RESEARCH ARTICLE

Editorial Process: Submission:01/18/2023 Acceptance:04/23/2023

Prediction and Detection of Cervical Malignancy Using Machine Learning Models

Seeta Devi[1]*, Sachin Ramnath Gaikwad[2], Harikrishnan R[2]

Abstract

Objective: Human papillomavirus and other predicting factors are responsible causing cervical cancer, and early prediction and diagnosis is the solution for preventing this condition. The objective is to find out and analyze the predictors of cervical cancer and to study the issues of unbalanced datasets using various Machine Learning (ML) algorithm-based models. **Methods:** A multi-stage sampling strategy was used to recruit 501 samples for the study. The educational intervention was the video-assisted counseling which is consisted of two educational methods: a documentary film and face-to- face interaction with women followed by reminders. Following the collection of baseline data from these subjects, they were encouraged to undergo Pap smear screening. Women having abnormal Pap tests were sent for biopsy. Machine learning classification methods such as Decision Tree (DT), Random Forest (RF), Logistic Regression (LR), Multi-layer Perceptron (MLP) and Naive Bayes(NB) were used to evaluate the unbalanced input and target datasets. **Result:** Merely 398 women out of 501 showed an interest to participate in the study, but only 298 stated a willingness for cervical screening. Atypical malignant cells were discovered on the cervix of 26 women who had abnormal pap tests. These women had guided for further tests, such as a cervical biopsy, and seven women had been diagnosed with cervical cancer. LR in models 1, 2, and 4 showed 88% to 94% sensitivity with 84% to 89% accuracy, respectively for cervical cancer prediction, whereas DT in models 3, 5, and 6 algorithms exhibited 83% to 84% sensitivity with 84% to 88% accuracy, respectively. The NB and LR algorithms produced the highest area under the ROC curve for testing dataset, but all models performed similarly for training data. **Conclusion:** In current study , Logistic Regression and Decision Tree algorithms were identified as the best-performed ML algorithm classifiers to detect the significant predictors.

Keywords: Prediction and detection- Cervical malignancy- Machine learning algorithms- Prediction- Cervical screening

Asian Pac J Cancer Prev, **24 (4)**, 1419-1433

Introduction

Deep Learning (DL) and Machine Learning (ML) are effective algorithm classifiers in predicting brain tumours, breast cancer, thermal sensation, dementia evaluation, COVID-19, renal disorders, heart problems, and cervical cancer (Abbas et al., 2021; Ayoub et al., 2021; Khamparia et al., 2021). Because of technological improvements in the health care system, several medical disorders can now be predicted at an earlier stage based on identifying critical factors than traditional diagnostic approaches (Chen H et al., 2021; Javed et al., 2021; Javed et al., 2020; Sarwar et al., 2019). According to Global Statistics (2020), around 604,127 cervical malignant cases were detected (Parkin et al., 2005). Whereas, Cancer.Net Editorial Board survey stated that across the globe, about 341,831 women died from cervical cancer in 2020. Cervical cancer is primarily found in developing countries, accounting for 83% of deaths (Salmi et al., 2019). Cervical malignant growths are the 4th utmost leading cause of death in women worldwide. It is one of the most serious malignancies that endangers women's health, and initial symptoms are difficult to detect until the disease has progressed to stage II (Aminisani et al., 2012). The diseased cells damage the cervical cells, and these cells migrate to other organs, including the lungs, heart, liver, kidneys etc. (Do et al., 2001). The most common cause of the rising prevalence of cervical cancer among women, particularly in developing countries, is a lack of awareness about screening measures that aid in early diagnosis (MacCosham et al., 2020).

The burden of cervical cancer is being reduced in many countries, particularly in developed countries, by implementing predictive, detective, preventive, and systematic treatment techniques. Cervical cancer deaths decreased from 4% per year in 1996-2003 to 1% per year from 2010 to 2019. (Cancer.Net Editorial Board). Screening tests enable clinicians to treat precancerous lesions at an early stage, preventing them

*[1]Symbiosis College of Nursing (SCON), Symbiosis International Deemed University (SIDU), Pune- 412115, India. [2]Symbiosis Institute of Technology (SIT), Symbiosis International Deemed University (SIDU), Pune- 412115, India. *For Correspondence: drseetadevi1981@gmail.com*

from progressing to malignant tumours. Although, the death rate in developing nations continues to rise due to a lack of resources, insufficient preventive methods, a lack of freely available Human papillomavirus (HPV) vaccine programmes, and a lack of awareness initiatives.

HPV contamination is one of the responsible factors for the development of cervical malignancy. HPV is primarily transmitted through sexual contact. It's acquisition has become increasingly associated with unusual sexual behaviours such as early sexual exposure, multiple sexual partners, and so on. Precancerous lesions take approximately 5 to 10 years to develop into malignant cells (MacCosham et al., 2020; Schiffman et al., 2007). Thereby providing an accessible time for women to go for cervical screening at least once in every three years with a Pap smear or visual inspection with acetic acid (VIA) or HPV DNA test, which assists them in diagnosing cervical cancer at an early stage. Cervical cancer is a preventable disease condition because it can be detected early by employing predictive and screening models. However, in developing nations, only 5% of women participate in cervical screening (Aminisani et al., 2012). Furthermore, cytological factors in Pap smear test are considered as diagnostic- predictive aspects because they identify the structure of the gland cell, squamous epithelial tissue, metaplastic cells, aberrant polymorphic cells, and dysplasia cells, as well as the existence of blood, bacteria, and fungus in the client's samples (Do et al., 2001).

Numerous susceptible predictors of cervical carcinoma have been identified, include smoking by the individual or their partner, inadequate nutrition, immunosuppression, use of immunomodulatory drugs, prolonged contraception utilization, racial groups, deficiency of vitamin A, C and folate, having multiple sex partners, subsequent pregnancies, childbearing at a young age, sexually-transmitted genital tract infections, low socioeconomic status, and illiteracy (Latha et al., 2014; Mandelblatt et al., 1991; Randall et al., 2016; Workowski et al., 2015). Daily, health sector creates huge volumes of data that can be utilized to forecast future sickness based on a patient's treatment history and health information.

Furthermore, by incorporating essential healthcare data, these areas can be enhanced. Non-invasive classification technologies, such as supervised machine learning (ML), are critical for cervical carcinoma prediction (Ramondetta, 2013; Workowski et al., 2015). ML in health care allows researchers to process massive amounts of complex medical data and evaluate it for curative ideas. Physicians use this material to extend medical treatment to patients. Patient satisfaction may enhance due to ML in healthcare coverage. Among the methods utilized are the Decision Tree (DT), Random Forest (RF), Logistic Regression (LR), Neural Network (NN), Naive Bayes, Artificial Neural Network Decision tree, and Support Vector Machines (SVM) (Mukhopadhyay et al., 2016; Nematollahi et al., 2017; Rezaianzadeh et al., 2019).

The aim of this study was to deploy the various ML algorithm based models to understand the issues in uneven datasets.

Following were the objectives of this study

1. To find out and analyze the predictors of cervical cancer using machine learning classifiers.
2. To study the issues of unbalanced datasets using various ML algorithm-based models.
3. To conduct the survey in view to find out the concerns of women regarding cervical cancer screening and that provides an accurate message to the readers.

This paper includes the various sections namely, Related work, Research methodology, Results, Discussion and Conclusion.

Related work

As per the literature section criteria (LSC), the authors examined the relevant research papers from various databases. The current study investigated several electronic resources, including SCOPUS, PubMed, Institute of Electrical and Electronics Engineers (IEEE) Xplore, and Springer. The article listed below provides existing studies on the current study.

Literature Selection Criteria

It is one of the most critical aspects of the literature selection criterion since it enables researchers to arrange their work systematically, especially while downloading articles. Regarding search criteria, the authors ensured that the chosen publication was at least published in SCOPUS indexed journals.

Inclusion criteria followed in the review of the current study

- The purpose of the study was included in the researcher article.
- The duration of this survey was established between 2015 and Nov 2022, and it attempted to comprehend the insights of earlier studies.

Recent studies focused on several methodologies based on standard machine learning approaches, such as k-nearest neighbors (KNN), K-means clustering and RF, for early detection of cervical malignancy. A study was conducted to examine the effectiveness of several means in artificial neural networks to detect malignancy, and non-malignant cells (Singh et al., 2020). The researchers aimed to demonstrate a way of screening for cervical cancer utilizing cervigram pictures and the directed local histogram methodology (OLHT) (Asadi et al., 2020). Nithya et al., 2019 attempted to determine the level of cervical infection using the UCI data repository and six categorization models. In this study, physicians verified pre-processing data to confirm some risk factors and perform validation. The models' performance was further evaluated using 10-fold cross-validation . A study had examined various measurements, including accuracy, sensitivity, specificity, and area under the curve (AUC). In that sequence, the QUEST values were 95.55, 90.48, 100, and 9.20, respectively. The authors employed an integrated learning technique to diagnose machinery faults. Each model training was applied at an indigenous level to improve learning performance (Lu et al., 2020).

