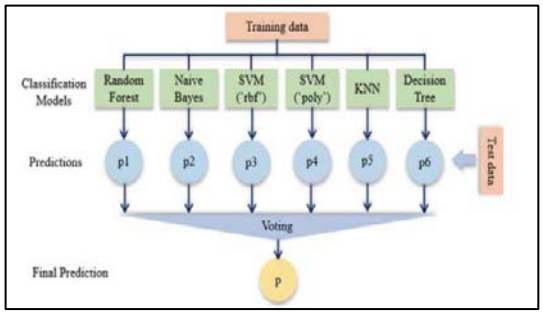

Fig. 1. Overall Structure of the Voting Method.

- PCA: Principal Component Analysis Feature Selection

PCA is a transformation process that can be used to decrease the number of features by extracting new, small, independent features without decreasing the model performance while maintaining the most critical required information contained in the original dataset. The correlated features can be combined as principal components in the statistical dimensionality reduction technique [22]. PCA is a mathematical process that defines the feature orientation based on the advantage of the eigenvector. Where the x-dimension feature space is converted into y-dimension, where $y < x$ and the y-dimension feature space is known as a principal component. Then, the result of the covariance matrix is used to calculate eigenvectors and eigenvalues. the eigenvector with the highest eigenvalue is selected and this is the principal component of the cervical cancer dataset because it determines the important relationships between features the least important data is ignored. Finally, the data is shrunk from a high dimension to a lower one [11].

D. Validation (Cross Validation)

Cross validation (CV) technology refers to a resampling procedure for a limited data sample that can be used for validation and testing ML models. Cross-validation k-fold technology splits the dataset randomly into k (number of folds) identical parts. Then, one part is kept as validation data for model testing, while the residual k-1 parts are utilized as training data. The CV process is then repeated k times as various folds are used each time as the test set. The average of the k results from k-folds is then calculated to obtain a single result [26], [11].

Stratified K-Fold is different from k-fold, and it helps in dealing with an unbalanced set of data. First, stratified k-fold shuffles the data once before splitting and keeps each row with its label. Then, it splits data into k parts. The aim is to have the percentage of samples for each class to be similarly distributed across folds [17].

E. Evaluation Metrics

In biomedical data, the correct diagnosis of a cancer patient becomes important for ensuring a person's health, thus total accuracy is not used alone to evaluate the model. Consequently, several measurements, together with overall accuracy, are used to compare different models for the prediction of cervical cancer and to obtain explanations of diagnosed conditions. In this section, each of these measurements is reviewed.

1) Confusion matrix: The confusion matrix is a technique used for measuring performance in the form of a Table that contains information about both actual and predicted classes, as shown in Table I. If the proposed problem to be studied consists of an n row, this would result in the size of the confusion matrix being n*n, where the rows represent the actual row and the columns represent the expected row. The matrix describes actual and predicted values for two or more classes [27].

- True Positive (TP) indicates the number of correctly classified positive records.
- False Positive (FP) indicates the number of not correctly classified negative samples as positive records.
- True Negative (TN) indicates the number of correctly classified negative records.
- False Negative (FN) indicates the number of not correctly classified positive samples as negative records.

In the confusion matrix, a number of different metrics can be accessed that constitute the essential criteria for measuring model performance, including sensitivity (recall), specificity, f1-score, positive predictive accuracy (PPA) and negative predictive accuracy (NPA), together with overall accuracy.

Accuracy is a common metric, but it is inaccurate when used to measure the performance of an unbalanced dataset. It is the total number of correct predictions that have been achieved over the number of total predictions [24]. It is calculated by equation (2):

$$Total\ Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

Sensitivity, also called the recall is the percentage of positives that are correctly identified from all the positives [8]. It is calculated by equation (3):

$$Sensitivity = \frac{TP}{TP+FN} \quad (3)$$

Specificity is the proportion of negatives that are correctly identified from all the negatives [8]. It is calculated by equation (4):

$$Specificity = \frac{TN}{TN+FP} \quad (4)$$

Precision, also known as Positive Predictive Accuracy (PPA), is the percentage of positive results that are true positive [11]. It is calculated by equation (5):

$$PPA = \frac{TP}{TP+FP} \quad (5)$$

Negative Predictive Accuracy (NPA) is the percentage of negative results that are true negative [11]. It is calculated by equation (6):

$$NPA = \frac{TN}{TN+FN} \quad (6)$$

F1-score is a harmony metric of Precision and Sensitivity on a single parameter and its range values are between 0 and 1, and it is better when it's closer to 1 [16]. It is calculated by equation (7):

$$F1 = 2 * \frac{\text{Precision} * \text{recall}}{(\text{Precisio} + \text{recall})} \quad (7)$$

2) Receiver operating characteristic (ROC) curve and area under curve (AUC): In clinical epidemiology, receiver operating characteristic (ROC) analysis is used to ascertain the accuracy of diagnostic medical tests that can distinguish between two cases of patients: the "diseased" and "non-diseased". It has received increasing attention in evaluating the performance of machine learning algorithms. The ROC curve depends on the idea of a separator scale, which results in outcomes for patients and non-patients.

