Research Article

Automated invasive cervical cancer disease detection at early stage through suitable machine learning model

Sohely Jahan[1] · M. D. Saimun Islam[1] · Linta Islam[2] · Tamanna Yesmin Rashme[3] · Ayesha Aziz Prova[4] · Bikash Kumar Paul[5,6,7] · M. D. Manowarul Islam[2] · Mohammed Khaled Mosharof[8]

Received: 27 April 2021 / Accepted: 31 August 2021
Published online: 16 September 2021

Abstract

Cervical cancer is a common cancer that affects women all over the world. This is the fourth leading cause of death among women and has no symptoms in its early stages. At the cervix, cervical cancer cells develop slowly. If it can be detected early, this cancer can be successfully treated. Health professionals are now facing a major challenge in detecting such cancer until it spreads rapidly. This study applied various machine learning classification methods to predict cervical cancer using risk factors. The main aim of this research work is to be described of the performance variation of eight most classifications algorithm to detect cervical cancer disease based on the selection of various top features sets from the dataset. Multilayer Perceptron (MLP), Random Forest and k-Nearest Neighbor, Decision Tree, Logistic Regression, SVC, Gradient Boosting, AdaBoost are examples of machine learning classification algorithms that have been used to predict cervical cancer and help in early diagnosis. A variety of approaches are used to avoid missing values in the dataset. To choose the various best features, a combination of feature selection techniques such as Chi-square, SelectBest and Random Forest was used. The performance of those classifications is evaluated using the accuracy, recall, precision and f1-score parameters. On a variety of top feature sets, MLP outperformed other classification models. The majority of classification models, on the other hand, claim to have the highest accuracy on the top 25 features in dataset splitting ratio (70:30). For each model, the percentage of correctly classified instances has been presented and all of the results are then discussed. Medical professionals will be able to use the suggested approach to perform research on cervical cancer.

Keywords Cervical cancer · Classification · Early-stage detection · Features selection · SVC · Multilayer perceptron

1 Introduction

Invasive happens in the woman's cervix by affecting the deeper tissues of the cervix. The cervical cancer can spread to other parts of their body lungs, liver, bladder, vagina and rectum. Healthy cells in the cervix develop changes (mutations) growing, multiplying at a set rate and eventually dying at a set time in their DNA. Moreover, abnormal mutational activities of unhealthy cell are growing, multiplying cells out of control as well as they do not die

✉ Bikash Kumar Paul, bikash.k.paul@ieee.org | [1]Department of Computer Science and Engineering, University of Barisal, Kornokathi, Barisal 8200, Bangladesh. [2]Department of Computer Science and Engineering, Jagannath University, Dhaka 1100, Bangladesh. [3]Department of Computer Science and Engineering, Uttara University, Dhaka 1230, Bangladesh. [4]Department of Computer Science and Engineering, Central Women's University, Dhaka 1203, Bangladesh. [5]Department of Software Engineering (SWE), Daffodil International University (DIU), Sukrabad, Dhaka 1207, Bangladesh. [6]Group of Biophotomatiχ, Mawlana Bhashani Science and Technology University, Santosh, 1902 Tangail, Bangladesh. [7]Department of Information and Communication Technology, Mawlana Bhashani Science and Technology University, Santosh, 1902 Tangail, Bangladesh. [8]Department of ICT, ICT Division, Telecommunication & Information Technology, Dhaka, Bangladesh.

accumulating abnormal cells form a mass (tumor). The most well-informed symptoms of cervical cancer unusual pain after sex, vaginal bleeding after sex, between periods, after menopause or after a pelvic examination as well as vaginal discharge. The most common diverse risk factors are many sexual partners, early sexual activity, sexually transmitted infections (STIs), weakened immune system, smoking, exposure to miscarriage prevention drugs and the like [1].

The sexually transmitted human papillomavirus (HPV) plays a vital role in cervical cancer and genital warts. Among 100 different strains of HPV, HPV-16 and HPV-18 are the most well-known strains for cancer. The cancer-causing strain of HPV is also responsible for vulvar cancer, vaginal cancer, penile cancer, anal cancer, rectal cancer, throat cancer [2]. Depending on spreading level cervical cancer has four stages: stage 1 only spreads to the lymph nodes, stage 2 larger cancer spreads outside of the uterus and cervix or to the lymph nodes, stage 3 spreads to the lower part of the vagina or the pelvis along with blocks the ureters, the tubes that carry urine from the kidneys to the bladder as well as final stage spreads outside of the pelvis to organs like your lungs, bones or liver [3]. The easiest initiatives to prevent cervical cancer are the HPV vaccine that have routine Pap tests (early-stage detection), practice safe sex, don't smoke and so on [4, 5].

The "National Strategy for Cervical Cancer Prevention and Control" program is launched by the Bangladesh Ministry of Health & Family Welfare (MoHFW) extending over five years from 2017 to 2022. Although World Health Organization (WHO) identified invasive cervical cancer as the fourth most common cancer in women, this is the second most common type of cancer among women between 15 and 44 years of age in Bangladesh. Every year new cases are diagnosed at approximately 12,000, and the severity of the disease brings over 6000. About 4.4% of women in the general population have a high inclination to cervical HPV16/18 infection at a given time, and 80.3% of invasive cervical cancers are attributed to HPVs 16 or 18 in the region Bangladesh of Southern Asia [6–9].

Various types of ample prevention measures are practiced, but the occurrence of cervical cancer cannot stop only using screening tests. The early-stage detection of this cancer can play an important role to control death due to invasive cervical cancer. Currently, computer vision, artificial intelligence (AI), machine learning (ML), deep learning (DL) are the most used popular term for detecting various diseases. Among them, the various effective algorithm of ML model creates great attention by rapidly detecting the targeted diseases. The suitable ML algorithms can be applied to the targeted diseases dataset that preprocessed form by applying several preprocessing activities such as data cleaning, dimensionality reduction, feature selection. The desired analyzed result of this algorithm may assist the medical officers in diagnosing diseases rapidly and provide the best medications for their patients.

This research selects a variety of different top features using a combination of feature selection techniques, reducing training time and assisting oncologists in quickly detecting cervical cancer, and the main objectives of this research are to improve classification performance using machine learning classification techniques and provide the performance results analysis based on different top features set.

The remaining part of the paper is organized as follows: Sect. 2 reviews all relevant works in classification algorithms for cervical cancer. The proposed methodology for detecting cervical cancer using various classification algorithms and feature selection methods is explained in Sect. 3. Sections 4 and 5 focus on result and discussion. The conclusion of the paper is discussed in Sect. 6.

2 Related works

The incredible machine learning (ML) has diverse application scope for the lion portion of diseases detection of all kinds of animals and plants. Currently, a plethora of ML models are proposed and applied for the targeted application areas to accelerate and enrich the research purposes. In [10], four target variables—Hinselmann, Schiller, Cytology, and Biopsy—with 32 risk factors are considered for analyzing pernicious cervical cancer leading to unexpected death. These major culprits are sorted out by developing an effective model and applying the most popular ML methods: Logistic Regression, Decision Tree, Random Forest, and Ensemble Method. The analyzed result demonstrates cervical cancer prediction accuracy with ROC curve, AUC curve, respectively, 98.56%, 99.50% for Ensemble method using SMOTE-Voting-PCA proposed model. Moreover, SMOTE-Voting PCAM model executes the prediction of Schiller target variable, respectively, 98.49%, 98.60% and 99.80%.

Jiayi Lu et al. [11] studied that the ghastly prevalent cervical cancer is a burdensome matter in both developing and developed countries due to lack of awareness and health diagnostic facilities. As a result, the late detection of the occurred cervical cancer and the high cost of this diseases treatment creates more perilous situation for patients. The UCI dataset and the private dataset are used for this research work. The most interesting ensemble approach containing LR, DT, SVM, MLP, KNN to this issue has been discussed to predict the risk of cervical cancer. The predicted performance of the proposed approach is improved by adopting data correction mechanism and gene-assistance module in this work. The obtained accuracy of the proposed method was 83.16%.

In the last decade, the most occurring cancer of women body has attracted much attention from research teams for early-stage detection and treatment. In this work [12], machine learning with classification algorithms like Multilayer Perceptron, Decision Trees, Random Forest, K-Nearest Neighbor, Naïve-Bayes are applied for early-stage cervical cancer detection using risk factor collected dataset from UCI. The combined several machine learning techniques into one model demonstrated the accuracy of 87.21% using training and validation operations.