For the prediction of cervical cancer (Jahan et al., 2021), authors applied various types of ML classification algorithms. This study aimed to identify the topmost features that can cause cervical cancer utilizing eight well-known classification algorithm methods, including Multilayer Perceptron, Random Forest and k-Nearest Neighbor, Decision Tree, Logistic Regression, SVC, Gradient Boosting, and AdaBoost. The measures used to assess the performance of those classifications were accuracy, recall, precision, and f1-score. The MLP classification method worked admirably in detecting a wide range of relevant features in the datasets (Jahan et al., 2021). Authors (Jaswinder Singh et al., 2019) attempted to provide a model for cervical cancer prediction utilizing the UCI data repository and ML classification models. The data was pre-processed, and then the repository data was updated by extraction and validation. This study model included ten data elements connected with four stages. The pre-processed data were made available to the physician for verification before training the ML classifiers. Six classifiers were utilized in this research, with the decision tree classifier confirming the suitable stage prediction in terms of false-positive rate, f1-measure, and precision (Jaswinder Singh et al., 2019).

A study had described the numerous LM classifiers for the early prediction of cervical cancer, including multi-layer perceptron, decision trees, random forest, K-Nearest Neighbor, and Naive-Bayes. The authors of this study examined the performance of various ensemble methods (AdaBoost, Stochastic Gradient Boosting, Random Forests, and Extra Trees) and ML classifiers (SVM and K-Nearest Neighbor) for predicting cervical cancer based on risk factors. This study's measurement metrics are F1 score, Area Under Curve, and Recall. The extra trees classifier performed the best, with 96% accuracy (Ahishakiye et al., 2021). Al Mudawi et al.,2022 presented a report in which they used ML classifiers to predict cervical cancer. The study is divided into four phases: dataset, data pre-processing, predictive model selection, and pseudo-code assignment. This work utilized algorithm classifiers such as decision tree, logistic regression, K-nearest neighbor's algorithm, adaptive boosting, gradient boosting, random forest, and XGBoost. Adaptive boosting, Random forest (RF), decision tree, and gradient boosting algorithms performed best with the highest accuracy score of 100%, while SVM also performed with 99% accuracy.

Studies attempted to employ five machine learning algorithms in those investigations: random forest, KNN, C5.0, SVM, and RPart, they achieved the highest accuracy rates of 97, 96.9, 96, 88, and 88, respectively. ML approaches such as decision tree, random forest, and logistic regression have been utilized in combination with the voting method (Alyafeai et al., 2020; Mukama et al., 2017) and assimilated carcinomas of the cervix prediction models and a cervical screening pipeline based on cervical imaging. A deep neural network-learning model was used to automate the detection and diagnosis of cervical malignant tumours. This finds the union intersection (IoU) 1,000 times faster than state-of-the-art data-driven simulation, with a detecting accuracy of 0.68. By automating the diagnosis of cervical cancer from Pap-drug pictures, (William et al., 2019) used a model to reduce the likelihood of errors. For image improvement, a local adaptive histogram was applied. Authors (Unlersen et al., 2017) collected data from 858 patients with 33 variables employed in predictive analysis to detect cervical cancer. A unique machine learning techniques such as Multilayer Perceptron, BayesNet, and k-Nearest Neighbor were used for accurate prediction. The performance of this algorithm is measured using a confusion matrix and the proportion of correctly recognized occurrences. The performance measurements of confusion matrix are precision call, recall score, F1 score and accuracy. The confusion matrix and cost-effectiveness in terms of CPU time are used to analyze the performance of numerous approaches. This proposed approach obtains the best feature while decreasing process time to aid clinicians in the early detection of cervix carcinoma. Logistic Regression (LR) yields 100% accuracy but requires additional CPU time. Nonetheless, 99% accuracy is obtained in exchange for reduced CPU time (Singh et al., 2020).

This paper deals with unbalanced datasets using a combination of attribute selection methods and evaluates the performance of ML algorithms-based models. In contrast, previous studies focused on splitting the balanced dataset, and some studies focused on finding the best ML algorithms for predicting cervical cancer. The authors of this paper also employed the Gini index analysis to determine the percentage of predictors that appeared in the target categories.

Materials and Methods

A prospective study was conducted, which consists of four phases, as listed below

1. Researchers reviewed various databases, recent research studies, and other library resources to identify risk factors and critical variables likely to lead to cervical cancer. Machine learning techniques were also found as important for predicting cervical cancer.

2. Researchers developed a questionnaire based on a literature review and expert opinion, focusing on risk factors for cervical cancer. Cronbach's alpha is used to calculate and confirm the reliability, and its value was 0.92, indicating that the tool's reliability was determined to be effective.

3. Researchers sought permission from various Pimpri Chinchwad Municipal Corporation (PCMC) hospitals to collect data and obtain ethics approval from Symbiosis Independent Ethics Committee, Pune. The hospital and volunteers were recruited using a multistage sampling strategy from various PCMC facilities. A total of 501 women participated in the study; after collecting baseline data from the women, they were motivated to undergo the Pap smear screening. Of 501 women, 298 expressed interest in the study and participated in cervical screening. Atypical malignant cells were found on the cervix among 26 women with abnormal pap tests. These women had guided further investigations, such as a cervical biopsy, of those 7 women were diagnosed with cervical cancer.

4. Data Preprocessing: The data was normalized using the Continuize discrete variables technique with a single feature value. Due to class imbalance in the dataset, we

adopted the methodology of dividing the data into six models. Initially, we had 501 samples; out of this sample, 298 samples had undergone screening. Seven samples were diagnosed with cervical cancer, while others had negative screening results. Due to the class imbalance, this data cannot be used for accurate prediction as per medical considerations. This imbalanced class may provide wrong prediction values. To overcome this issue, we have reduced the class imbalance by dividing the dataset into six models and keeping the less-numbered samples. Samples in these models were selected randomly from 298 participants, the five models consisted of 50 samples, while 6th model entailed 45 samples, and each of these models mechanized with seven positive samples. Figure 1 indicates a similar workflow for all six models. The prediction model was evaluated using 5 folds of cross validation with stratification method and the target class was chosen as the average over classes. 5 algorithms' hyperparameters are fine-tuned.

Challenges of data collection

The collection of data from women for cervical screening (Pap test) became challenging due to several reasons.

Some of the common challenges are included below

Lack of awareness

Many women were not aware of the importance of regular cervical screening and were not ready for testing. This resulted in a low participation percentage; despite the fact that the authors recruited 501 subjects for the study, only 298 had undertaken screening.

Fear and embarrassment

Many women felt uncomfortable or embarrassed about undergoing a Pap test.

Accessibility and cost

In all health centres, cervical screening services were not readily accessible, or the cost of screening was too high. Thus, the authors had decided to collect the data only from YCM hospital where the testing charges were less and for some women charges are waved off.

Lack of trust

Some women had the problems of trusting healthcare providers, the healthcare system, and results of test.

To overcome these challenges, we had adopted some strategies such as education and awareness campaigns, outreach programs, and the provision of language support were implemented to encourage women to participate in cervical screening programs.

Machine learning algorithms for prediction of medical dieases

This section explains the methods of machine learning techniques used to predict cervical cancer in the current study.

Decision Tree (DT) algorithm

DT trails the rules of divide and conquer. In DT algorithm, the features will take up the different values called as classification trees. To resolve classification and regression issues, the classification and regression tree to be used. DT purports to have a lot of tree branches, which is why it has the Tree in its name. The DT starts with the root nodes, just as a tree originates through its roots. The leaves represent special categories, while the branches represent the mix of characteristics that result in the categorical variables. DT can also accept the continuous variables known as the regression trees. The commonly used DT algorithms in medical field are C4.5 and EC4.5 (Lilhore et al., 2022; Tiwari et al., 2018).

Random Forest (RF)

Utilizing various learners, ensemble methods improve performance of the models. RF is a form of ensemble intelligence as well. The RF tagging approach decreases the possibility of aberrations influencing findings. This is effective for both categorical and continuous data. In this, scaling of the datasets is not required. For the more learners, the higher computational resources are needed

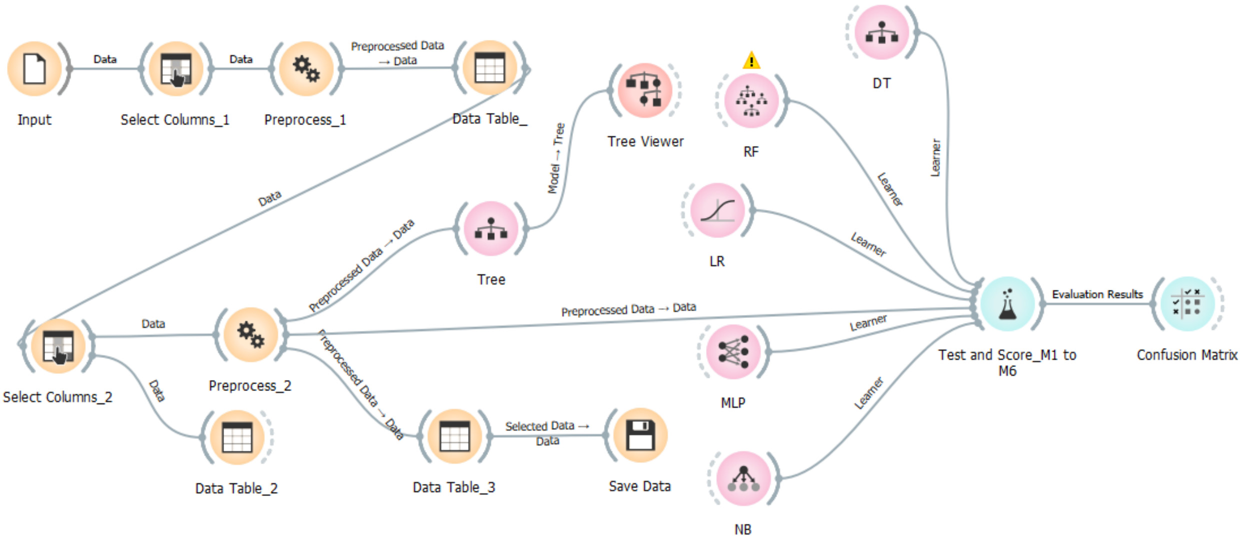

Figure 1. Workflow Diagram for Classification Prediction Models

for complicated methods. The decision in this method is determined by polling. This type of method is known as ensemble learning. Random forests are composed of an array of trees and plants. Random forests have numerous decision trees, similar to the number of trees in the forest. The decision made by the majority of trees is regarded as the right conclusion (Al Mudawi et al., 2022; Kaur et al., 2018).