The ROC curve is a graphical plot, where the Y-axis represents the (sensitivity) that given by equation (8), in contrast to the X-axis, where the (1-specificity) is given by equation (9). The closer the curve is to the upper and left borders of the ROC area, the more accurate the test [28], as shown in Fig. 2.

$$TPR = \frac{TP}{TP+FN} \quad (8)$$

$$FPR = \frac{FP}{TN+FP} \quad (9)$$

While the ROC curve is a perfect visual tool for recognizing a classifier's performance, sometimes a numerical value is needed for comparison purposes. The simplest way to calculate the value of the ROC is to measure the area under curve (AUC). The AUC is the percentage of a box's area under the ROC curve, where its values range from 0 to 1. The classifier's performance increases as the AUC value approaches 1 [24]. It is used to evaluate the performance of classifiers on data with an unbalanced distribution because it is unbiased against a minority class [29]. Also, the AUC of a classifier is equal to the chance that the classifier will rank a positive record as randomly chosen higher than a negative record [30].

TABLE I. THE CONFUSION MATRIX

		Predicted Values	
		Positive (1)	Negative (0)
Actual Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

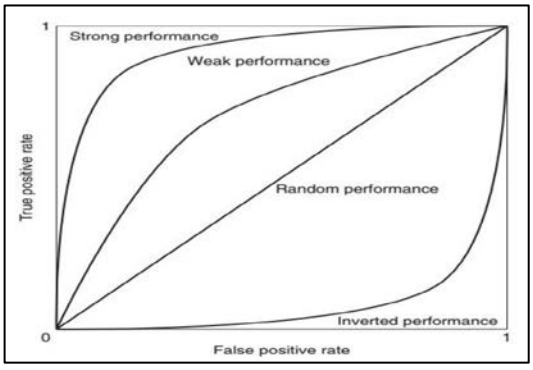

Fig. 2. The ROC Curve.

IV. Data and Analysis

A. Cervical Cancer (Risk Factors) Dataset

Cervical cancer dataset used in this paper was provided publicly by the University of California, Irvine (UCI) Machine Learning Repository, which is a collection of datasets and data generators that are employed by the ML community for the empirical analysis of ML algorithms [31]. David Aha created this archive and fellow graduate students at the university in 1987, as an FTP archive.

The dataset at the Hospital Universitario de Caracas in Caracas, Venezuela was collected in 2017. It focuses on predicting the diagnosis of cervical cancer for 858 cases, while it contains 32 features that display demographic information, habits and historical medical records for these patients as well as four target variables—Hinselmann, Schiller, Cytology and Biopsy—which constitute the main diagnostic methods for cervical cancer [32].

These four target variables were used as classification labels to classify the dataset by machine algorithms [33]. The descriptions and types of the 32 features are shown in Table II.

TABLE II. Descriptionof Data Attributes

No.	Attribute	Type
1	Age	Int
2	Number of sexual partners	Int
3	First sexual intercourse (age)	Int
4	Number of pregnancies	Int
5	Smokes	Bool
6	Smokes (Years)	Bool
7	Smokes (pack/Years)	Bool
8	Hormonal Contraceptives	Bool
9	Hormonal Contraceptives (Years)	Int
10	IUD	Bool
11	IUD (Years)	Int
12	STDs	Bool
13	STDs (number)	Int
14	STDs: condylomatosis	Bool
15	STDs: cervical condylomatosis	Bool
16	STDs: vaginal condylomatosis	Bool
17	STDs: vulvo-perineal condylomatosis	Bool
18	STDs: syphilis	Bool
19	STDs: pelvic inflammatory	Bool
20	STDs: genital herpes	Bool
21	STDs: molluscum contagiosum	Bool
22	STDs: AIDS	Bool
23	STDs: HIV	Bool
24	STDs: Hepatitis B	Bool
25	STDs: HPV	Bool
26	STDs: Number of diagnosis	Int
27	STDs: Time since first diagnosis	Int
28	STDs: Time since last diagnosis	Int
29	Dx: Cancer	Bool
30	Dx:CIN	Bool
31	Dx:HPV	Bool
32	Dx	Bool

B. Cervical Cancer and Risk Factors Dataset Concerns

In this paper, a small cervical cancer and risk factors dataset has been used with a heavy class imbalance. Fig. 3 shows the ratio of positive and negative results for cervical cancer obtained through the main diagnostic methods that represent the four target variables. The synthetic minority oversampling (SMOTE) technique was used to solve the imbalance problem. It is a statistical method that aims to increase the number of records in a balanced way. This method generates new cases in the dataset based on existing minority cases provided as inputs. With SMOTE, majority instances remain unchanged [34]. The use of repeated K-fold cross-validation also plays an active role in the dataset with limited observations [35]. Mitigating the problem of a small dataset is one of the advantages of ensemble technologies. Ensemble learning is an effective method that combines multiple learning algorithms and classification models, which helps improve overall prediction accuracy and reduces the possibility of overfitting the training data [22].