In this work [13], the dataset is introduced by surveying 858 patients with a number of attributes like 33 to make a predictive analysis about extinguishing cervical cancer. This dataset was split into training and test portion where various machine learning with classification methods like Multilayer Perceptron, BayesNet and k-Nearest Neighbor are applied for correct prediction. The performance analysis of this algorithm is based on confusion matrix using correctly classified instances and percentage.

This paper [14] was published in the Special Section on Fault Diagnosis, Data-Driven Monitoring, and Control of Cyber-Physical Systems in IEEE Access. According to the authors [14], several data-driven approaches such as principal component analysis (PCA), particle swarm optimization (PSO), fuzzy positivistic C-means clustering, linear regression (LR), artificial neural network (ANN) and support vector machine (SVM) have been proposed and implemented in the field in recent years to provide timely diagnosis.

A plethora of effective machine learning algorithm is applied on the most popular Pap smear image dataset to diagnose cervical cancer at an early stage [15]. The performance metrics of confusion matrix are precision call, recall score, F1 score and accuracy. Using confusion matrix and cost-effectiveness in terms of CPU time is used for the performance analysis of various algorithms. The selected best feature and reduced processing time are obtained by this proposed method to help oncologists in early detection of cervical cancer. In this work, Logistic Regression (LR) exhibits 100% accuracy requiring more CPU time. However, 99% accuracy is obtained in the exchange of less CPU time.

This paper deals with various selections of top feature sets using a combination of feature selection techniques, where previous related papers focused on only one selection of top features. All previous related works deal with only one dataset splitting ratio; on the other hand, this paper deals with three different dataset splitting ratios.

3 Proposed method

Our main goal is early-stage cervical cancer detection through classification algorithms on various top features of the dataset. A model has been proposed for this reason. The proposed methodology is divided into the following subtasks as shown in Fig. 1. In this approach, some important stages need extra concentration: (1) pre-processing and features selection and (2) model building and analysis. Each step is discussed in detail below:

3.1 Dataset description

The dataset was collected from the University of California, Irvine (UCI) Machine Learning dataset repository. The data were gathered at the Universitario Hospital de Caracas in

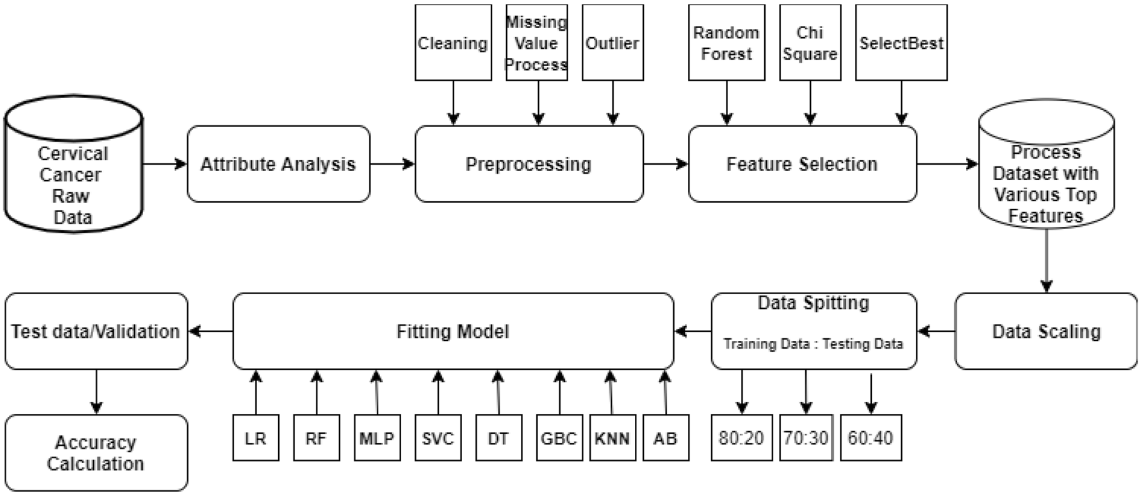

Fig. 1 The proposed model

Caracas, Venezuela, in 2017. A total of 858 patients' cases for 32 features as well as four target variables—Hinselmann, Schiller, Cytology and Biopsy displaying demographics information, behaviors and medical records—are included in the dataset. There are many missing values in this dataset, due to many patients not answering questions because of privacy concerns [16]. The dataset's attributes are listed in Table 1 (Figs. 2 and 3).

3.2 Pre-processing

Data preprocessing is crucial step in data mining and machine learning (ML) field to handle missing value, noisy data, incomplete data, inconsistent data, outlier data of raw data. Preprocessing consists of a number of basic steps including cleaning, integration, transformation and reduction. The main objectives of data preprocessing are reducing data size, determining relationship among data, normalizing data, removing outlier and extracting features. Before applying ML models, basic six steps need to be performed for dealing with the targeted dataset. These steps are orderly followed such as importing library, importing dataset, taking care of missing data in dataset, encoding categorical data, splitting dataset into training set and test set [17]. Some columns in the dataset have few missing cells with question marks that the patients had skipped due to privacy issues. Initially, all the question marks were replaced with null value in the dataset. Some of the columns have categorical values, while others have distinct values. To solve this curse, missing values of categorical and discrete columns are replaced by the mode and mean values of that column, respectively. Feature scaling or data normalization is often used in data processing to normalize the range of independent variables or attributes of data values. As a result, before implementing classification algorithms for early-stage cervical cancer detection, we scale our data [Standard Scaling] and [MinMaxScaling] so that all features contribute equally to the result.

3.3 Feature selection

When developing a predictive model, feature selection is the process of selecting the best amount of features from the dataset that effectively contribute the most to the forecast variable or output. Irrelevant attributes in the dataset will reduce the model's prediction efficiency as well as the classifiers' overall performance in terms of accuracy and complexity. Many more benefits can be obtained by using feature selection, such as reduced training time, improved model accuracy and reduced over-fitting when modeling the dataset [18]. In classification techniques, there are several different types of feature selection methods. In this paper, we used the Chi-square test and SelectKBest methods, as well as Random Forest feature selection, to determine which features were more relevant. After applying those feature selection methods, we determined various top features such as 10, 15, 20, 25, 30. The various top features are detailed in the figure below:

Table 1 Attribute information

Feature	Value range	Type	Feature	Value range	Type
Age	13–84	Int	STDs: pelvic inflammatory	0/1	Bool
Number of sexual partners	0–28	Int	STDs: genital herpes	0/1	Bool
First sexual intercourse (age)	10–32	Int	STDs: molluscum contagiosum	0/1	Bool
Number of pregnancies	0–11	Int	STDs: AIDS	0/1	Bool
Smokes	0/1	Bool	STDs: HIV	0/1	Bool
Smokes (Years)	0–37	Bool	STDs: Hepatitis B	0/1	Bool
Smokes (pack/Years)	0–37	Bool	STDs: HPV	0/1	Bool
Hormonal Contraceptives	0/1	Bool	STDs: Number of diagnosis	0–3	Int
Hormonal Contraceptives (Years)	0–30	Int	STDs: Time since first diagnosis	1–22	Int
IUD	0/1	Bool	STDs: Time since last diagnosis	1–22	Int
IUD (Years)	0–19	Int	Dx: Cancer	0/1	Bool
STDs	0/1	Bool	Dx:CIN	0/1	Bool
STDs (number)	0–4	Int	Dx:HPV	0/1	Bool
STDs: condylomatosis	0/1	Bool	Dx	0/1	Bool
STDs: cervical condylomatosis	0/1	Bool	Hinselmann: target variable	0/1	Bool
STDs: vaginal condylomatosis	0/1	Bool	Schiller: target variable	0/1	Bool
STDs: vulvo-perineal condylomatosis	0/1	Bool	Cytology: target variable	0/1	Bool
STDs: syphilis	0/1	Bool	Biopsy: class or target variable	0/1	Bool