Logistic Regression

Logistic regression (LR) is a machine learning (ML) method used to tackle class imbalance problem. The LR model is built on a conditional framework, with data values that ranges from zero to one. Email spam detection, fraudulent financial transactions identification, and malignant tumour diagnosis are all instances of LR-based ML. LR employs the cost function, sometimes known as the sigmoid function. Any actual number from zero and one is transformed by the sigmoid function (Al Mudawi et al., 2022; Wright, 1995).

Naïve Bayes (NB)

The NB model is a classification algorithm based on Bayesian concepts. It anticipates membership likelihood for every category according to a specific record or data point. The most likely category is the one with the highest likelihood. Rather than just making predictions, the NB classification projects probabilities (Ahsan et al., 2022).

The existence of a feature in a category is assumed to be unconnected with any other characteristic in that category, and if they're connected, then remain independent of each other's occurrence. Because each characteristic contributes to the likelihood individually. An advantage of this approach is that it can be used on both binary and multi-dimensional data. Also, unlike other machine learning methods, we need fewer training datasets (Khalil et al., 2019).

Table 1. Important Variable Collected from the Various Library Resources

Raw	Factors	Type of the data	Role
1	Age	Categorical	Input
2	Marital status	Categorical	Input
3	Ag a marriage	Categorical	Input
4	Age at onset of the sex	Categorical	Input
5	Age at the first child	Categorical	Input
6	Life time sexual partners	Categorical	Input
7	Life time pregnancies	Categorical	Input
8	Screened for cervical cancer	Categorical	Input
9	Number of times screened	Categorical	Input
10	Reasons for not screening	Categorical	Input
11	Smoking consumption	Categorical	Input
12	Duration of smoking	Numerical	Input
13	Temporary family planning methods	Categorical	Input
14	Type of the family planning methods	Categorical	Input
15	Family history cervical cancer	Categorical	Input
16	Infection of the cervix	Categorical	Input
17	Screening results	Categorical	Target

Multi-layer Perceptron (MLP)

A Multi-layer Perceptron is a type of neural network that consists of multiple layers of interconnected nodes or neurons. The MLP receives input data, processes it through multiple hidden layers, and produces an output. MLP is a supervised learning algorithm that is used for classification and regression tasks (Mohd Fakharuddin Zorkafli et al., 2019) .

The prediction models are created using the input dataset and fine-tuning each algorithm's hyper-parameter. After clicking on each algorithm, we could fine-tune the hyper-parameters to maintain better prediction accuracy. To develop prediction models, a 5-fold cross-validation model was adopted.

Process of descriptive statistics

One of the data mining methods involving mathematics and related data collection and elucidation is descriptive statistics for predicting a high cervical cancer data set (Ramnath et al., 2021). The JASP is a simple tool to use for researchers with limited computer skills. Figure 2 depicts the four sections of descriptive statistics in this tool. The below image depicts descriptive statistics such as central tendency (median), and dispersion.

Table 1 presents the major risk factors used for cervical cancer prediction based on data from recent research studies and library resources. Table 1 shows that the majority of the data is categorical data, which is organised by a set of categories rather than evaluated on a continuous numerical scale. Just smoking duration falls within the category of numerical data. All components' roles were treated as input variables, while screening results were considered as target variables for data processing.

Results

Thus seventeen significant variables were used in

Table 2. Distribution of the Samples based on Their Risk Factors

Predictors	Median	Dispersion	Missing
Age	30-35 year	1.65	0%
Marital status	Married and living with partner	0.53	0%
Age marriage	> 15 years	0.53	0%
Age at onset of the sex	16-20 year	1.04	0%
Age at first child	15-20 years	0.919	0%
Life time sexual partners	0-1	0.318	0%
Life time pregnancies	0-2	0.772	0%
Screened for cervical cancer	No	0.246	0%
If no why	Not informed / no awareness	1.01	0%
No. of times screening	0	0.229	0%
Smoking consumption	No	0.447	0%
Usage of temporary family planning methods	No	0.582	0%
Family history of cervical cancer	No	0.41	0%
Infection of the cervix	Candidiasis , and bacterial vaginosis	1.32	0%

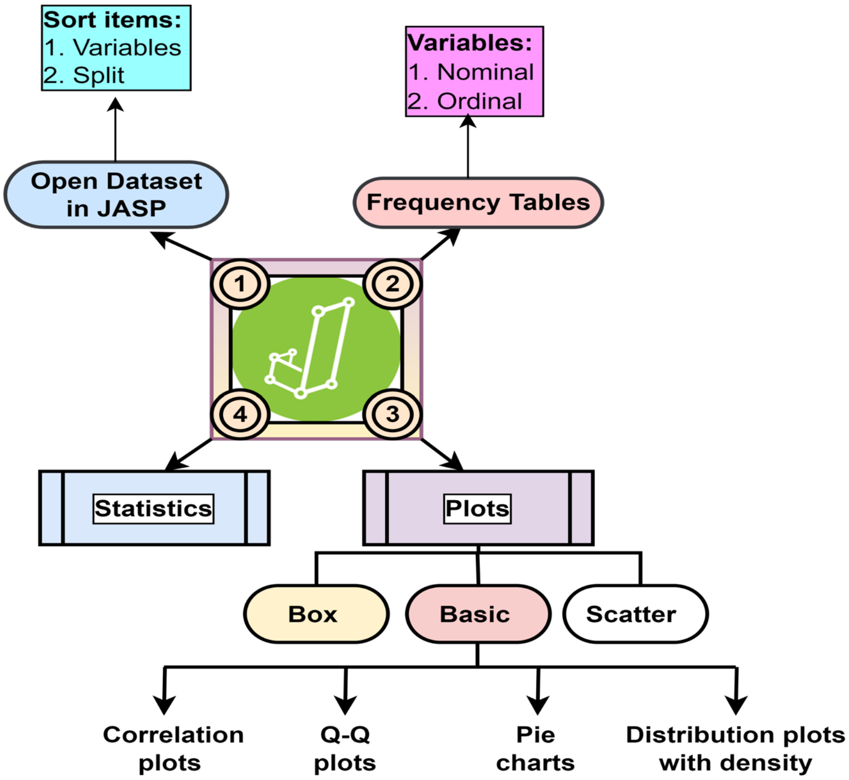

Figure 2. Descriptive Statistics in JASP Tool

the study to predict cervical cancer. These variables were measured in each participant. The numerical and categorical value of each target variable susceptible to cervical carcinoma was either negative/positive results.

Results of descriptive statistics

Table 2 illustrates the median, dispersion and missing values of the predictors used in the current study. The data provided appears to be a set of predictors and their corresponding median and dispersion values. Each predictor represents a factor or characteristic that may be related to the risk of developing cervical cancer. The study's median age was between 30 and 35 years, and the majority of participants had married and living with their partners, with a median age at marriage larger than 15 years and a dispersion value of 0.53. The median age of first sexual activity is between 16 and 20 years' age, and the majority of participants had 0-1 lifetime sexual partners, with a dispersion value of 0.318. Surprisingly, the majority of participants had never been screened for cervical cancer, with a dispersion value of 0.24. There was no missing data among the selected predictors. Figure 3 Showed the data of distribution of the positive cases for cervical cancer. A total of 298 women participated in

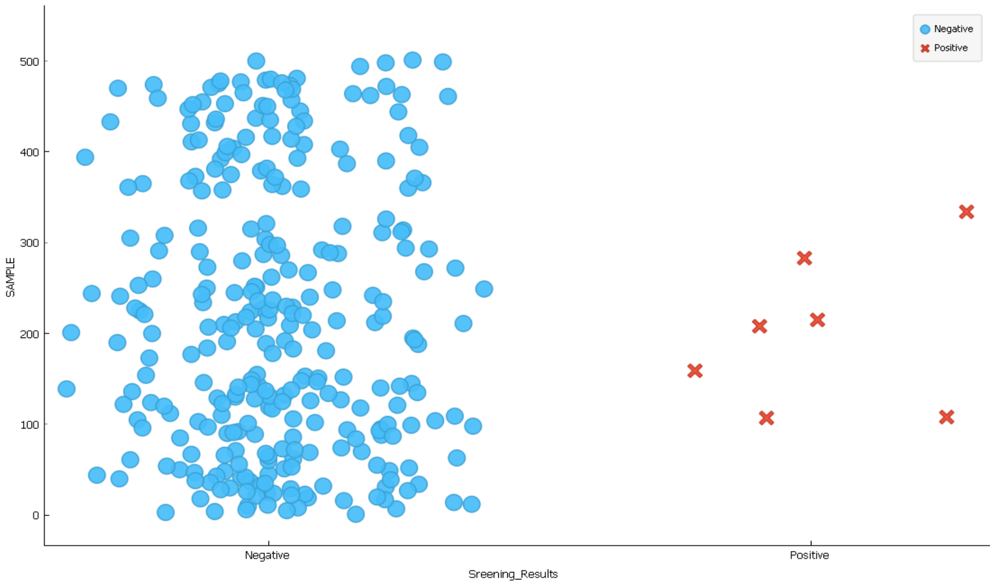

Figure 3. Distribution of the Positive Cases for Cervical Cancer

the current study for cervical screening (Pap smear and Biopsy). The scatter diagram depicts, 7 positive and 291 negative results for cervical cancer.