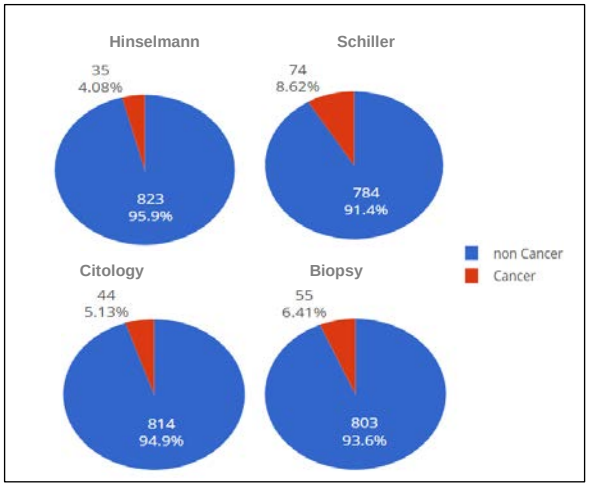

Fig. 3. Summary of the Percentage of the Four Target Variables in the Cervical Cancer Dataset.

C. Data Pre-Processing

The extraction of valuable information and results depends mainly on the quality of the data, while the medical data is affected by some factors that affect its quality, such as missing values, noisy data, inconsistencies, and outliers. Therefore, it is necessary to process the data before starting the machine learning process, where the data pre-processing is an essential step for raising data efficiency. Data pre-processing includes data preparation and dataset transformation that makes knowledge discovery more effective [34]. In this paper, the following steps were used to pre-process data:

1) Missing data: Missing data refer to the data values, which are not stored for a variable or attribute in the dataset [36]. Missing data pose a significant problem in the data analysis process because it is very common for data to be lost, especially in medical data [37]. Missing data is very critical because most analytical methods cannot be applied to an incomplete dataset as this greatly affects the quality of the machine learning model. Therefore, missing values must be

dealt with in calculating missing values with reasonable values [38]. Some algorithms, such as scikit-learn methods assume that all values are not missing and have a meaningful value [39].

The cervical cancer dataset contains many blank values, and this is due to some patients who did not answer some questions because of their individual privacy [31]. There are two approaches to handling missing data in the cervical cancer dataset: ignore (remove) missing values or impute (fill) missing values.

2) Ignore missing values: Ignoring some features contributes to making data consistent due to the high percentage of missing values in them. This approach is useful because some features have missing values in the dataset, such as the "Time since first diagnosis" and "Time since last diagnosis", where their missing values were greater than 80 % of all data in these two attributes. Due to the difficulty in filling in such a large proportion of missing values with meaningful values and not finding any attribute dependencies that can be used to derive values for the missing data, these two attributes were excluded [11], [12], [16].

3) Impute missing values: Imputation methods is one of the common methods in the field of missing data that fill missing values with appropriate values [40]. There are many features in cervical cancer dataset with the missing values less than 20%. These missing values were recorded as "?" in the dataset and imputed in one of the following two ways:

- Imputation using the mean values: This is the most common of imputation techniques [41]. This method is conducted by calculating the mean value of non-missing data in a specific column and then the missing values are replaced with this value in that column.
- Imputation using decision tree: In 1982, Kalton and Kasprzyk were the first proposing the use of a decision tree to handle lost data [42]. In this method, the sklearn Iterative Imputer class was used with a decision tree regressor for numerical data and a decision tree classifier for categorical data. Instead of ignoring a feature that has missing values, the decision tree imputation was used to convert the lost value of that feature to some calculated value. Thus, the decision tree imputation predicts the imputation value based on other values in the dataset. Where the feature that contains missing values is used as a target, the remaining attributes are used as training data. After fitting the model, the missing values are identified as if they were class labels [43], [44]. The advantages of this method are it produces more accurate values and is available for both categorical and numeric variables; however, it is also more time-consuming [45].

4) Data transformation: In data transformation step, the data is converted or consolidated so that the processing results are more efficient, and it is easy to understand the existing patterns. Then, the data becomes suitable for processing and applying machine learning algorithms [23].

Normalization is one of the data transformation strategies, which refers to the process of scaling the values of features to be within a small specified range or common range, such as [0,1] or [-1,1]. There are many normalization methods, such as Min-Max, decimal and Z-score normalization. Min-max normalization was applied to this dataset.

5) Outliers: In a dataset, finding outliers is a challenging and complicated process, especially for high-dimensional datasets. The outlier refers to an observation or a subset of observations that appear to be inconsistent with the rest of the dataset, while the outlier detection refers to searching for objects in the dataset that are not subject to the laws that are valid for the main part of the remaining data [46].

In medical data, the leading causes of outliers are malfunctions of medical devices, human errors, patient-specific behaviors, natural change in the patient, medication intake, food or alcohol, stress, and others [46]. In some cases, the outlier value provides useful information because it may indicate a rare disease and, therefore, outlier values are usually treated by keeping, removing, or modifying them. In this paper, outlier values were preserved because these values explain the situation of people in society and their differences [47].

V. Implementation and Results

A. Implementation

As we mentioned previously in the research methodology section, three main steps were used. The first step was choosing and understanding the dataset. The second step was pre-processing the original data for classification and handling data imbalance. The last step involved feature selection and building a model based on useful prediction classification (see Fig. 4).