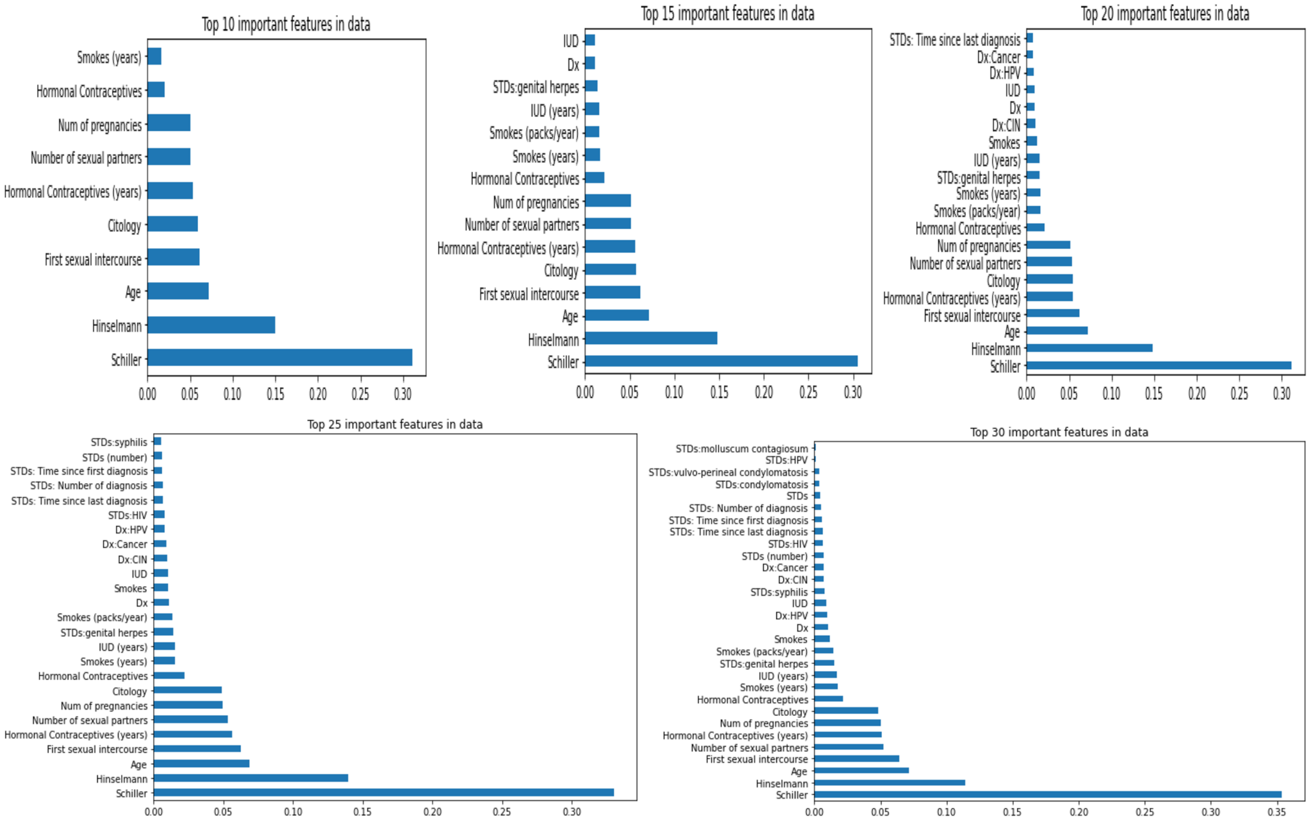

Fig. 2 Various top features

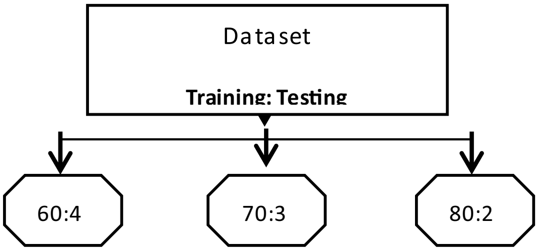

Fig. 3 Dataset splitting ratio

3.3.1 Chi-square and SelectKBest

The Chi-square statistic is used to compare two types of data: tests of independence and tests of goodness of fit. The Chi-square test of independence is used in feature selection to determine whether the class mark is independent of a feature [19]. For categorical features in a dataset, the Chi-square test is used. We measure the Chi-square between each feature and the target and choose the various top features with the highest Chi-square scores using SelectKBest model. It decides whether the sample's association between two categorical variables reflects their true association in the population. The successful application of this feature selection method has already found in plenty of research works [20]. The Chi-square score is calculated by Eq. (1):

$$X^2 = \frac{\sum \left(O_i - E_i\right)^2}{E_i} \tag{1}$$

where O_i is the observed value or no. of observations of class and E_i is the expected value or no. of expected observations of class if there was no relationship between the feature and the target.

3.3.2 Random Forest

Random Forest is a supervised learning algorithm that can be used for regression as well as classification. Because of their relative accuracy, robustness and ease of use, Random Forests are one of the most common machine learning methods. They also give two simple feature selection methods: mean decrease impurity and mean decrease accuracy [21]. The greatest appropriate suitable features are generated with a high score after applying the Random Forest algorithm to the dataset. These features received the highest score out of all the features in the dataset.

3.4 Data splitting

To determine the performance of machine learning algorithms, datasets are split into training and testing sets. During the dataset splitting process, the training set receives the majority of the data, while the testing set receives a smaller portion. During the training of the proposed model, the train dataset was used. By running the trained model on the test dataset, the success rates were determined. In this study, the dataset was split in various ways to compare the accuracy, precision, recall and f1-score values of various classification algorithms.

3.5 Classification algorithms

3.5.1 Logistic Regression (LR)

Logistic Regression is a machine learning classification algorithm used to assign observations to a discrete set of classes. Unlike linear regression, which produces continuous number values, Logistic Regression produces a probability value that can be mapped to two or more distinct groups using the logistic sigmoid equation. The function converts every real number into a number between 0 and 1. We use sigmoid to map predictions to probabilities in machine learning.

Equation of sigmoid function is given by Eq. (2):

$$f(x) = 1/\left(1 + e^{-(x)}\right) \tag{2}$$

where

- f(x) = output between 0 and 1 (probability estimate)
- x = input to the function (algorithm's prediction, e.g., mx + b)
- e = base of natural log

LR has a number of advantages, including the ability to have probabilities naturally and the ability to handle multi-class classification problems [22]. Another advantage is that most LR model analysis approaches are based on the same concepts as linear regression [23]. Logistic Regression was shown to be the most consistent with its accuracy scores.

3.5.2 Random Forest (RF)

Random Forest (RF) is a data mining technique for dealing with classification and regression issues. Growing an ensemble of trees and voting to determine the class category greatly improved classification accuracy. To develop these ensembles, random vectors are created. Each tree is made up of one or more random vectors. Classification and regression trees make up RF. The production of trees is analyzed to solve classification problems. The RF prediction is determined by a majority of class votes.

Although the generalization error is reduced to a limiting value, since over-fitting does not occur in large RFs, more trees are added to RF [24]. To achieve higher accuracy, low bias and correlation are important. Trees are grown without pruning to achieve low bias, and variables are randomly distributed at each node to achieve low correlation. After a tree has been developed, an out-of-bag (OOB) dataset is used to evaluate its performance. While more trees grow, RF uses the OOB dataset to measure an unbiased error estimate. In RF classification, the OOB dataset is often used to calculate the significance of variables [25].

3.5.3 K-nearest neighbor

The k-NN is a classification problem-solving supervised learning algorithm. The key point is to study the characteristics of each category in advance [26]. According to the k-NN algorithm used in the classification, the distance between the new individual and all previous individuals is considered, and the nearest k class is used, based on the attributes drawn from the classification level. As a result of this procedure, test data are assigned to the k-nearest neighbor group, which contains the most members in a given class. The research uses experiments to determine the best k number, and the Euclidean distance calculations method is used to calculate distance.

Euclidean calculation method is presented in Eq. 3 [27]

$$d\left(p_i, p_j\right) = \sqrt{\sum_{s=1}^{n} (p_{is} + p_{js})^2} \tag{3}$$

Where the p_i and p_j are different points belonging to the n-dimensional space in which the Euclidean process is applied for points distance calculation.

3.5.4 Multilayer perceptron (MLP)

It is an artificial neural network model that maps input sets to appropriate output sets using a feedforward method. A multilayer perceptron (MLP) is made up of several layers of nodes, each of which is connected to the one before it. Except for the input nodes, each node is a processing unit or a neuron with a nonlinear activation function. It uses a supervised learning technique as backpropagation that is used for training the network. The alteration of the standard linear perceptron, MLP is capable of distinguishing data which are not linearly separable [28].