Finding Predictors of cervical cancer using machine learning classifiers

Table 3 shows that, in the initial stages of modeling, four variables are categorized under exclusion criteria based on at least one variable being depicted in each ML algorithm; none of those listed in this table fit this criterion. As a result, researchers have exempted these four variables from further data analysis.

Table 4 presents the classification performance of algorithms in the second stage of modelling with tuned hyper-parameters. The comparison of all six models is shown in Table 4. Models 1, 2, and 4 of the logistic regression algorithm and models 3, 5, and 6 of the Decision Tree algorithm produced better classification prediction outcomes for cervical cancer prediction. Table 4. Classification performance of algorithms in the second stage of modelling with tuned hyper-parameters. Figure 4. Presents the ROC curve algorithm classification. The relevant variables considered for cervical cancer prediction were carried out at the second stage of modeling

Table 3. Categorization of Predictors in Exclusion Criteria during 1st Stage of the Modeling

Predictor	DT	RF	LR	MLP	Naïve Bayes	Occurrences
Age	Nil	Nil	Nil	Yes	Nil	1
Age at marriage	Nil	Yes	Nil	Nil	Nil	1
Number of times screened	Nil	Nil	Nil	Nil	Nil	0
Life time pregnancies	Nil	Nil	Nil	Nil	Nil	0

Table 4. Classification Performance of Algorithms in the Second Stage of Modelling with Tuned Hyper-Parameters

Model	Algorithm	AUC	CA	F1	Precision	Recall	Sensitivity	Specificity	Accuracy
Model 1	Decision Tree	0.59	0.79	0.79	0.79	0.79	0.88	0.14	0.79
	Logistic Regression	0.55	0.84	0.82	0.81	0.84	0.94	0.14	0.87
	Multi-layer Perceptron	0.54	0.82	0.81	0.8	0.82	0.92	0.14	0.82
	Naive Bayes	0.68	0.72	0.76	0.84	0.72	0.74	0.57	0.72
	Random Forest	0.47	0.68	0.72	0.77	0.68	0.76	0.14	0.68
Model 2	Decision Tree	0.49	0.74	0.76	0.78	0.74	0.74	0.14	0.74
	Logistic Regression	0.7	0.88	0.85	0.84	0.88	0.88	0.14	0.84
	Multi-layer Perceptron	0.64	0.84	0.82	0.81	0.84	0.84	0.14	0.84
	Naive Bayes	0.69	0.68	0.74	0.86	0.68	0.68	0.71	0.68
	Random Forest	0.48	0.82	0.81	0.8	0.82	0.82	0.14	0.82
Model 3	Decision Tree	0.67	0.84	0.82	0.81	0.84	0.84	0.14	0.84
	Logistic Regression	0.51	0.84	0.82	0.81	0.84	0.84	0.14	0.84
	Multi-layer Perceptron	0.44	0.79	0.79	0.79	0.79	0.79	0.14	0.79
	Naive Bayes	0.75	0.75	0.79	0.85	0.75	0.75	0.57	0.75
	Random Forest	0.51	0.79	0.8	0.81	0.79	0.79	0.29	0.79
Model 4	Decision Tree	0.81	0.86	0.83	0.82	0.86	0.86	0.57	0.86
	Logistic Regression	0.83	0.89	0.89	0.89	0.89	0.89	0.14	0.89
	Multi-layer Perceptron	0.77	0.81	0.8	0.79	0.81	0.81	0.14	0.81
	Naive Bayes	0.71	0.7	0.75	0.84	0.7	0.7	0.57	0.7
	Random Forest	0.54	0.82	0.81	0.8	0.82	0.82	0.14	0.82
Model 5	Decision Tree	0.59	0.84	0.82	0.81	0.84	0.84	0.14	0.84
	Logistic Regression	0.56	0.74	0.76	0.78	0.74	0.74	0.14	0.74
	Multi-layer Perceptron	0.47	0.75	0.77	0.78	0.75	0.75	0.14	0.75
	Naive Bayes	0.66	0.72	0.76	0.82	0.72	0.72	0.43	0.72
	Random Forest	0.66	0.82	0.81	0.8	0.82	0.82	0.14	0.82
Model 6	Decision Tree	0.48	0.83	0.82	0.81	0.83	0.83	0.29	0.88
	Logistic Regression	0.69	0.75	0.78	0.83	0.75	0.75	0.57	0.75
	Multi-layer Perceptron	0.49	0.79	0.79	0.79	0.79	0.79	0.29	0.79
	Naive Bayes	0.72	0.75	0.78	0.83	0.75	0.75	0.57	0.75
	Random Forest	0.57	0.73	0.75	0.77	0.73	0.73	0.29	0.73