In the modelling stage, four predictive models were implemented for each target variable to compare their results and then to determine the best model based on its ability to detect cervical cancer. These predictive models were conducted using the Jupyter Notebook, which is an open-source environment that allows editing and running of Python 3.3 programming language. There are several Python libraries that have been used to build these models, such as Scikit-learn, matplotlib, NumPy and pandas.

After data pre-processing, the feature reduction technique, PCA with 11 principle components, was used to decrease the number of features and processing time. Then, the dataset was split into training and testing sets. Due to the unbalanced dataset, SMOTE technology was applied to the training set to achieve balance to the minority class highest accuracy and avoid classification mislead.

The voting classifier was applied, which is one popular ensemble method. The voting classifier combined the prediction outputs of three classifiers: logistic regression, random forest, and decision tree, as shown in Fig. 5, and extremely predicted classes were selected as class variables of test samples. Then, the stratified 10-fold cross-validation (CV) method was used to prevent the overfitting problem and for the validation and testing of data. Subsequently, the result of the

model was assessed using different evaluation metrics, such as accuracy, sensitivity, specificity, precision (PPA), NPA, f1-score and ROC_AUC.

For the four predictive models, the voting classifier was applied to them all to focus on the impact of SMOTE and PCA technologies on model performance and to compare them. The first model was built without applying any of these two technologies, the second model contained PCA only, the third one contained SMOTE only and both techniques were applied to the fourth model.

B. Evaluation and Results

The results of the four predictive models for each target variable are discussed in the following sections.

1) Target variable: Hinselmann: With the Hinselmann test, there were 823 benign and 35 malignant samples. The voting model before SMOTE achieved a total accuracy of 95.69 %, while after using SMOTE it achieved an accuracy of 96.62 %. The accuracy increased after using SMOTE by 0.93 %, the sensitivity ratio increased from 50 % to 96.97 % and the ROC_AUC metric rate increased from 64.17 % to 97.75 %, as shown in Table III.

SMOTE-voting-PCA works well with 11 principal components. In this case, the negative predictive accuracy, and the ROC_AUC scale already reached 99%. In comparison with the voting before SMOTE, SMOTE-voting-PCA increased the sensitivity and precision rates by 46.50 % and 30.99 % respectively. Also, the overall accuracy of SMOTE-voting-PCA was nearly 97%. Accordingly, SMOTE and PCA methods can basically actualize the action of the voting classifier.

Fig. 6 shows the superiority of the voting classifier performance compared to the three classifiers—logistic regression, random forest, and decision tree—in the SMOTE-Voting -PCA model of the Hinselmann target variable.

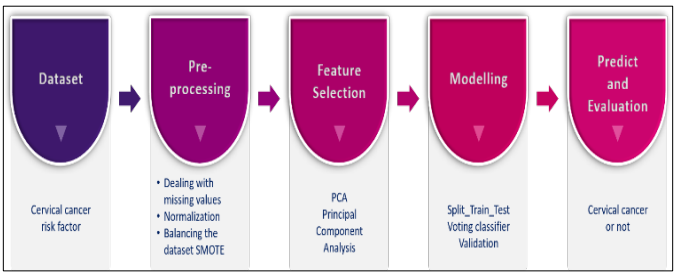

Fig. 4. Methodology Process.

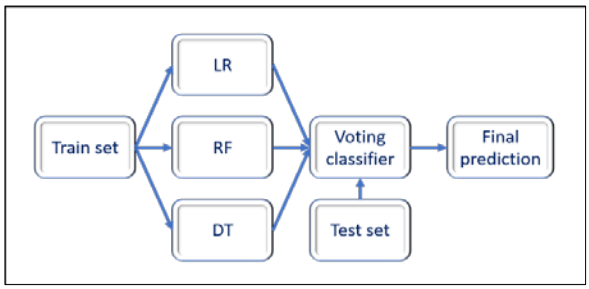

Fig. 5. Voting Classifier Workflow.

TABLE III. EVALUATION OF PREDICTIVE MODELS OF THE HINSELMANN TARGET VARIABLE

Evaluation Metrics %	Before SMOTE		After SMOTE	
	Voting	Voting-PCA	SMOTE-Voting	SMOTE-Voting -PCA
Accuracy	95.69	95.93	96.62	96.73
Sensitivity	50.00	50.00	96.97	96.50
Specificity	100	100	97.69	97.69
PPA	48.00	48.00	77.00	78.99
NPA	95.92	95.92	98.53	98.77
F1-score	49.00	49.00	96.39	96.85
ROC_AUC	64.17	57.04	97.75	98.56

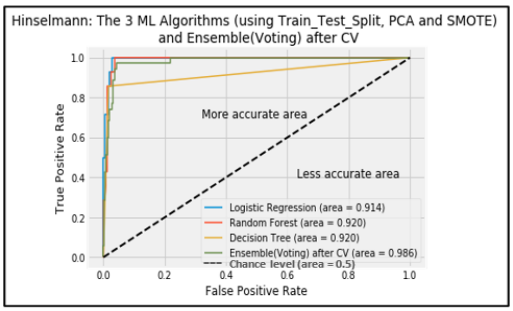

Fig. 6. The ROC Curve for the SMOTE-Voting-PCA Model of the Hinselmann Target Variable.