3.5.5 Support vector machine

The support vector machine (SVM) is a supervised learning system and one of the kernel-based machine learning techniques. Vapnik at Bell Laboratories is primarily responsible for the creation of the SVM. SVM creates a high-dimensional space and then divides it based on the training details [29]. SVC was chosen for its ability to handle high-dimensional input [30] and specifies the kernel type to be used in the SVC algorithm. It must be one of ‘linear,’ ‘poly,’ ‘rbf,’ ‘sigmoid,’ ‘precomputed’ or a callable. If none is given, ‘rbf’ will be used. We applied kernel ‘linear.’

3.5.6 Decision tree

One of the classification methods is the decision tree [31], which classifies labeled trained data into a tree or rules. To test the accuracy of a classifier, test data are randomly selected from training data after the tree or rules are derived in the learning process. Unlabeled data are classified using the tree or rules learned during the learning phase after accuracy is checked. The structure of a decision tree is similar to the tree with a root node, a left subtree and right subtree. The leaf nodes in a tree represent a class label. The attribute splits depend on the impurity measures such as information gain, gain ratio, Gini index. The information gain is defined by Eq. (4):

$$\mathrm{Gain}(a,b) = \mathrm{Entrppy}(a) - \sum\nolimits_{v\in \mathrm{Values}(b)} \frac{|a_v|}{|a|}.\mathrm{Entropy}(a_v) \tag{4}$$

Following the construction of the tree, it is pruned to inspect for overfitting and noise. Finally, the tree has been optimized. The benefit of a tree-structured approach is that it is simple to understand and interpret, and it can handle both categorical and numerical attributes. It is also resistant to outliers and missing values. Decision tree classifiers are commonly used to diagnose diseases like breast cancer, ovarian cancer and heart sound diagnosis [32].

3.5.7 Gradient boosting classifiers (GBC)

Gradient boosting (GB) [33] generates new models sequentially from an ensemble of poor models with the aim of minimizing the loss function with each new model. The gradient descent method is used to calculate this loss function. Each new model fits more accurately with the observations when the loss function is used, and thus, the overall accuracy is increased. Boosting, on the other hand, must be stopped at some point; otherwise, the model would appear to over fit. A threshold for prediction accuracy or a maximum number of models produced may be used as a stopping criterion.

3.5.8 AdaBoost classifier

Adaptive boosting gives each training observation an equal weight at first. It uses a series of weak models and gives higher weights to observations that have been misclassified. The accuracy of the misclassified findings is increased since it uses several poor models, integrating the effects of the decision boundaries obtained during multiple iterations. Since it uses several poor models, integrating the outcomes of multiple iterations’ decision boundaries, the accuracy of the misclassified findings is increased, as is the accuracy of the overall iterations [34].

The weak models are evaluated using the error rate given in Eq. (5):

$$\varepsilon_t = pr_{i-D_t}\left[h_t(x_i) \neq y_i\right] = \sum_{i:h_t(x_i)\neq y_i} D_t \tag{5}$$

where ε_t is the weighted error estimate, pr_{i-D_t} is the probability of the random example i to the distribution D_t, h_t are the hypotheses of the weak learner, x_i is the training observation, y_i is the target variable, and t is the iteration number. The prediction error is one if the classification is wrong and 0 if the classification is correct.

3.6 Evaluation measures

The confusion matrix is a common method used to solve classification problems. It can be used to solve problems involving multiclass classification as well as binary classification. The confusion matrix is a N*N-dimensional matrix that describes the classification performance of a classifier in relation to some test data. Only if the true value for test data is defined can it be determined. Confusion matrices represent counts from predicted and actual values.

A number of evaluation criteria are listed below:

Accuracy: Accuracy is the most immersive performance measure. It is the ratio of correctly expected observation to total observation. The sum of true positive and true negative is divided by the total number of subjects in the sample to determine accuracy. An Eq. (6) is used to calculate it:

$$\mathrm{Accuracy} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}} \times 100 \tag{6}$$

Recall: It refers to the proportion of system-generated results that correctly predicted positive observations compared to all actual positives. It is also known as sensitivity. An Eq. (7) is used to calculate it:

$$\text{Recall} = \frac{TP}{TP + FN} \times 100 \tag{7}$$

Precision: It specifies the proportion of system-generated positive observations that are correctly predicted compared to the total number of predicted positive observations. An Eq. (8) is used to calculate it:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \times 100 \tag{8}$$

F1-Score: F1-score is defined by the weighted average of precision and recall. Hence, F1-score takes both false positive (FP) and false negative (FN) into account to convey the balance between recall and precision. An Eq. (9) is used to calculate it:

$$\text{F1} - \text{score} = \frac{2 * \text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}} \times 100 \tag{9}$$

4 Result and analysis

This section discusses the results of the experimental analysis and the detection of cervical cancer disease. For the purpose of analysis, various classification algorithms were used, and the result showed that the outcome was dependent on a number of various suitable features from the dataset.

4.1 Experimental setup

The proposed method is evaluated on a system with 8 GB of RAM and a 3.0 GHz Intel Core i-7 processor using the cloud-based web application environment named Google Colab was used to create the model and use classification algorithms to detect cervical cancer disease on the dataset. The Google Colab provides free Jupyter notebook environment without setup requirements. Moreover, this ensures simultaneous working environment and frequently used-required machine learning libraries [35]. Various suitable features chosen for the detection of cervical cancer disease from 36 attributes in the dataset using the combination of various feature selection technique, and target attributes named biopsy were used to classify the dataset as healthy or cancer results for cervical cancer diagnostics.

4.2 Result evaluation

The results and accuracies we hope to achieve by using various classification algorithms are shown here. The classification algorithm's results and performance are summarized in the various tables, and figures are listed. Using Chi-square, SelectBest and Random Forest algorithms is to select various top features with a higher score for detecting cervical cancer disease.

4.2.1 Analysis of model performance on various top features

On top 10 features, MLP claims best accuracy 97.40% and f1-score of 97.20%. The RF classifier has the highest precision and recall, while the DT classifier has the lowest

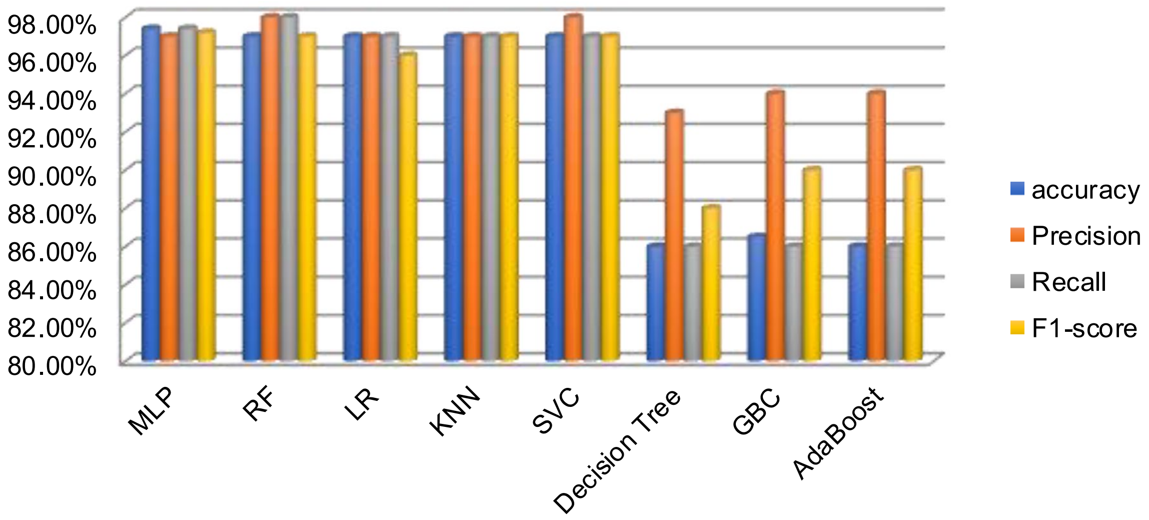

Fig. 4 Highest performance comparison of different classification algorithms on top 10 features

accuracy reporting because decision tree is a technique of machine learning, while MLP is exclusive to deep learning. The MLP is made up of a network of processing units that resemble neurons. For these top 10 features, MLP performs better than any other classifiers. Highest performance comparison of different classification algorithms on top 10 features is shown in Fig. 4. Various performance comparison of different classification algorithms on top 10 features based on various dataset splitting ratio is shown in Table 2.