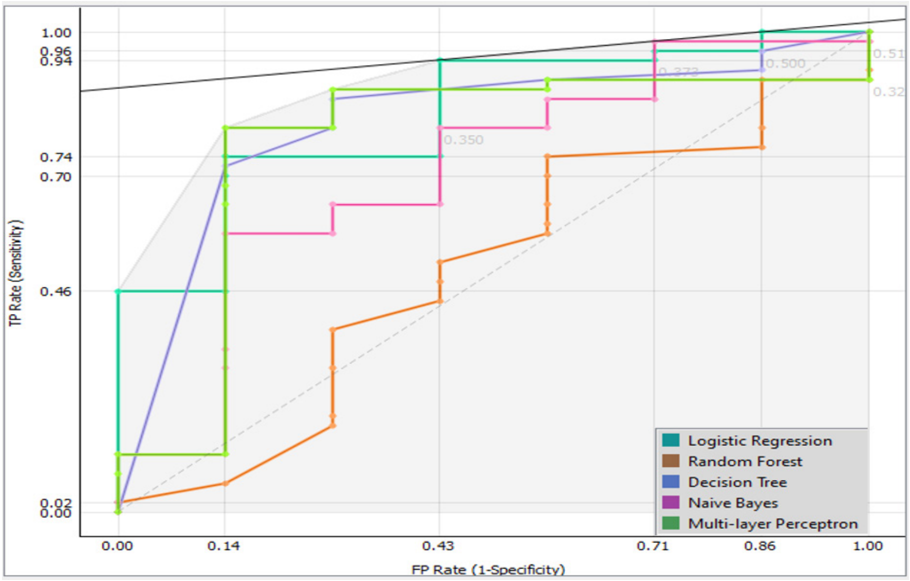

Figure 4 (a). Model 1- ROC Curve-based Classification Prediction

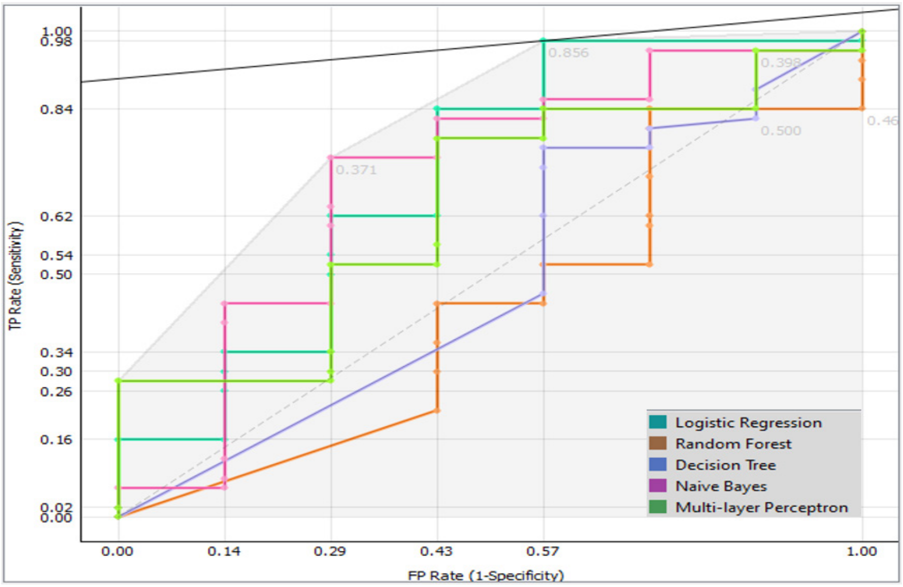

Figure 4 (b). Model 2- ROC Curve-based Classification Prediction

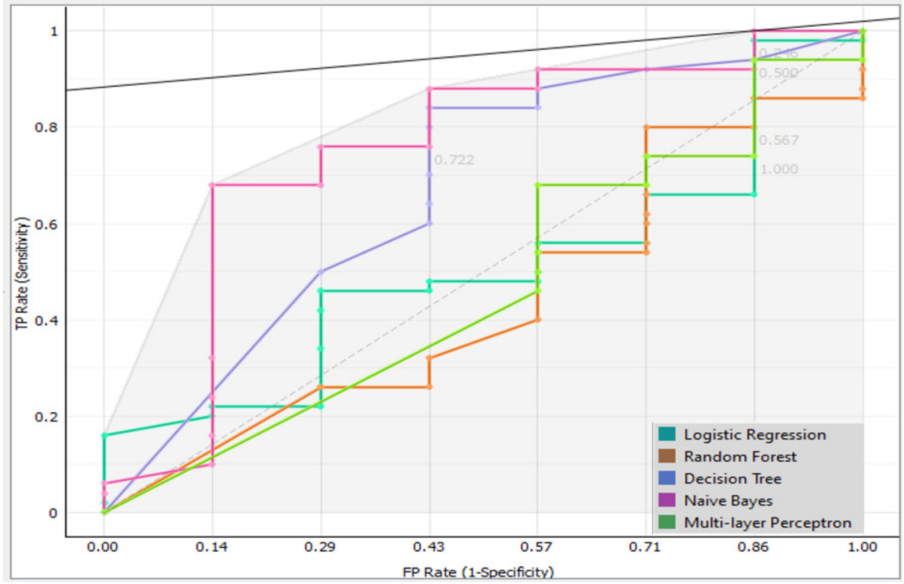

Figure 4 (c). Model 3-ROC Curve-based Classification Prediction

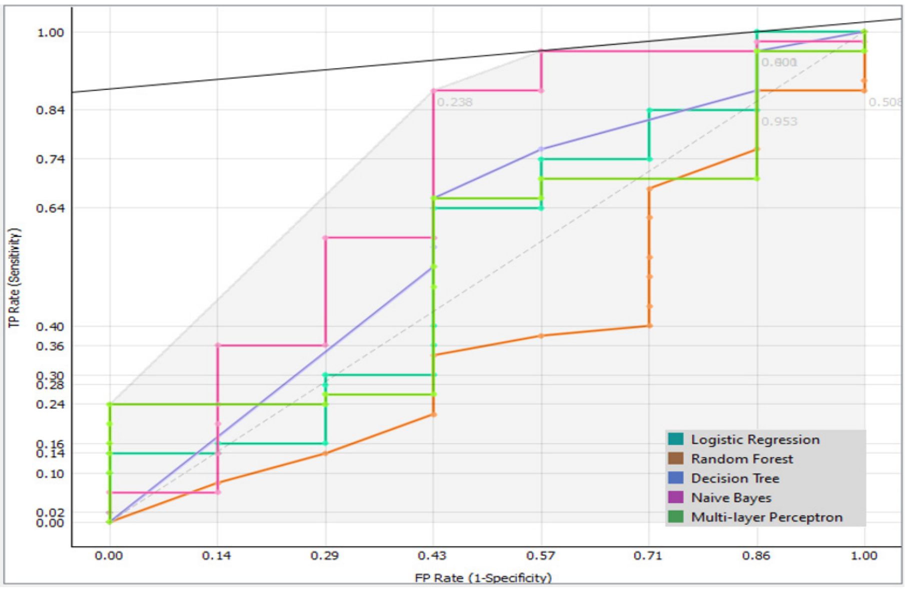

Figure 4 (d). Model 4- ROC Curve-based Classification Prediction

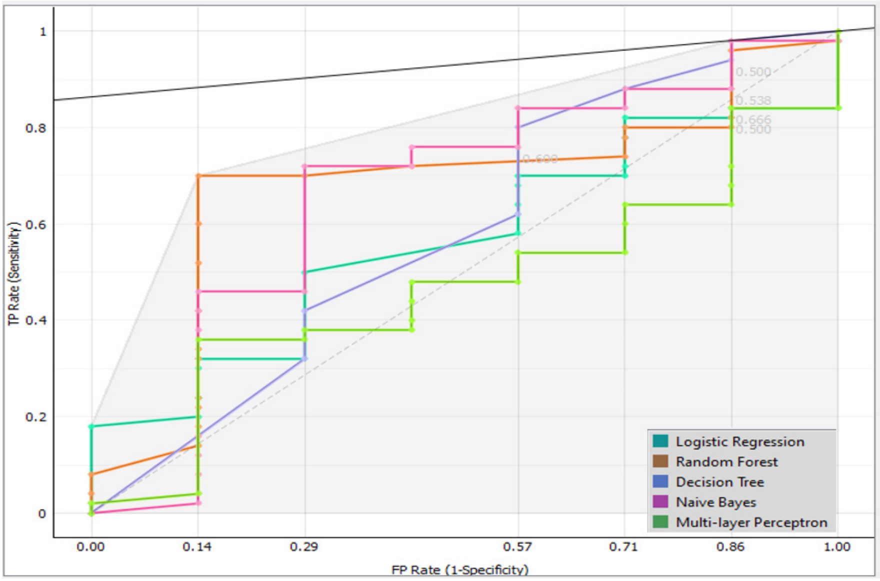

Figure 4 (e). Model 5- ROC Curve-based Classification Prediction

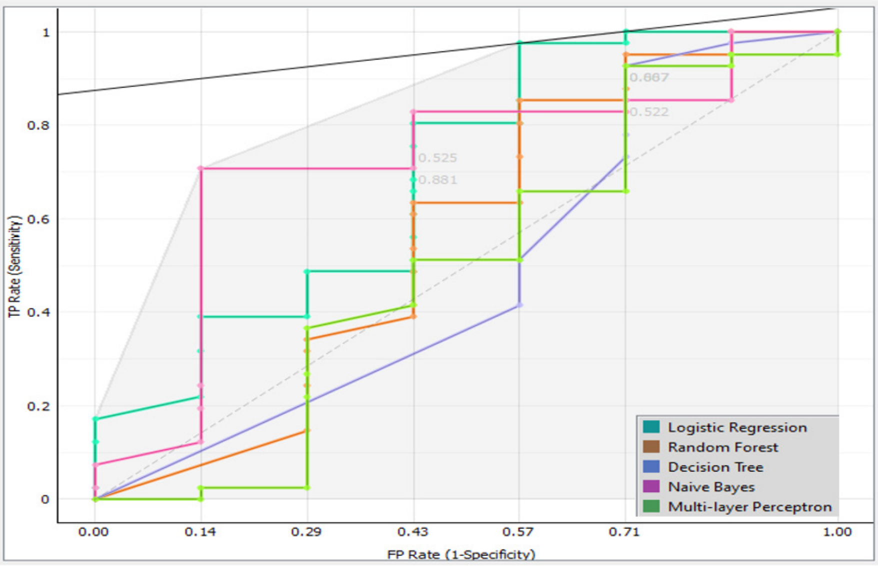

Figure 4 (f). Model 6- ROC Curve-based Classification Prediction

(Table 4), and the parameters were evaluated individually for each model. Based on the evaluation criteria of specificity, sensitivity, and area under the ROC curve, Random Forest, Naive Bayes, MLP, Logistic Regression, and Decision Tree machine algorithms correspondingly performed the finest. Figure 4 exhibits the results of evaluating ROC curve and the area under the ROC curve for the algorithms run in the second stage of modelling revealed that the Naive Bayes (in 1, 3, and 5 models), Logistic Regression, (in 2,4, and 6 models) had the utmost area under the ROC curve for the test data while other others have performed similarly for the training data.

Figure 5 Shows the significant predictors appeared by tree viewer Gini Index. Table 5 depicts the significant final predictors used in the current study. The 14 variables that are confirmed in the second stage of modeling

Table 5. Significant Predictors Appeared in the Second Stage of the Modeling

Sr. No.	Predictor	Random Forest	Naive Bayes	MLP	Logistic Regression	Decision Tree	Occurrence
1.	Age at onset of the sex	Yes	Yes	Yes	Yes	Yes	5
2.	Life time sexual partners	Yes	Yes	Yes	Yes	Yes	5
3.	Screened for cervical cancer	Yes	Yes	Yes			3
4.	Reasons for not screening	Yes	Yes	Yes	Yes	Yes	5
5.	Smoking consumption	Yes	Yes	Yes	Yes	Yes	5
6.	Family history cervical cancer		Yes	Yes	Yes		3
7.	Infection of the cervix		Yes	Yes	Yes	Yes	4
8.	Duration of contraceptives used	Yes	Yes	Yes	Yes	Yes	5
9.	Type of the family planning methods	Yes	Yes	Yes	Yes		4
10.	Family history cervical cancer	Yes		Yes	Yes	Yes	4
11.	Marital status	Yes	Yes	Yes			3
12.	Age at the first child	Yes		Yes	Yes	Yes	4
13.	Duration of smoking	Yes		Yes	Yes	Yes	4
14.	Smoking Consumption	Yes		Yes	Yes	Yes	4