2) Target variable: Schiller: Concerning Schiller's test, the voting classifier before SMOTE achieved an overall accuracy of 90.09% with 74 patients and 784 non-patient samples. After SMOTE, SMOTE-Voting achieved an accuracy of 95.22 %, while the sensitivity increased by 39.87 %, PPA by 23.47 %, NPA by 5.46 % and ROC_AUC by 23.39 % in comparison to the voting model.

In voting-PCA with 11 principal components, the sensitivity, NPA and f1-score decreased by 2.00%, 0.42% and 4.00%, respectively, in comparison with the voting model. In contrast, the SMOTE-voting-PCA model for Schiller test with 11 components obtained the highest ratios in accuracy, sensitivity, PPA, NPA and f1-score, as shown in Table IV. Likewise, the ROC curve of SMOTE-voting-PCA in Fig. 7 shows that the model has the highest ROC_AUC in comparison to other models.

TABLE IV. EVALUATION OF PREDICTIVE MODELS OF THE SCHILLER TARGET VARIABLE

Evaluation Metrics %	Before SMOTE		After SMOTE	
	Voting	Voting-PCA	SMOTE-Voting	SMOTE-Voting-PCA
Accuracy	90.09	90.33	95.22	98.49
Sensitivity	55.00	53.00	94.87	98.60
Specificity	96.94	98.60	97.19	98.98
PPA	61.00	62.00	84.47	95.16
NPA	92.23	91.81	97.69	99.49
F1-score	57.00	53.00	95.11	98.37
ROC_AUC	68.17	61.78	91.56	99.80

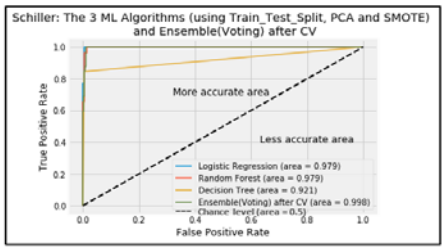

Fig. 7. The ROC Curve for the SMOTE-Voting-PCA Model of the Schiller Target Variable.

3) Target variable: Cytology: In the Cytology screening test, the voting model before SMOTE achieved a total accuracy of 94.29% with 44 malignant and 814 benign samples. This is considered to be a better result than the voting model after SMOTE, as the accuracy after SMOTE decreased to 91.72%, but the sensitivity and ROC_AUC rate increased to 91.26% and 72.30%, respectively, in the SMOTE-voting experiment.

In the SMOTE-voting-PCA model with 11 principal components, the ratio of four measures increased—NPA, f1-score, ROC_AUC and sensitivity—in comparison to the other models, reaching 97.84 %, 93.35 %, 93.90 % and 93.12 %, respectively, as shown in Table V. The remaining measures—accuracy, specificity, and PPA—decreased by 1.98 %, 5.16 % and 3.21%, respectively in comparison to the voting-PCA model.

It can be concluded that, for the Cytology test, the SMOTE-voting-PCA model obtained the highest ROC_AUC, sensitivity, PPA and NPA ratios in comparison to the rest of the models, as shown in Fig. 8.

4) Target variable: Biopsy: In a biopsy test, the accuracy of the voting model without SMOTE reached 93.24 % with 55 malignant and 803 benign samples. Table V shows that the performances of the voting and voting-PCA before SMOTE models were somewhat similar in most evaluation metrics.

After SMOTE, when comparing the models SMOTE-voting and SMOTE-voting-PCA, the accuracy, sensitivity, PPA, NPA, and ROC_AUC increased in the SMOTE-voting-PCA model by 2.22%, 1.99%, 6.6%, 0.89% and 4.64%, respectively. Thus, according to the evaluation results in Table VI and the ROC curves shown in Fig. 9, the SMOTE-voting-PCA model with 11 principal components was able to predict cervical cancer via a Biopsy test better than other models. This clarifies the role of the two technologies in raising the performance of the model, whether with a biopsy test or with the previous three tests—Hinselmann, Schiller and Cytology.

C. Discussion and Comparison

From the previous results, the voting method helped increase the performance of the models in comparison to other classifiers and to obtain a good accuracy in the classification of cervical cancer data. However, the somewhat high accuracy rate during the classification was offset by a low sensitivity rate, ranging between 50% and 53%, in all previous experiments of the four target tests. Where many patients were classified as non-patients that is incorrect and medically unacceptable classification. This defect is due to the limited and unbalanced dataset. Hence, the SMOTE algorithm was used to solve this problem and create new samples synthetically, thus increasing the data of cervical cancer patients. After using SMOTE technology in the SMOTE-voting model, accuracy, sensitivity and PPA ratios improved by 0.93% to 5.13%, 39.26% to 46.97% and 2% to 29%, respectively, for all target variables. PCA technology was also used to reduce the features to 11 principal components, thereby reducing computational processing time and increasing model efficiency. Experiments showed that SMOTE and PCA technologies have greatly helped classify cervical cancer data correctly for all target variables.

