I.J. Intelligent Systems and Applications, **2021, 5, 46-63**
Published Online October 2021 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijisa.2021.05.05

A Multi-Stage Approach Combining Feature Selection with Machine Learning Techniques for Higher Prediction Reliability and Accuracy in Cervical Cancer Diagnosis

Avijit Kumar Chaudhuri
Department of Computer Application, SEACOM SKILLS UNIVERSITY, Kendradangal, Bolpur, Birbhum, 731 236, West Bengal, India
E-mail: c.avijit@gmail.com

Arkadip Ray
Department of Information Technology, Government College of Engineering and Ceramic Technology, Kolkata, West Bengal, 700010, India
E-mail: arka1dip2ray3@gmail.com

Dilip K. Banerjee
Department of Computer Application, SEACOM SKILLS UNIVERSITY, Kendradangal, Bolpur, Birbhum, 731236, West Bengal, India
E-mail: dkbanrg@gmail.com

Anirban Das
University of Engineering & Management, Kolkata, West Bengal, India
E-mail: anirban-das@live.com

Received: 07 July 2021; Revised: 09 August 2021; Accepted: 23 August 2021; Published: 08 October 2021

Abstract: Cervical cancer is the fourth most prevalent cancer in women which has claimed 3,41,831 lives and accounted for 6,04,127 new cases in 2020 worldwide. To reduce such a vast mortality rate, early detection of the disease is essential. A fast, accurate, and interpretable machine learning model is a research subject. Fewer features reduce the computational effort and improve interpretation. A 3-Stage Hybrid feature selection approach and a Stacked Classification model are evaluated on the cervical cancer dataset obtained from the UCI Machine Learning Repository with 35 features and one outcome variable. Stage-1 uses a Genetic Algorithm and Logistic Regression Architecture for Feature Selection and selects twelve features well correlated with the class but not among themselves. Stage-2 utilizes the same Genetic Algorithm and Logistic Regression Architecture for Feature Selection to select five features. In Stage-3, Logistic Regression (LR), Naïve Bayes (NB), Support Vector Machine (SVM), Extra Trees (ET), Random Forest (RF), and Gradient Boosting (GDB) are used with the five features to identify patients with or without cancer. Data splitting, several metrics, and statistical tests are used, along with 10-fold cross validation, to do a comparative analysis. LR, NB, SVM, ET, RF, and GDB demonstrate improvement across performance measures by reducing the number of features to five. In the 66-34 split, all five machine learning methods except NB recorded 97% accuracy with 5 features. Also, the Stacked model produced higher than 96% accuracy with five features in 66-34 and 80-20 splits, and in 10-fold cross validation. Various performance aggregators have shown improved results with reduced features when compared to previous studies. Finally, with approximately 100% performance in classification results, the suggested ensemble model showed its promise. The output results were compared to those of other studies on the same dataset, and the proposed classifiers were found to be the most effective across all performance dimensions.

Index Terms: Cervical Cancer, Feature Selection, Genetic Algorithm (GA), Logistic Regression (LR), Gradient Boosting (GDB).

1. Introduction

Cancer is the second leading cause of death globally, accounting for an estimated 9,958,133 deaths with new cases amounting to 19,292,789 in 2020 [1]. Health systems in low- and middle-income countries are least equipped to cope with the cancer burden, so much so, in 2017, only 26% of developing countries informed having screening services available for the populace compared to 90% of developed countries having treatment services. The anticipated cancer frequencies will reach up to 22 million in 2030 [2,3]. In India, new cancer incidence in 2020 is 6.9% (1,324,413 cases) of the world tally. Women are more likely to be diagnosed with cancer than men amongst the Indian population [4]. Cervical cancer is one type of gynecological cancer that reports the fourth highest cancer incidence (6.5%) and mortality (7.7%) in women globally [1]. Flouting world trend, cervical cancer in India stands second in terms of incidence (18.3%) and mortality (18.7%) out of total female patients suffering from different cancers like breast, ovary, esophagus, lung, oral, stomach, etc.