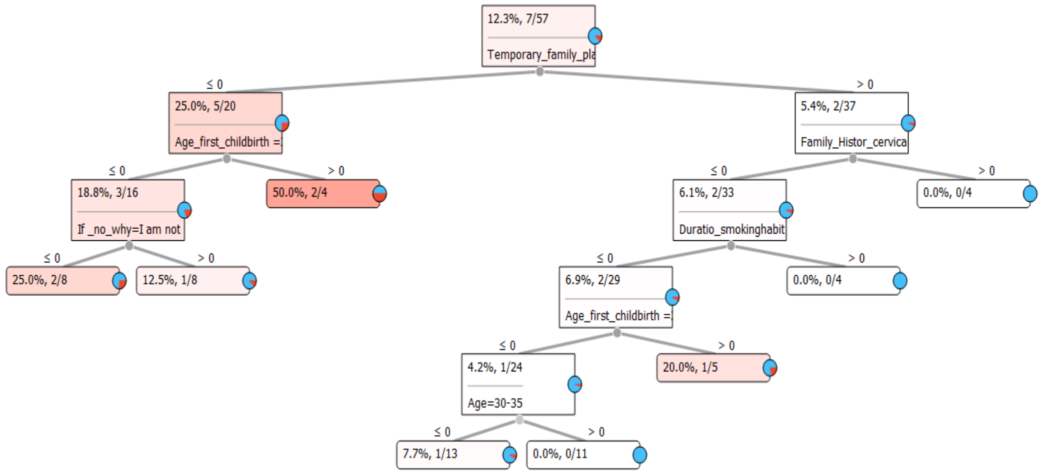

Figure 5 (a). Model 1- Significant predictors appeared by tree viewer Gini Index

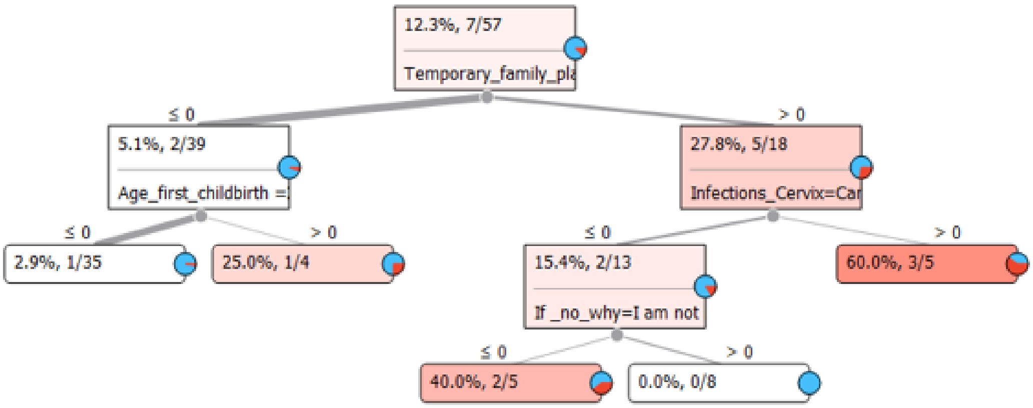

Figure 5 (b). Model 2- Significant Predictors Appeared by Tree Viewer Gini Index.

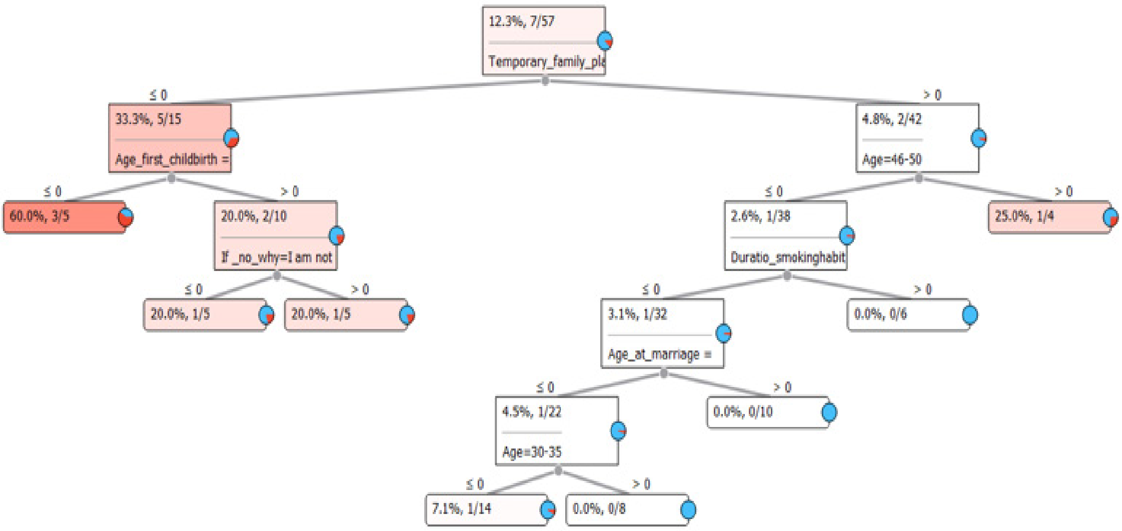

Figure 5 (c). Model 3- Significant Predictors Appeared by Tree Viewer Gini Index.

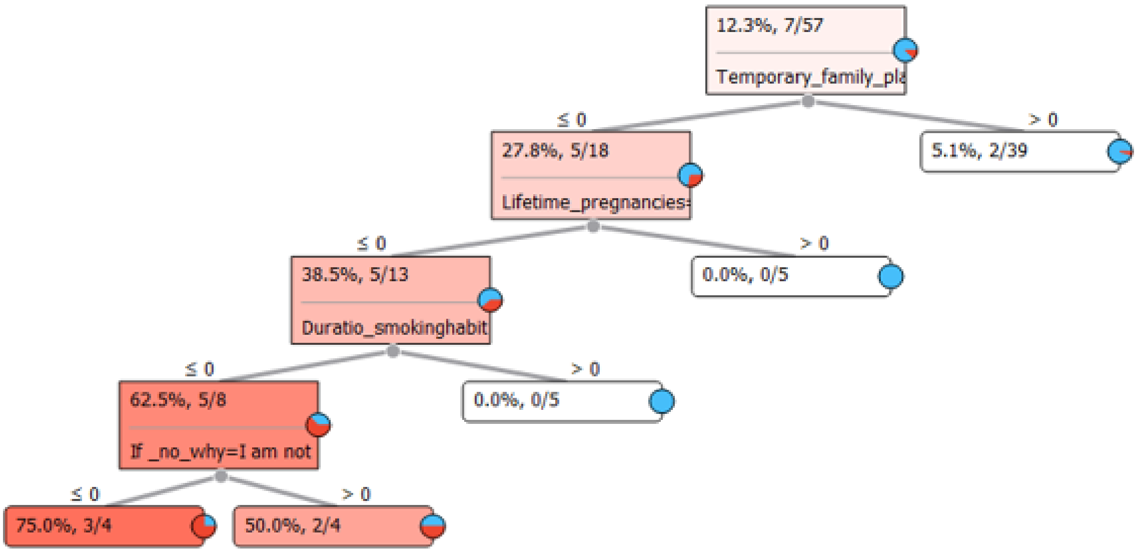

Figure 5 (d). Model 4- Significant Predictors Appeared by Tree Viewer Gini Index.

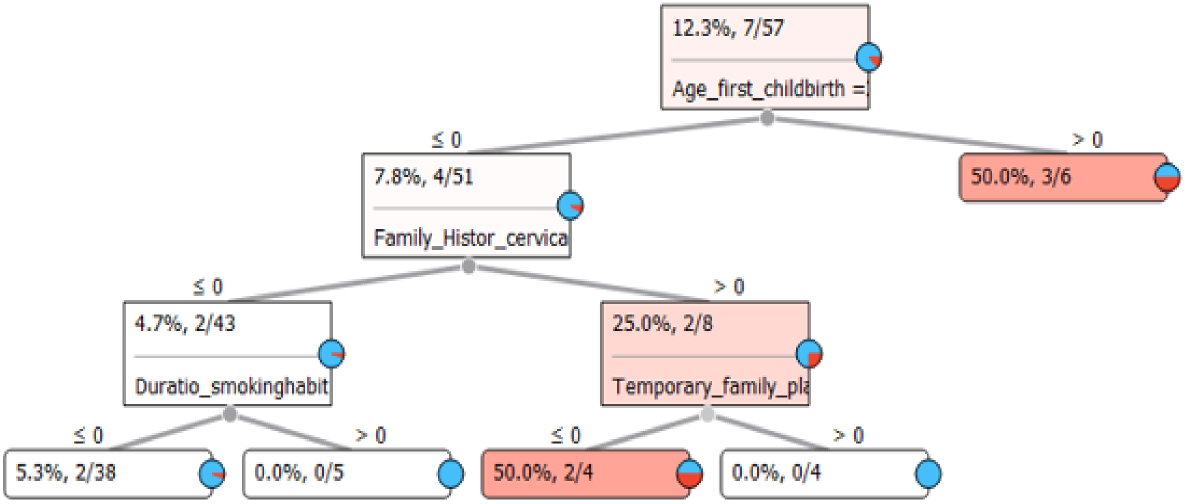

Figure 5 (e). Model 5- Significant Predictors Appeared by Tree Viewer Gini Index.