The top 10 relevant risk factors of cervical cancer of these target tests are the features 0 and 2 that indicate age and first sexual intercourse according to Table II, and these features appear in the first three ranks for all target tests, while feature 8, which indicates hormonal contraceptives, appears in three of the four tests. Fig. 10 shows the top 10 relevant risk factors for the Biopsy target variable.

When comparing our results to the results of Wu and Zhou [11] shown in Fig. 11, we found that the SMOTE-Voting-PCA model outperforms the accuracy and specificity of the SVM-PCA model in the Hinselmann and Cytology tests. In contrast, our models of the Schiller target variable obtained better results in four measures: accuracy, specificity, PPA and NPA. With the Biopsy test, the accuracy, specificity, and PPA of our model increased after SMOTE in comparison to the non-SMOTE models proposed by Wu and Zhou [11].

TABLE V. EVALUATION OF PREDICTIVE MODELS OF THE CYTOLOGY TARGET VARIABLE

Evaluation Metrics %	Before SMOTE		After SMOTE	
	Voting	Voting-PCA	SMOTE-Voting	SMOTE-Voting-PCA
Accuracy	94.29	94.87	91.72	92.89
Sensitivity	52.00	52.00	91.26	93.12
Specificity	99.14	99.75	95.21	94.59
PPA	59.00	73.00	61.00	69.79
NPA	95.05	95.08	96.27	97.84
F1-score	52.00	53.00	91.14	93.35
ROC_AUC	60.48	48.02	72.30	93.90

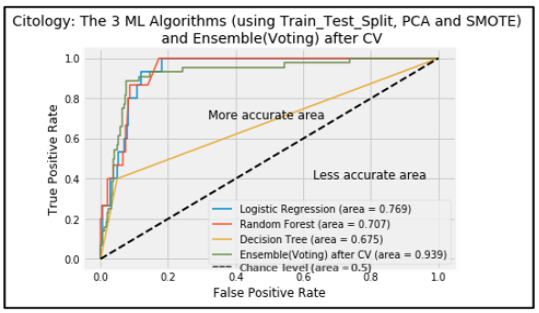

Fig. 8. The ROC Curve for the SMOTE-Voting-PCA Model of the Cytology Target Variable.

TABLE VI. EVALUATION OF PREDICTIVE MODELS OF THE BIOPSY TARGET VARIABLE

Evaluation Metrics %	Before SMOTE		After SMOTE	
	Voting	Voting-PCA	SMOTE-Voting	SMOTE-Voting-PCA
Accuracy	93.24	93.36	95.22	97.44
Sensitivity	51.00	52.00	95.80	97.79
Specificity	99.25	99.50	96.64	98.01
PPA	59.00	64.00	83.01	89.61
NPA	93.76	93.78	98.10	98.99
F1-score	51.00	52.00	95.22	97.44
ROC_AUC	65.55	52.47	94.86	99.50

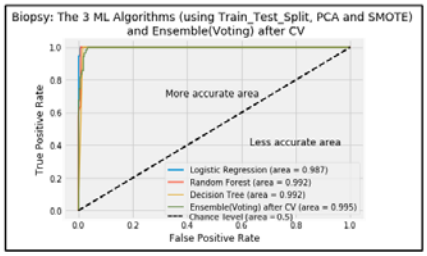

Fig. 9. The ROC Curve for the SMOTE-Voting-PCA Model of the Biopsy Target Variable.

Fig. 10. The Importance of Features for the Biopsy Target Variable.

Fig. 11. The Comparison of the Results of SMOTE And non-SMOTE Models for the Four Target Variables.

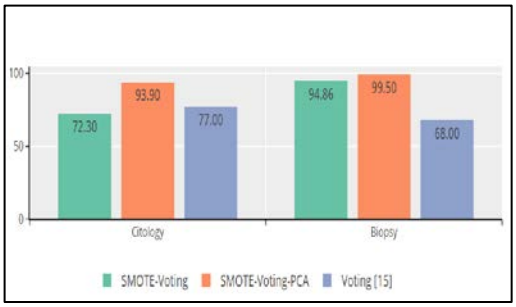

Fig. 12. Comparison of the ROC_AUC Measure for the Cytology and Biopsy Target Variables.

In Fig. 12, ROC_AUC was compared between the models we propose (SMOTE-Voting and SMOTE-Voting-PCA) and the voting model proposed by Rayavarapu et al. [15]. This comparison confirms the roles of SMOTE and PCA techniques in raising model performance.

VI. CONCLUSIONS

An early detection procedure provides the best opportunity for diagnosing cervical cancer at an early stage of the disease when the treatment is more beneficial. Cervical cancer, if detected early, is one of the most successfully treatable types of cancer. The paper is focused on finding a model capable of diagnosing cervical cancer with high accuracy and sensitivity using machine learning algorithms, as well as on trying to find a method for dealing with an unbalanced dataset, where the imbalance problem reduces predictive efficiency and increases misleading classification. In this paper, we combined the best three classifications of machine learning algorithms to predict cervical cancer and obtain the highest results using one of the ensemble approaches, which is the voting method. Four predictive models were created using the UCI cervical cancer risk factors dataset for each of the targeted variables: Hinselmann, Schiller, Cytology and Biopsy. The proposed models introduce new built-in classifications, which collect certain techniques, such as the SMOTE to increase the number of minority cases to rebalance the dataset and the PCA technique to reduce the dimensions that do not affect the accuracy of the model. From the results obtained, the voting method with SMOTE and PCA technologies helped classify cervical cancer data correctly for all target variables and raise the accuracy, sensitivity, and ROC_AUC of predictive models to high rates as in the Schiller target variable, they reached to 98.49%, 98.60%, and 99.80%, respectively.