On top 15 features, MLP and LR each claim best accuracy 98.00% and recall 98.00%. The MLP classifier has the highest precision and recall, while the AdaBoost classifier has the lowest accuracy reporting. For these top 15

Table 2 Performance comparison of different classification algorithms on top 10 features

Model	Training data and testing data ratio	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
MLP	70:30	97.40	97.00	97.40	97.20
	60:40	97.00	97.00	97.00	97.00
	80:20	97.00	97.00	97.00	97.00
RF	70:30	97.00	98.00	98.00	97.00
	60:40	96.00	97.00	96.00	96.00
	80:20	95.00	96.00	95.00	96.00
LR	70:30	96.80	92.00	96.80	94.00
	60:40	97.00	97.00	97.00	96.00
	80:20	96.60	96.00	96.60	96.00
KNN	70:30	96.80	94.00	96.80	95.00
	60:40	96.00	96.00	96.00	96.00
	80:20	97.00	97.00	97.00	97.00
SVC	70:30	97.00	98.00	97.00	97.00
	60:40	97.00	97.00	97.00	97.00
	80:20	97.00	97.00	97.00	97.00
Decision Tree	70:30	86.00	93.00	86.00	88.00
	60:40	84.00	94.00	84.00	88.00
	80:20	81.00	91.00	81.00	95.00
GBC	70:30	86.50	94.00	86.00	90.00
	60:40	85.00	94.00	85.00	88.00
	80:20	86.00	94.00	86.00	89.00
AdaBoost	70:30	86.00	94.00	86.00	90.00
	60:40	84.00	93.00	84.00	88.00
	80:20	86.00	94.00	86.00	90.00

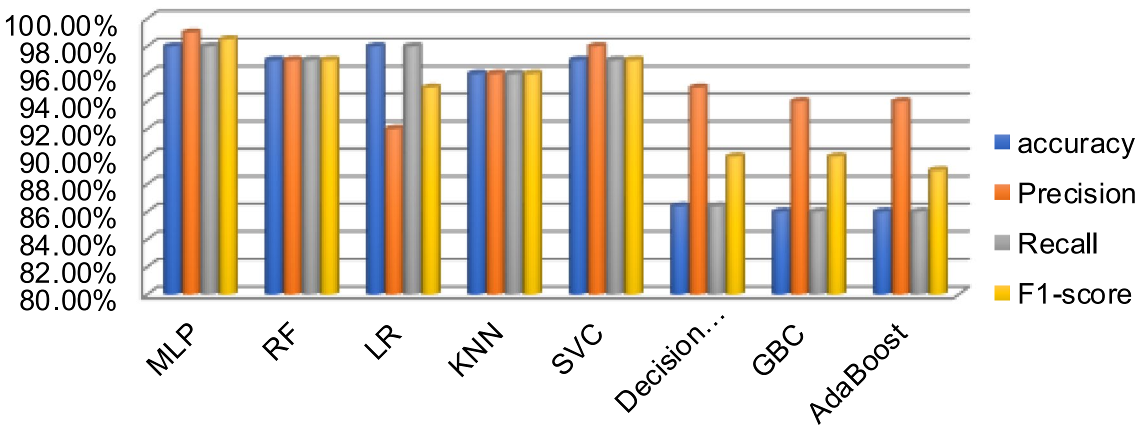

Fig. 5 Highest performance comparison of different classification algorithms on top 15 features

Table 3 Performance comparison of different classification algorithms on top 15 features

Model	Training data and testing data ratio	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
MLP	70:30	98.00	99.00	98.00	98.50
	60:40	97.00	97.00	97.00	97.00
	80:20	97.00	97.00	97.00	97.00
RF	70:30	97.00	97.00	97.00	97.00
	60:40	96.80	97.00	96.80	96.00
	80:20	96.00	96.00	96.00	96.00
LR	70:30	98.00	92.00	98.00	95.00
	60:40	97.00	97.00	97.00	97.00
	80:20	96.00	95.00	96.00	95.50
KNN	70:30	95.40	94.00	95.40	94.50
	60:40	96.00	96.00	96.00	96.00
	80:20	96.00	95.00	96.00	95.00
SVC	70:30	97.00	98.00	97.00	97.00
	60:40	97.00	97.00	97.00	97.00
	80:20	97.00	97.00	97.00	97.00
Decision Tree	70:30	86.40	95.00	86.40	90.00
	60:40	86.00	95.00	86.00	89.00
	80:20	85.00	96.00	81.00	89.00
GBC	70:30	86.00	94.00	86.00	90.00
	60:40	85.00	94.00	85.00	88.00
	80:20	83.00	93.00	8300	87.00
AdaBoost	70:30	86.00	94.00	86.00	89.00
	60:40	86.00	94.00	86.00	89.00
	80:20	82.00	92.00	82.00	86.00

features, MLP and LR each perform better than other classifiers. Highest performance comparison of different classification algorithms on top 15 features is shown in Fig. 5. Performance comparison of different classification algorithms on top 15 features based on various dataset splitting ratio is shown in Table 3.

On top 20 features, MLP and RF each claim best accuracy over 97.00%. The MLP classifier has the highest precision and recall, while the GBC and AdaBoost have the lowest accuracy reporting. For these top 20 features, MLP performs better than other classifier. Highest performance comparison of different classification algorithms on top

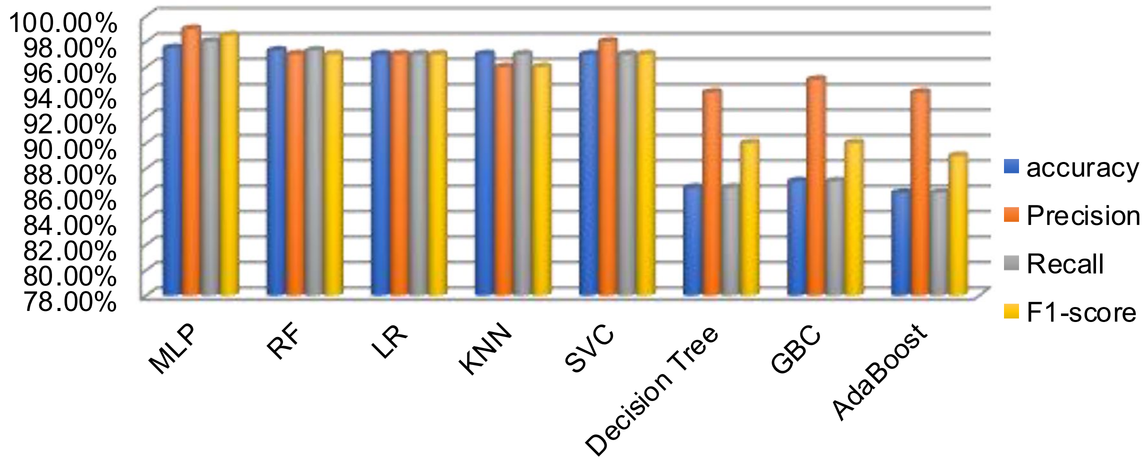

Fig. 6 Highest performance comparison of different classification algorithms on top 20 features

Table 4 Performance comparison of different classification algorithms on top 20 features

Model	Training data and testing data ratio	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
MLP	70:30	97.50	99.00	98.00	98.50
	60:40	97.00	97.00	97.00	97.00
	80:20	97.00	97.00	97.00	97.00
RF	70:30	97.30	97.00	97.30	97.00
	60:40	96.00	96.00	96.00	96.00
	80:20	96.00	96.00	96.00	96.00
LR	70:30	97.00	96.00	97.00	96.00
	60:40	97.00	97.00	97.00	97.00
	80:20	97.00	96.00	97.00	96.00
KNN	70:30	97.00	96.00	97.00	96.00
	60:40	95.00	95.00	95.00	95.00
	80:20	95.00	95.00	95.00	95.00
SVC	70:30	97.00	98.00	97.00	97.00
	60:40	97.00	97.00	97.00	97.00
	80:20	97.00	97.00	97.00	97.00
Decision Tree	70:30	86.50	94.00	86.50	90.00
	60:40	84.00	94.00	84.00	88.00
	80:20	84.00	95.00	84.00	88.00
GBC	70:30	87.00	95.00	87.00	90.00
	60:40	85.00	94.00	85.00	88.00
	80:20	83.00	93.00	8300	87.00
AdaBoost	70:30	86.10	94.00	86.10	89.00
	60:40	85.00	94.00	85.00	89.00
	80:20	83.00	93.00	83.00	87.00

20 features is shown in Fig. 6. Performance comparison of different classification algorithms on top 20 features based on various dataset splitting ratio is shown in Table 4.