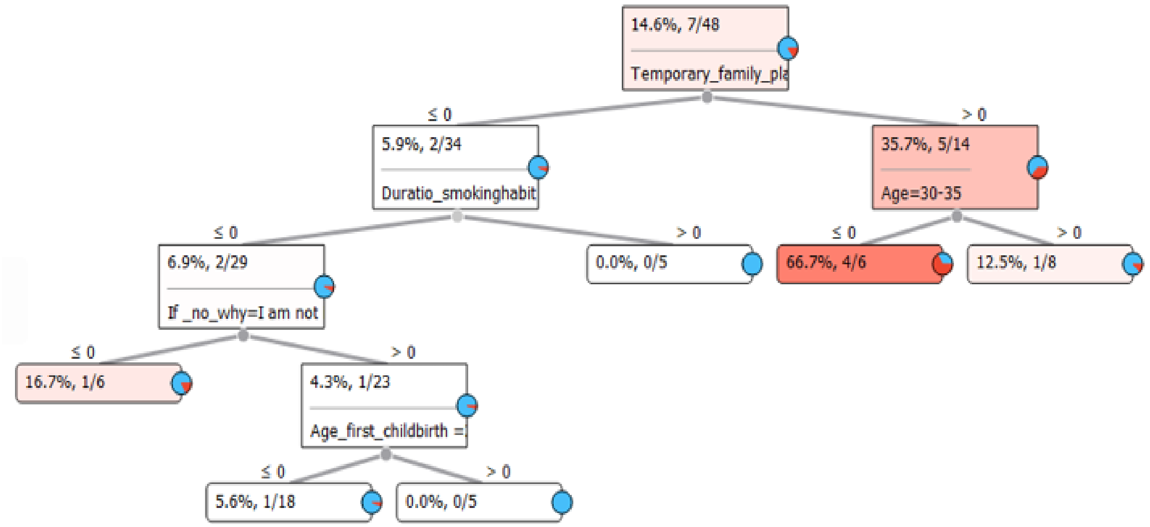

Figure 5 (f). Model 6- Significant Predictors Appeared by Tree Viewer Gini Index

as the predictors for prediction of the cervical cancer. The women's participation in cervical screening was also evaluated in the current study. Merely 29 (4.79%) women out of 501 actively participated in cervical cancer screening prior to the start of our study. The majority of women have never heard of cervical screening or its advantages.

Following the evaluation of women's participation in cervical screening, the researcher expressed a keen interest in analyzing the barriers to non-participation in cervical screening among the participants. The vast majority (51.70%) of women were unaware of cervical. cancer screening and had never had it done (94.21%). Study subjects had reported a variety of reasons for nonparticipation, including embarrassment over the test (40.32%), the belief that the test is expensive (33.93%), non-acceptance by family members (16.13%), not knowing where the test would be administered (30.14%), and the assumption that she is healthy and does not need the test (50.10%).

Discussion

The current study focused on developing a model for predicting cervical cancer, considering the most probable susceptible variables. Cervical cancer does not display signs or symptoms until it has progressed to the later stages, where the prognosis is poor; therefore, predicting and detecting cervical cancer at an early stage is critical to reducing morbidity and mortality rates among middle-aged women. As a result, the current study's researchers analyzed essential relevant predictors and the efficacy of the widely used algorithms in predicting cervical cancer. In our study, 17 risk factors were used as predictors of cervical cancer. Age, age at marriage, number of times screened, and lifetime pregnancies were excluded in the first step of modeling because they did not occur in more than two algorithms. However, in contrast, age, age at marriage, and lifetime pregnancies are influential factors in cervical cancer prediction (Kashyap et al., 2019).

In the present study, due to the imbalance in the dataset, the authors divided the data into six models, each having 57 samples (except the sixth model, which contained 45 samples), and these were compared with samples of positive screening findings. Following the data preparation, five algorithms were statistically evaluated: Decision Tree, Logistic Regression, Multi-layer Perceptron, Naive Bayes, and Random Forest. In the first, second, and fourth modeling of the data, the Logistic Regression displayed 88 % to 94% sensitivity with 84% to 89% of accuracy for classification prediction results for cervical cancer, while in the third, fifth, and sixth models of the data, the Decision Tree method demonstrated 83% to 84% sensitivity and 84 % to 88% accuracy in predicting cervical cancer. When analyzing the Receiver Operating Characteristic (ROC) curve and the area under the ROC curve for the algorithms run in the second phase of the modeling techniques, the greatest area under the ROC curve was discovered to be linked with the Naive Bayes and Logistic Regression algorithms machine for the testing dataset. In contrast, all the techniques performed similarly for the training data in the following study. Authors (Asadi et al., 2020) conducted a cross-sectional study in Iran with participants of 145 and 23 testing features. The data were analyzed by machine learning algorithms which contain SVM, QUEST, C&R tree, MLP, and RBF. Accuracy, sensitivity, specificity, and area under the curve (AUC) were the measurement criteria to evaluate the algorithms. The accuracy, sensitivity, specificity, and AUC of MLP were 90.9%, 90%, 91.67%, and 91.5% respectively. Personal health, relationship status, social status, contraception dosage, education level, and frequency of cesarean births were the significant predictors in the Algorithms. Another study, conducted by (Vidya et al., 2006). segmented the data with the attributes of 500 datasets and 100 testing datasets; it showed the greatest results related to MLP with 98% of accuracy, 98% of sensitivity, and 99 % of the area under ROC curve when matching with other algorithms. In the present study, all five algorithms showed better performance. However, Logistic Regression and Decision Tree have shown great performance classification results out of all six models. Compared to the current study, MLP-ANN and SVM obtained the greatest results in all indicators and the area under the ROC curve. This difference in findings can be addressed by choosing a greater sample size, such as 500 training data and 100 testing datasets (Vidya et al., 2006).

As per the results of Hemalatha et al., (2016). MLP algorithm showed the best results with 85.5% accuracy, a 78.94% sensitivity and a 60.72% precision, while in another study, Kusy et al., (2013) with a sample size of 107 displayed the results of 72% of accuracy, 69% of sensitivity, 74% of specificity, and 67% area under the ROC curve, where similar results are observed in the present study also. However, in all six models, the specificity percentage levels are less, possibly due to the significant difference in the data and target features. In Kusy et al., study, RBF neural network algorithm showed poorer performance with 55% of accuracy, 42% of sensitivity, 67% of specificity, and a 48% area under the ROC curve matched in the present study .

Last but not restricted, this paper gives a glimpse of how Artificial Intelligence can help in predicting

Cervical cancer. Despite recent breakthroughs in artificial intelligence (AI) and its applications in cancer, there are various constraints and obstacles that need to be overcome. Some of these are data access control, generalisation, building practical applications, explanations challenges, and obstacles related to education and competence in the subject. Even though numerous studies have shown Machine learning and artificial intelligence to be effective in the prediction of different types of cancers, the validation of specific ML algorithm is required to generalize the findings (Patil et al., 2020).

The short comings of the current study include, the small sample size with the imbalanced input and target datasets. Another types of ML algorithms might be administered as in our study; we have used only five algorithms to predict the cervical cancer. Nevertheless, as per the results of review studies and current study, it is concluded that ML algorithm approaches are quite beneficial in predicting cervical cancer. Large sample size is recommended to resolve the issues of imbalanced datasets.

In conclusion, the results of current study proved that machine learning algorithms could improve the cervical cancer predictions. Logistic Regression and Decision Tree have shown great performance classification results out of all six models. This study's results indicated that all five algorithms showed better performance. However, Logistic Regression had shown great performance classification results of 88 % to 94% sensitivity with 84% to 89% of accuracy and the Decision Tree method demonstrated 83% to 84% sensitivity and 84 % to 88% accuracy in predicting cervical cancer. For testing dataset, the highest area under the ROC curve was observed to be associated with the Naive Bayes and Logistic Regression methods machine. This study proved that we could obtain accurate sensitivity and accuracy prediction values even when there is an imbalanced dataset of input and output criteria.

Author Contribution Statement

Conceptualization, S.D., G.S.R., and H.R; methodology, S.D. and G.S.R.; software, G.S.R.; validation, S.D. and G.S.R.; formal analysis, S.D. and G.S.R.; investigation, S.D. and G.S.R.; resources, S.D., G.S.R., and H.R; data curation, S.D., G.S.R., and H.R; writing, S.D.; writing, review and editing, S.D., G.S.R., and H.R.; visualization, S.D., G.S.R.; supervision, S.D., G.S.R., and H.R; project administration, S.D.

Acknowledgements

General

The author expresses heartfelt gratitude to Symbiosis Centre for Research Innovation (SCRI) and Symbiosis International Deemed University for supporting authors throughout the study. The authors wish to express their sincere gratitude to the authorities of YCM hospital for data collection approval and also we would thank the women who participated in this study.

Funding Statement

Present study was funded by Symbiosis Centre for Research Innovation (SCRI) and Symbiosis International Deemed University (SIDU).

Approval

Obtained from Symbiosis College of Nursing, Symbiosis International Deemed University. Informed Consent Statement was obtained from the participants.

Ethical Declaration

Authors had obtained an ethical approval from Independent Ethics Committee (IEC) of Symbiosis International (Deemed University), Pune Maharashtra, India.