VII. FUTURE WORK

In our future work, the dataset used to detect cervical cancer can be improved and the efficiency of future prediction models can be increased by adding several essential attributes that assist early detection of cervical cancer. Some information could be collected and added to the dataset. For example, whether the Pap smear was performed recently and whether an HPV vaccination was given. This additional information can be collected from patients and clinics so that the extensive dataset will assist in building a better predictive model. Moreover, adding several essential attributes will also improve the prediction model for early detection of cervical cancer.

REFERENCES

[1] World Health Organization, "Cancer. " [Online].Available: https://www.who.int/health-topics/cancer#tab=tab_3. [Accessed: 24-Jan-2020].

[2] A. Alrawaji, W. Alzahrani, Z. Alshahrani, F. Alomran, and A. Almadouj, "Cancer Incidence Report Saudi Arabia 2015," 2015.

[3] C. Al-Eid, Haya S. , BDS, DFE, "Cancer Incidence Report Saudi Arabia 2010," 2010.

[4] The Global Cancer Observatory - World Health Organization, "Saudi Arabia Source: Globocan 2018," 2019.

[5] I. C. Scarinci et al., "Cervical cancer prevention: New tools and old barriers," Cancer, vol. 116, no. 11. pp. 2531–2542, 01-Jun-2010.

[6] Centers for Disease Control and Prevention, "Basic Information About Cervical Cancer | CDC," 07-Aug-2019. [Online]. Available: https://www.cdc.gov/cancer/cervical/basic_info/index.htm. [Accessed: 29-Jan-2020].

[7] M. Schiffman, P. E. Castle, J. Jeronimo, A. C. Rodriguez, and S. Wacholder, "Human papillomavirus and cervical cancer," Lancet, vol. 370, no. 9590. pp. 890–907, 08-Sep-2007.

[8] M. Sharma, "Cervical cancer prognosis using genetic algorithm and adaptive boosting approach," Health Technol. (Berl)., vol. 9, no. 5, pp. 877–886, Nov. 2019.

[9] W. J. Koh et al., "Cervical cancer, version 2.2015," JNCCN Journal of the National Comprehensive Cancer Network, vol. 13, no. 4. Harborside Press, pp. 395–404, 01-Apr-2015.

[10] S. F. Abdoh, M. Abo Rizka, and F. A. Maghraby, "Cervical cancer diagnosis using random forest classifier with SMOTE and feature reduction techniques," IEEE Access, vol. 6, pp. 59475–59485, 2018.

[11] W. Wu and H. Zhou, "Data-driven diagnosis of cervical cancer with support vector machine-based approaches," IEEE Access, vol. 5, pp. 25189–25195, Oct. 2017.

[12] B. Nithya and V. Ilango, "Evaluation of machine learning based optimized feature selection approaches and classification methods for cervical cancer prediction," SN Appl. Sci., vol. 1, no. 6, Jun. 2019.

[13] H. Teame et al., "Factors associated with cervical precancerous lesions among women screened for cervical cancer in Addis Ababa, Ethiopia: A case control study," PLoS ONE 13(1) e0191506, 2018.

[14] X. Deng, Y. Luo, and C. Wang, "Analysis of Risk Factors for Cervical Cancer Based on Machine Learning Methods," Proceedings of 2018 5th IEEE International Conference on Cloud Computing and Intelligence Systems, CCIS 2018. pp. 631–635, 2019.

[15] K. Rayavarapu and K. K. V. Krishna, "Prediction of Cervical Cancer using Voting and DNN Classifiers," Proceedings of the 2018 International Conference on Current Trends towards Converging Technologies, ICCTCT 2018. 2018.

[16] Abdullah, F. Bin Ashraf, and N. S. Momo, "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),2019, pp.1-6.

[17] K. Fernandes, D. Chicco, J. S. Cardoso, and J. Fernandes, "Supervised deep learning embeddings for the prediction of cervical cancer diagnosis," PeerJ Computer Science, vol. 2018, no. 5. 2018.

[18] J. Lu, E. Song, A. Ghoneim, and M. Alrashoud, "Machine learning for assisting cervical cancer diagnosis: An ensemble approach," Futur. Gener. Comput. Syst., vol. 106, pp. 199–205, May 2020.

[19] A. Ghoneim, G. Muhammad, and M. S. Hossain, "Cervical cancer classification using convolutional neural networks and extreme learning machines," Futur. Gener. Comput. Syst., 2020.

[20] E. Y. Boateng and D. A. Abaye, "A Review of the Logistic Regression Model with Emphasis on Medical Research," J. Data Anal. Inf. Process., vol. 07, no. 04, pp. 190–207, Sep. 2019.