On top 25 features, MLP claims best accuracy 98.00%. The best precision giving classifiers are MLP and SVC as 98.00%. The lowest accuracy reporting classifiers are GBC and AdaBoost. For these top 25 features, MLP performs better than other classifier. Highest performance comparison of different classification algorithms on top 25 features is shown in Fig. 7. Performance comparison of different classification algorithms on top 25 features based on various dataset splitting ratio is shown in Table 5.

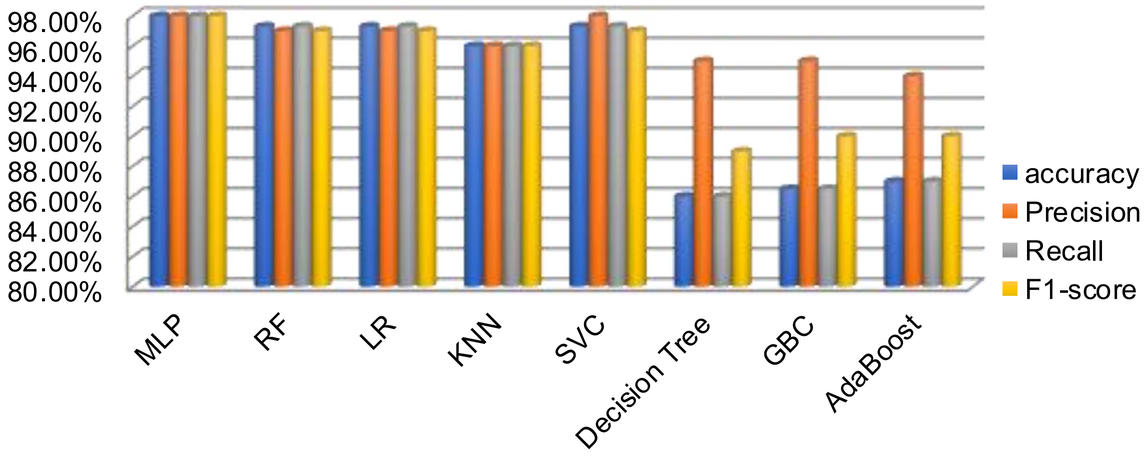

Fig. 7 Highest performance comparison of different classification algorithms on top 25 features

Table 5 Performance comparison of different classification algorithms on top 25 features

Model	Training data and testing data ratio	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
MLP	70:30	98.00	98.00	98.00	98.00
	60:40	97.00	97.00	97.00	97.00
	80:20	97.00	96.00	97.00	96.00
RF	70:30	97.30	97.00	97.30	97.00
	60:40	96.00	96.00	96.00	96.00
	80:20	96.60	96.00	96.60	96.00
LR	70:30	97.30	97.00	97.30	97.00
	60:40	97.00	97.00	97.00	97.00
	80:20	96.00	95.00	96.00	95.00
KNN	70:30	96.00	96.00	96.00	96.00
	60:40	95.00	95.00	95.00	95.00
	80:20	95.00	94.00	95.00	95.00
SVC	70:30	97.30	98.00	97.30	97.00
	60:40	97.00	97.00	97.00	97.00
	80:20	97.00	97.00	97.00	97.00
Decision Tree	70:30	86.00	95.00	86.00	89.00
	60:40	84.00	94.00	84.00	88.00
	80:20	84.00	95.00	84.00	88.00
GBC	70:30	86.50	95.00	86.50	90.00
	60:40	85.00	94.00	85.00	88.00
	80:20	83.00	93.00	8300	87.00
AdaBoost	70:30	87.00	94.00	87.00	90.00
	60:40	86.00	94.00	86.00	89.00
	80:20	83.00	93.00	83.00	87.00

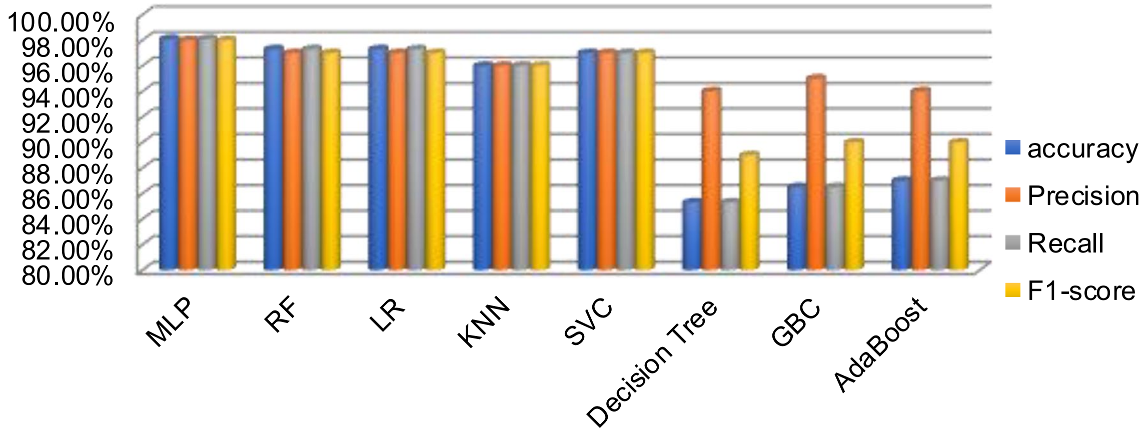

Fig. 8 Highest performance comparison of different classification algorithms on top 30 features

On top 30 features, MLP claims best accuracy 98.10%. The MLP classifier has the highest precision, while the GBC and AdaBoost have the lowest accuracy reporting. For these top 30 features, MLP performs better than any other classifiers. Highest performance comparison of different classification algorithms on top 30 features is shown in Fig. 8. Performance comparison of different classification algorithms on top 30 features based on various dataset splitting ratio is shown in Table 6.

Table 6 Performance comparison of different classification algorithms on top 30 features

Model	Training data and testing data ratio	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
MLP	70:30	98.10	98.00	98.10	98.00
	60:40	97.00	97.00	97.00	97.00
	80:20	96.00	95.00	96.00	95.00
RF	70:30	97.30	97.00	97.30	97.00
	60:40	96.00	96.00	96.00	96.00
	80:20	96.50	96.00	96.50	96.00
LR	70:30	97.30	97.00	97.30	97.00
	60:40	97.00	97.00	97.00	97.00
	80:20	96.00	95.00	96.00	95.00
KNN	70:30	96.00	96.00	96.00	96.00
	60:40	95.00	95.00	95.00	95.00
	80:20	95.00	95.00	95.00	95.00
SVC	70:30	97.00	97.00	97.00	97.00
	60:40	96.00	96.00	96.00	96.00
	80:20	97.00	97.00	97.00	97.00
Decision Tree	70:30	85.30	94.00	85.30	89.00
	60:40	84.00	93.00	84.00	88.00
	80:20	84.00	95.00	84.00	88.00
GBC	70:30	86.50	95.00	86.50	90.00
	60:40	85.00	94.00	85.00	88.00
	80:20	83.00	93.00	8300	87.00
AdaBoost	70:30	87.00	94.00	87.00	90.00
	60:40	86.00	94.00	86.00	89.00
	80:20	83.00	93.00	83.00	87.00

5 Discussion

The study of the obtained results using different classification algorithms supports the diagnosis of patients with cervical cancer disease. A combination of feature selection methods was used in this study for feature selection. Various suitable features/attributes with high scores among all features in the dataset have been chosen. Following that, eight different classification algorithms were used to detect cervical cancer disease at an early stage, including Multilayer Perceptron (MLP), Logistic Regression, Random Forest, K-nearest Neighbors, Decision tree, SVC, Gradient Boosting and AdaBoost. Among different dataset splitting ratios and various top features, the better accuracy rate obtained from Multilayer Perceptron (MLP) was 98.10% on top 30 features, Logistic Regression was 98.00% on top 15 features. On top 25 features Random Forest and Support Vector Machine each claim 97.30% and AdaBoost was 87.00%, On top 20 features K-Nearest Neighbors claim 97.00% and Gradient Boosting claims 87.00% and decision tree claims 86.50%. The study deals with various selections of top features sets, where previous related papers focused on only one selection of top features. All previous related works deal with only one dataset splitting ratio; on the other hand, this paper deals with three different dataset splitting ratios to measure performances variation of the model. In this paper [12] their model claims the highest accuracy is 87.21% on only one top features set; on the other hand, this model MLP claims up to 98.00% accuracy on various top features sets. The overall highest and lowest accuracy comparison of different classification algorithms is shown in Figs. 9 and 10. Also overall highest and lowest performance comparison of different classification algorithms is shown in Tables 7 and 8. Highest accuracy comparison on top suitable features of classification algorithms is shown in Fig. 11. Highest accuracy comparison on various top features of classification algorithms is shown in Fig. 12.