Data Availability

Data collected from the study subjects. The data are not publicly available due to privacy and ethical reasons.

Conflict of Interest

The authors declare no conflict of interest.

References

Abbas S, Jalil Z, Javed AR, et al (2021). BCD-WERT: a novel approach for breast cancer detection using whale optimization based efficient features and extremely randomized tree algorithm. *Peer J Comput Sci*, **7**, e390.

Ahishakiye E, Mwangi W, Muthoni P, et al (2021). Comparative performance of machine leaning algorithms in prediction of cervical cancer. IST-Africa Conference (IST-Africa), pp 1-13.

Ahsan MM, Luna SA, Siddique Z (2022). Machine-Learning-Based Disease Diagnosis: A Comprehensive Review. *Healthcare (Basel)*, **10**.

Al Mudawi N, Alazeb A (2022). A Model for Predicting Cervical Cancer Using Machine Learning Algorithms. *Sensors (Basel)*, **22**.

Alyafeai Z, Ghouti L (2020). A fully-automated deep learning pipeline for cervical cancer classification. *Expert Syst Appl*, **141**, 112951.

Aminisani N, Armstrong BK, Canfell K (2012). Cervical cancer screening in Middle Eastern and Asian migrants to Australia: a record linkage study. *Cancer Epidemiol*, **36**, e394-400.

Asadi F, Salehnasab C, Ajori L (2020). Supervised Algorithms of Machine Learning for the Prediction of Cervical Cancer. *J Biomed Phys Eng*, **10**, 509-13.

Ayoub A, Mahboob K, Javed AR, et al (2021). Classification and categorization of covid-19 outbreak in Pakistan. *Comput Mater Continua*, **69**, 1253-69.

Chen H, Liu J, Wen QM, et al (2021). CytoBrain: cervical cancer screening system based on deep learning technology. *J Comput Sci Technol*, **36**, 347-60.

Do HH, Taylor VM, Yasui Y, et al (2001). Cervical cancer screening among Chinese immigrants in Seattle, Washington. *J Immigr Health*, **3**, 15-21.

Hemalatha K, Rani DU (2016). Improvement of multilayer perceptron classification on cervical pap smear data with feature extraction. *Int J Innov Res Sci Eng Technol*, **20**, 419-24.

Jahan S, Islam MDS, Islam L, et al (2021). Automated invasive cervical cancer disease detection at early stage through suitable machine learning model. *SN Appl Sci*, **3**, 806.

Jaswinder Singh , Sandeep Sharma (2019). Prediction of cervical cancer using machine learning techniques. *Int J Appl Eng Res*, **14**, 2570-7.

Javed AR, Fahad LG, Farhan AA, et al (2021). Automated cognitive health assessment in smart homes using machine learning. *Sustain Cities Soc*, **65**, 102572.

Javed AR, Sarwar MU, Beg MO, et al (2020). A collaborative healthcare framework for shared healthcare plan with ambient intelligence. *Hum Centric Comput Inform Sci*, **2020**, 1-21.

Kashyap N, Krishnan N, Kaur S, et al (2019). Risk Factors of Cervical Cancer: A Case-Control Study. *Asia Pac J Oncol Nurs*, **6**, 308-14.

Kaur A, Mann KS (2018). Bone age classification using SVM' international journal of engineering science invention. *Int J Eng Sci Invention*, **7**, 38-45.

Khalil A, Prasad CSRK (2019). Introducing of mass rapid transit system (BRT) by using aggregate and disaggregate models. 1687-95, https://www.ijeat.org/wp-content/uploads/papers/v8i4/D6281048419.pdf.

Khamparia A, Gupta D, Rodrigues JJ, et al (2021). DCAVN: cervical cancer prediction and classification using deep convolutional and variational autoencoder network. *Multimedia Tools Appl*, **80**, 30399–415.

Kusy M, Obrzut B, Kluska J (2013). Application of gene expression programming and neural networks to predict adverse events of radical hysterectomy in cervical cancer patients. *Med Biol Eng Comput*, **51**, 1357-65.

Latha DS, Lakshmi PV, Fathima S (2014). Staging Prediction in Cervical Cancer Patients–A Machine Learning Approach. *Int J Innovative Res Pract*, **2**, 14-23.

Lilhore UK, Poongodi M, Kaur A, et al (2022). Hybrid Model for Detection of Cervical Cancer Using Causal Analysis and Machine Learning Techniques. *Comput Math Methods Med*, **2022**, 4688327.

Lu L, Song E, Ghoneim A, et al (2020). Machine learning for assisting cervical cancer diagnosis: An ensemble approach. Future Gener. *Comput Syst*, **106**, 199-205.

MacCosham A, El-Zein M, Burchell AN, et al (2020). Transmission reduction and prevention with HPV vaccination (TRAP-HPV) study protocol: a randomised controlled trial of the efficacy of HPV vaccination in preventing transmission of HPV infection in heterosexual couples. *BMJ Open*, **10**, e039383.

Mandelblatt J, Andrews H, Kerner J, et al (1991). Determinants of late stage diagnosis of breast and cervical cancer: the impact of age, race, social class, and hospital type. *Am J Public Health*, **81**, 646-9.

Mohd Fakharuddin Z, Muhammad Khusairi O, Iza Sazanita I, et al (2019). Classification of Cervical Cancer Using Hybrid Multi-layered Perceptron Network Trained by Genetic Algorithm. *Procedia Comput Sci*, **163**, 494-501.

Mukama T, Ndejjo R, Musabyimana A, et al (2017). Women's knowledge and attitudes towards cervical cancer prevention: a cross sectional study in Eastern Uganda. *BMC Womens Health*, **17**, 9.

Mukhopadhyay S, Kurmi I, Dey R, et al (2016). Optical diagnosis of colon and cervical cancer by support vector machine. Biophotonics: Photonic Solutions for Better Health Care V, 98870U; Brussels, Belgium: International Society for Optics and Photonics.

Nematollahi M, Akbari R, Nikeghbalian S, et al (2017). Classification Models to Predict Survival of Kidney Transplant Recipients Using Two Intelligent Techniques of Data Mining and Logistic Regression. *Int J Organ Transplant Med*, **8**, 119-22.

Nithya B, Ilango V (2019). Evaluation of machine learning based optimized feature selection approaches and classification methods for cervical cancer prediction. *SN Appl Sci*, **1**, 641.

Parkin DM, Bray F, Ferlay J, et al (2005). Global cancer statistics, 2002. *CA Cancer J Clin*, **55**, 74-108.

Patil S, Moafa IH, Mosa Alfaifi M, et al (2020). Reviewing the role of artificial intelligence in cancer. *Asian Pac J Cancer Biol*, **5**, 189-99.

Ramnath GS, Harikrishnan R (2021). Households electricity consumption analysis: a bibliometric approach. *Libr Philos Pract*, **5098**. https://digitalcommons.unl.edu/libphilprac/5098.

Ramondetta L (2013). What is the appropriate approach to treating women with incurable cervical cancer?. *J Natl Compr Canc Netw*, **11**, 348-55.

Randall TC, Ghebre R (2016). Challenges in Prevention and Care Delivery for Women with Cervical Cancer in Sub-Saharan Africa. *Front Oncol*, **6**, 160.

Rezaianzadeh A, Dastoorpoor M, Sanaei M, et al (2019). Predictors of length of stay in the coronary care unit in patient with acute coronary syndrome based on data mining methods. *Clin Epidemiol Glob Health*, **8**, 383-8.

Salmi N, Rustam Z (2019). Naive Bayes classifier models for predicting the colon cancer. In: IOP Conference Series: Materials Science and Engineering. *IOP Publishing*, **546**, 052068.

Sarwar MU, Javed AR (2019). Collaborative health care plan through crowdsource data using ambient application. In: 2019 22nd International Multitopic Conference (INMIC), IEEE, pp 1-6.

Schiffman M, Castle PE, Jeronimo J, et al (2007). Human papillomavirus and cervical cancer. Lancet, 370, 890-907.

Singh SK, Goyal A (2020). Performance analysis of machine learning algorithms for cervical cancer detection. *Int J Healthcare Inf Syst Inf*, **15**, 1-21.

Tiwari S, Lilhore U, Singh A (2018). Artificial neural network and genetic clustering based robust intrusion detection system. *Int J Comput Appl*, **179**, 36-40.

Unlersen MF, Sabanci K, Özcan M (2017). Determining cervical cancer possibility by using machine learning methods. *Int J Latest Res Eng Technol*, **3**, 65-71.

Vidya R, Nasira GM (2006). Knowledge extraction in medical data mining: a case based reasoning for gynecological cancer an expert diagnostic method. *ARPN J Eng Appl Sci*,

10, 3997-4001.

William W, Ware A, Basaza-Ejiri AH, et al (2019). Cervical cancer classification from Pap-smears using an enhanced fuzzy C-means algorithm. *Inform Med Unlocked*, **14**, 23-33.

Workowski K, Bolan GA (2015). Sexually Transmitted Diseases Treatment Guidelines. *Morb Mortal Wkly Rep*, **64**, 1-137.

Wright RE (1995). Reading and Understanding Multivariate Statistics. American Psychological Association; Washington, DC, USA . Logistic regression.