[21] S. Mishra, "Handling Imbalanced Data : SMOTE vs . Random Undersampling" Int. Res. J. Eng. Technol., vol. 4, no. 8, pp. 317–320, 2017.

[22] P. Yang, Y. Hwa Yang, B. B. Zhou, and A. Y. Zomaya, "A Review of Ensemble Methods in Bioinformatics," Curr. Bioinform., vol. 5, no. 4, pp. 296–308, 2016.

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

[24] U. K. Kumar, M. B. S. Nikhil, and K. Sumangali, "Prediction of breast cancer using voting classifier technique," in 2017 IEEE International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials, ICSTM 2017 - Proceedings, 2017, pp. 108–114.

[25] A. Yuniarti and M. A. Fauzi, "Ensemble method for indonesian twitter hate speech detection," Indones. J. Electr. Eng. Comput. Sci., vol. 11, no. 1, pp. 294–299, 2018.

[26] N. T. M. Sagala, "A Comparative Study of Data Mining Methods to Diagnose Cervical Cancer," J. Phys. Conf. Ser., vol. 1255, no. 1, 2019.

[27] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques Third Edition. Morgan Kaufmann, Elsevier, 2011.

[28] K. Hajian-Tilaki, "Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation," Casp. J. Intern. Med., vol. 4, no. 2, pp. 627–635, 2013.

[29] D. Wang and M. Reynolds, "AI 2011: Advances in Artificial Intelligence | SpringerLink," 2011.

[30] T. Fawcett, "ROC Graphs: Notes and Practical Considerations for Researchers," 2004.

[31] "UCI Machine Learning Repository: Cervical cancer (Risk Factors) Data Set." [Online]. Available: http://archive.ics.uci.edu/ml/datasets/Cervical +cancer+%28Risk+Factors%29. [Accessed: 18-Oct-2019].

[32] K. Fernandes, J. S. Cardoso, and J. Fernandes, "Transfer learning with partial observability applied to cervical cancer screening," in Iberian Conference on Pattern Recognition and Image Analysis. Springer International Publishing, 2017, vol. 10255 LNCS, pp. 243–250.

[33] K. Akyol, "A Study on Test Variable Selection and Balanced Data for Cervical Cancer Disease," Inf. Eng. Electron. Bus., vol. 5, pp. 1–7, 2018.

[34] T. M. Alam, M. Milhan, A. Khan, M. A. Iqbal, A. Wahab, and M. Mushtaq, "Cervical Cancer Prediction through Different Screening Methods using Data Mining, " IJACSA) Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 2, 2019.

[35] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," Computational and Structural Biotechnology Journal, vol. 13. Elsevier, pp. 8–17, 2015.

[36] H. Kang, "The prevention and handling of the missing data," Korean J. Anesthesiol., vol. 64, no. 5, pp. 402–406, May 2013.

[37] P. Royston, "Multiple Imputation of Missing Values,"Stata J. Promot. Commun. Stat. Stata, vol. 4, no. 3, pp. 227–241, Aug. 2004.

[38] V. Audigier, F. Husson, and J. Josse, "A principal component method to impute missing values for mixed data," Adv. Data Anal. Classif., vol. 10, no. 1, pp. 5–26, Mar. 2016.

[39] R. Garreta and G. Moncecchi, Learning scikit-learn : Machine Learning in Python. Packt Publishing, 2013.

[40] C. T. Tran, M. Zhang, and P. Andreae, "A genetic programming-based imputation method for classification with missing data," in Lecture Notes in Computer Science, Springer, Cham, 2016, vol. 9594, pp. 149–163.

[41] P. J. García-Laencina, J. L. Sancho-Gómez, and A. R. Figueiras-Vidal, "Classifying patterns with missing values using Multi-Task Learning perceptrons," Expert Syst. Appl., vol. 40, no. 4, pp. 1333–1341, Mar. 2013.

[42] T. Rockel, D. W. Joenssen, and U. Bankhofer, "Decision Trees for the Imputation of Categorical Data, " kit Sci. Publ. , vol. 2, no. 1, 2017.

[43] H. Barlow, S. Mao, and M. Khushi, "Predicting High-Risk Prostate Cancer Using Machine Learning Methods, " Data, vol. 4, no. 3, p. 129, Sep. 2019.

[44] M. S. Santos, J. P. Soares, P. H. Abreu, H. Araújo, and J. Santos, "Influence of data distribution in missing data imputation," Lect. Notes Artif. Intell. Med. AIME 2017. Lect. Notes Comput. Sci. vol 10259. Springer, Cham, vol. 10259 LNAI, pp. 285–294, 2017.

[45] G. Svolba and SAS Institute., Data quality for analytics using SAS. SAS Institute, 2012.

[46] C. M. Salgado, C. Azevedo, H. Proença, and S. M. Vieira, "Noise versus outliers," in Secondary Analysis of Electronic Health Records, Springer International Publishing, 2016, pp. 163–183.

[47] M. S. Koti, "Chapter-3 Outlier Mining in Medical Databases," in Automation of Data Mining (DM) in Hospital Information System (HIS) for Quality Improvement, Coimbatore, Bharathiar University, 2014.