6 Conclusion

Nowadays, cervical cancer is a widespread disease and screening also entails lengthy clinical tests. The aim of the research is to develop a model that can accurately diagnose and analyze risk factors for cervical cancer data using classification algorithms such as Multilayer Perceptron

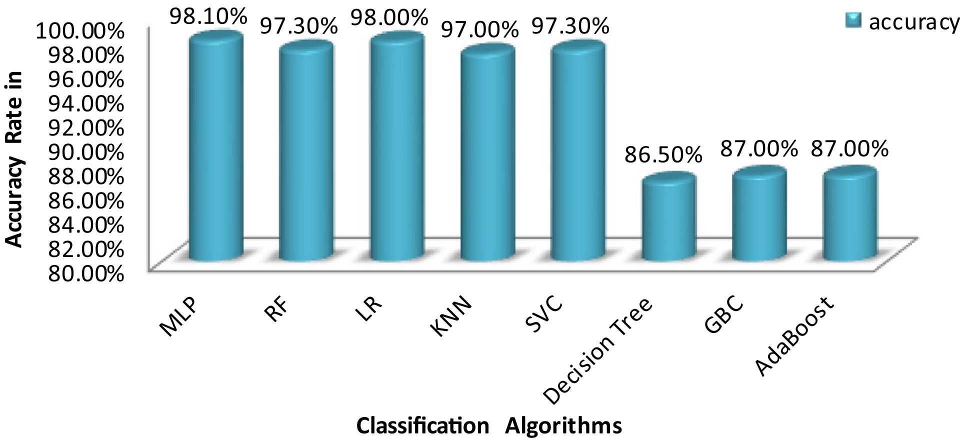

Fig. 9 Highest accuracy comparison of different classification algorithms

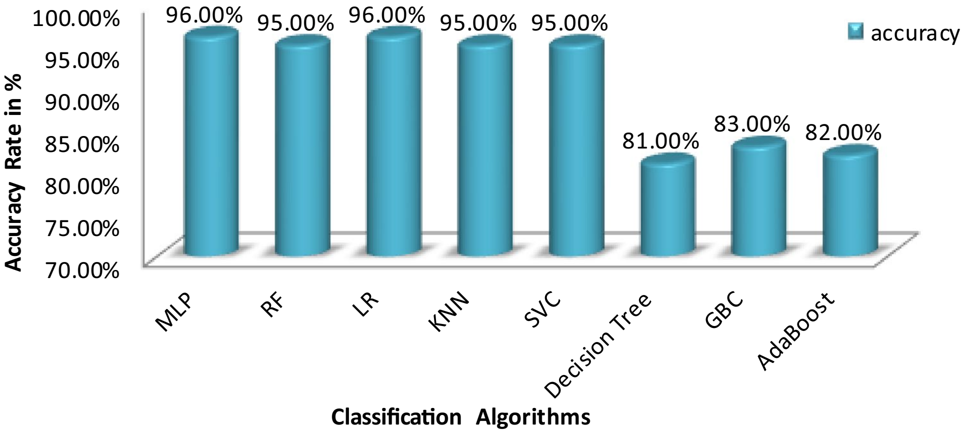

Fig. 10 Lowest accuracy comparison of different classification algorithms

Table 7 Highest performance comparison of different classification algorithms

Model	Training data and testing data ratio	No. of top feature	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
MLP	70:30	30	98.10	98.00	98.10	98.00
RF	70:30	25	97.30	97.00	97.30	97.00
LR	70:30	15	98.00	92.00	98.00	97.00
KNN	70:30	20	97.00	96.00	97.00	96.00
SVC	70:30	25	97.30	98.00	97.300	97.00
Decision Tree	70:30	20	86.50	94.00	86.50	90.00
GBC	70:30	20	87.00	95.00	87.00	90.00
AdaBoost	70:30	25	87.00	94.00	87.00	90.00

(MLP), Random Forest and k-Nearest Neighbor, Decision Tree, Logistic Regression, SVC, Gradient Boosting and AdaBoost. A variety of approaches are used to prevent missing values. A combination of feature selection techniques, such as Chi-square, SelectBest and Random Forest, was used to pick different top suitable features. Accuracy, recall, precision and f1-score parameters are used to analyze those classifications performance. As a result, when compared to the other eight classification algorithms, MLP has the highest rate of accuracy, precision, recall and f1-score, while Decision Tree has the lowest rate of accuracy, precision, recall and f1-score. Classification models claim highest accuracy on specific top features such as Multilayer Perceptron (MLP) was 98.10% on

Table 8 Lowest performance comparison of different classification algorithms

Model	Training data and testing data ratio	No. of top feature	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
MLP	80:20	30	96.00	95.00	96.00	95.00
RF	80:20	10	95.00	96.00	95.00	96.00
LR	80:20	30	96.00	95.00	96.00	95.00
KNN	80:20	25	95.00	94.00	95.00	95.00
SVC	60:40	30	95.00	96.00	95.00	96.00
Decision Tree	80–20	10	81.00	91.00	81.50	86.00
GBC	80–20	30	83.00	93.00	83.00	87.00
AdaBoost	70:30	15	82.00	92.00	82.00	86.00

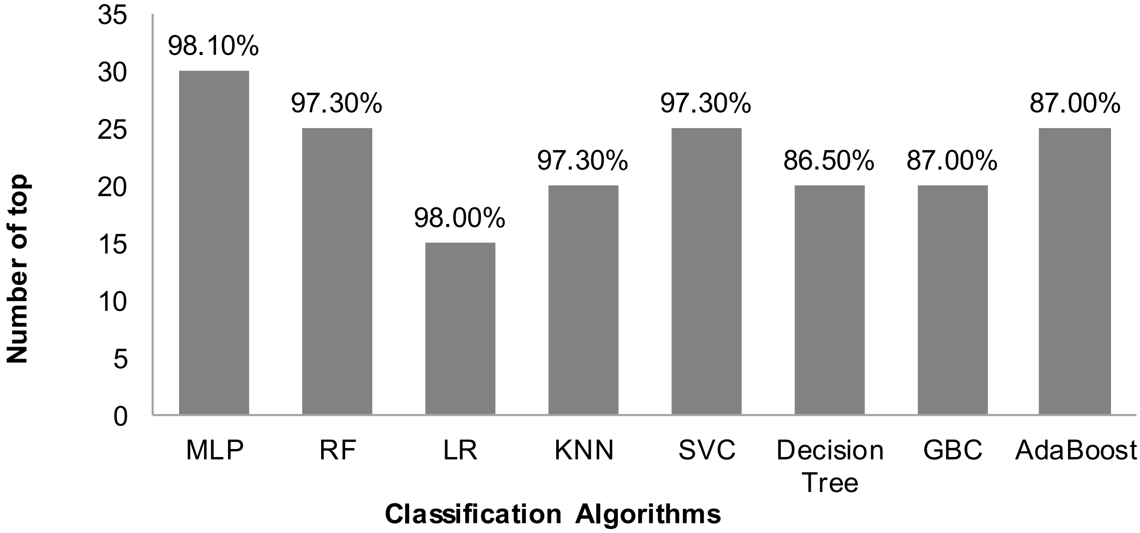

Fig. 11 Highest accuracy comparison on top suitable features of different classification algorithms

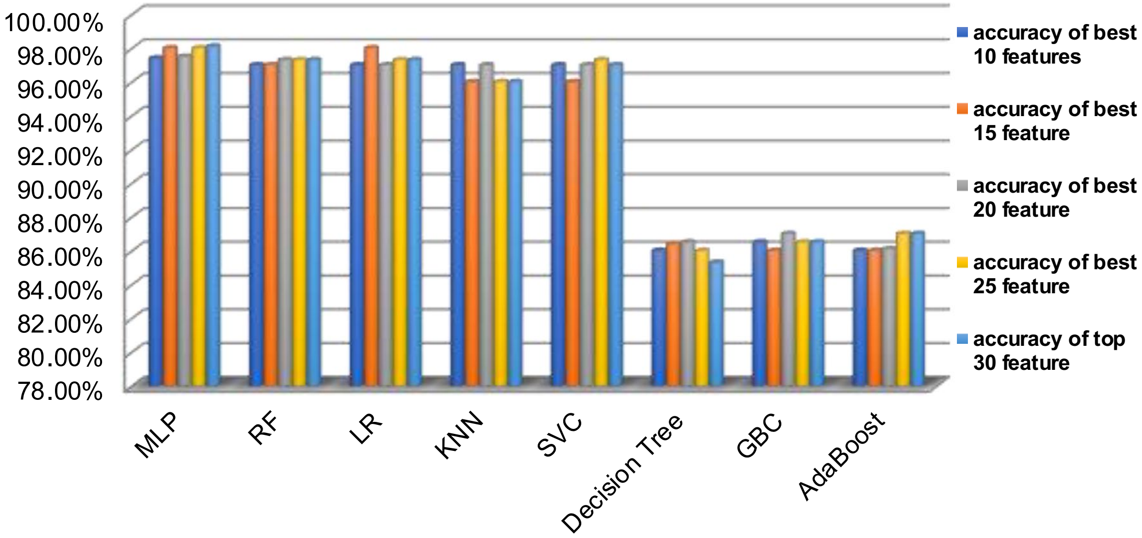

Fig. 12 Highest accuracy comparison on various top features of different classification algorithms

top 30 features; Logistic Regression was 98.00% on top 15 features. These models are reasonably accurate, and their accuracies are similar. On average, MLP proved to be better than others classification model on various top features. On the other hand, most of the classification models claim the highest accuracy on the top 25 features in dataset splitting ratio (70:30). After comparing our findings to those of many previous studies, we discovered that our models were more effective at diagnosing cervical cancer based on certain evaluation criteria.

Acknowledgements The authors are thankful to those who have participated in this research work

Declarations

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. El-Nashar, Manal Ahmed, Rawan Yasseen Bamjboor, Ammar Mansour, and Banan Aied Althobaity. Awareness of the women about the vaginal infection as a risk factor for cervical cancer in Taif city, Saudi Arabia
2. Silvia de Sanjosé, Beatriz Serrano, Sara Tous, Maria Alejo, Belén Lloveras, Beatriz Quirós, Omar Clavero, August Vidal, Carla Ferrándiz-Pulido, Miquel Ángel Pavón, Dana Holzinger, Gordana Halec, Massimo Tommasino, Wim Quint, Michael Pawlita, Nubia Muñoz, Francesc Xavier Bosch, Laia Alemany, (2018) RIS HPV TT, VVAP and Head and Neck study groups, Burden of Human Papillomavirus (HPV)-Related Cancers Attributable to HPVs 6/11/16/18/31/33/45/52 and 58. JNCI Cancer Spectrum, 2(4): pky045, doi: https://doi.org/10.1093/jncics/pky045
3. Cervical-cancer, https://www.webmd.com/cancer/cervical-cancer [Access Date: 3/16/2021]
4. Cervical-cancer symptoms, https://www.healthline.com/health/cervical-cancer#symptoms [Access Date: 3/16/2021]
5. Early-detection,https://www.who.int/bangladesh/news/detail/10-11-2020-who-supports-early-detection-and-control-of-cervical-and-breast-cancer-in-bangladesh [Access Date: 3/17/2021]
6. Banik R, Naher S, Rahman M et al (2020) Investigating Bangladeshi rural women's awareness and knowledge of cervical cancer and attitude towards HPV vaccination: a community-based cross-sectional analysis. J Canc Educ. https://doi.org/10.1007/s13187-020-01835-w
7. Bangladesh Human Papillomavirus and Related Cancers, Fact Sheet 2018 (2019–06–17), https://hpvcentre.net/statistics/reports/ BGD_FS.pdf [Access Date: 3/17/2021]
8. BangladeshHuman Papillomavirus and Related Cancers, Fact Sheet 2018, https://hpvcentre.net/statistics/reports/BGD.pdf [Access Date: 3/17/2021]
9. USA_FS, https://hpvcentre.net/statistics/reports/USA_FS.pdf [Access Date: 3/18/2021]
10. Alsmariy R, G Healy, and H Abdelhafez. (2020) Predicting cervical cancer using machine learning methods. IJACSA thesia.org
11. Lu J et al (2020) Machine learning for assisting cervical cancer diagnosis: an ensemble approach. Future Gener Comput Syst 106:199–205
12. Ahishakiye E et al. (2020) Prediction of cervical cancer basing on risk factors using ensemble learning. In: 2020 IST-Africa conference (IST-Africa). IEEE
13. Unlersen MF, Sabanci K, Özcan M (2017) Determining cervical cancer possibility by using machine learning methods. Int J Latest Res Eng Technol 3(12):65–71
14. Wu W, Zhou H (2017) Data-driven diagnosis of cervical cancer with support vector machine-based approaches. IEEE Access 5:25189–25195. https://doi.org/10.1109/ACCESS.2017.2763984
15. Singh SK, Goyal A (2020) Performance analysis of machine learning algorithms for cervical cancer detection. Int J Healthcare Inf Syst Inf (IJHISI) 15(2):1–21
16. K. Fernandes, J. S. Cardoso, and J. Fernandes, (2017) Transfer learning with partial observability applied to cervical cancer screening. In: Iberian conference on pattern recognition and image analysis, LNCS. Springer International Publishing, vol. 10255, pp. 243–250
17. Data Preprocessing basic steps, https://medium.datadriveninvestor.com/data-preprocessing-for-machine-learning-188e9eef1d2c [Access Date:3/20/2021]
18. Verma AK, Pal S, Kumar S (2019) Comparison of skin disease prediction by feature selection using ensemble data mining techniques. Inf Med Unlock 16:100202
19. Rachburee N, and W Punlumjeak (2015) A comparison of feature selection approach between greedy, IG-ratio, Chi-square, and mRMR in educational mining. In: 2015 7th international conference on information technology and electrical engineering (ICITEE), IEEE
20. Ijaz MF, Attique M, Son Y (2020) Data-driven cervical cancer prediction model with outlier detection and over-sampling methods. Sensors 20(10):2809
21. Han H., X Guo, and H Yu. (2016) Variable selection using mean decrease accuracy and mean decrease gini based on random forest. In: 2016 7th IEEE international conference on software engineering and service science (icsess). IEEE
22. P. Karsmakers, K. Pelckmans, and J. A. K. Suykens (2007) Multi-class kernel logistic regression: a fixed-size implementation. In: international joint conference on neural networks, pp. 1756–1761
23. Hosmer DW, Lemeshow S (2000) Applied logistic regression, 2nd edn. Wiley, London
24. Breiman L (2001) Random forests. Mach Learn 45(1):5–32
25. Verikas A, Gelzinis A, Bacauskiene M (2011) Mining data with random forests: a survey and results of new tests. Pattern Recognit 44(2):330–349
26. Wang J, Neskovic P, Cooper LN (2007) Improving nearest neighbor rule with a simple adaptive distance measure. Pattern Recogn Lett 28(2):7

27. Zhou Y, Li Y, Xia S (2009) An improved KNN text classification algorithm based on clustering. J Comput 4(3):8
28. Gardner MW, Dorling SR (1998) Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. Atmos Environ 32(14):2627–2636
29. Ashok B, Aruna P (2016) Comparison of feature selection methods for diagnosis of cervical cancer using SVM classifier. Int J Eng Res Appl 6:94–99
30. Szlosek, Donald A., and Jonathan Ferrett. Using machine learning and natural language processing algorithms to automate the evaluation of clinical decision support in electronic medical record systems. eGEMs 4.3 (2016)
31. Han J, Kamber M (2000) Data mining; concepts and techniques. Morgan Kaufmann Publishers, United States
32. Stasis, A.C. Loukis, E.N. Pavlopoulos, S.A. Koutsouris, D. (2003) Using decision tree algorithms as a basis for a heart sound diagnosis decision support system. In: 2003 4th International IEEE EMBS Special Topic Conference Information Technology Applications in Biomedicine
33. Friedman JH (2002) Stochastic gradient boosting. Comput Stat Data Anal 38:367–378
34. Rahman S et al (2020) Performance analysis of boosting classifiers in recognizing activities of daily living. Int J Env Res Public Health 17(3):1082
35. Google Colab, https://www.tutorialspoint.com/google_colab/what_is_google_colab.htm [Access Date: 3/20/2021]

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.