Cervical cancer is a malicious tumor that occurs in the cells lining the cervix (opening of the uterus to the vagina or, birth canal) due to abnormal cell growth and reproduction without controlled cell division and cell death. Early-stage symptoms of cervical cancer include increased vaginal discharge and postmenopausal bleeding. As cervical cancer progresses, severe symptoms like lasting pelvic pain, loss of appetite, weight loss, fatigue, swelling in legs, blockage of the urinary bladder, kidney failure, etc. may arise. Cervical cancer is mainly of two types, squamous cell carcinoma, and adenocarcinoma [5]. A staging system for cervical and other cancers of the female reproductive organs (staging from I to IVb) is provided by the International Federation of Gynecology and Obstetrics (FIGO) [6]. Cervical cancer takes 15 to 20 years in women with usual immune systems to advance from precancerous lesion (cervical intraepithelial neoplasia - CIN) in normal cells to invasive malignancy.

Lion's share of cervical cancer cases (99%) is associated with the infection of human papillomavirus (HPV), a widespread virus transmitted through sexual intercourse. Globally, 70–80% of cervical cancers are attributed to HPV (mainly genotypes HPV-16 and HPV-18) and in India, HPV prevalence is 88–97% among women with cervical cancer. Within a few months after getting infected with HPV, it heals spontaneously, and about 90% of patients prevent cervical cancer progression by early detection and treatment of precancerous lesions within 2 years. HPV is a class of viruses that typically infect the reproductive tract of sexually active men and women [7]. The body's immune system which is resistant to HPV tends to inhibit this infected women's virus attack, but for a trifling fraction of women, the virus has a long-life cycle before cells on the cervical surface become cancer cells [7].

It remains imprecise whether the stated risk factors are stand-alone or, whether they are subjected to act as cofactors to HPV infection for the inception of cervical cancer [8-15]. The equivocal relationship between patients' risk factors and multifaceted screenings requiring prominent pathological and medical expertise can lead to delayed and prolong treatment processes in a country like India, thus causing increased fatalities. In the era of high-performance computing and artificial intelligence, screening tests can be automated by data mining and machine learning systems to comprehend patterns in clinical datasets for examining the risk factors of cervical cancer. Efficacy of classification, clustering, and prediction through various optimization, statistical and probabilistic techniques is essential for data-driven clinical diagnosis. Identifying women with a higher risk of developing cervical carcinoma will help to avail themselves of preferential cancer screening and medical treatment straight away. Testing and early diagnosis of cancer contribute to a reduction in the peril of perishing cancer sufferers.

In this study, several machine learning algorithms are instigated to shorten the diagnostic period and streamline the assessment process for physicians. This paper addresses the critical questions – (1) What are the crucial features that explain the cause and effect of cervical cancer? and (2) Which data mining tool yields higher accuracy.

The authors propose a hybrid feature selection and stacked generalization model (HFSSGM) approach to sequentially determining the significant features, predict using different proven data mining methods and simulate with different train-test partitions as depicted in Fig. 1. The features are determined using Genetic algorithm (GA) and Logistic Regression (LR) and the dataset is revised keeping the significant features. The Genetic algorithm and Logistic Regression are applied to the revised dataset and further refined. The iteration continues till the accuracy of prediction using LR increases. The final dataset with fewer features is then split to create the train-test sets and subject to well-proven data mining techniques (DMTs), namely, Logistic Regression (LR), Naïve Bayes (NB), Support Vector Machine (SVM), Extra Trees (ET), Random Forest (RF), Gradient Boosting (GDB). The step is repeated with a different train-test ratio and prediction using the DMTs. This iteration continues till the accuracy, specificity, sensitivity, and test result improve.

Furthermore, the proposed classifier has been compared to RF, NB, SVM with radial basis function kernel, ET, and GDB, which are all state-of-the-art machine-learning classifiers. Different performance metrics were used to test the classifiers, including accuracy, sensitivity, specificity, receiver operating characteristic, the area under the curve, and statistical tests like kappa statistics. To test the credibility of the classification models in handling unbalanced data, this study used various splits of training and testing data, including 50%–50%, 66%–34%, 80%–20%, and 10-fold cross-validation. The architecture for the proposed classifier for cervical cancer prediction has been depicted in Fig. 2 below.

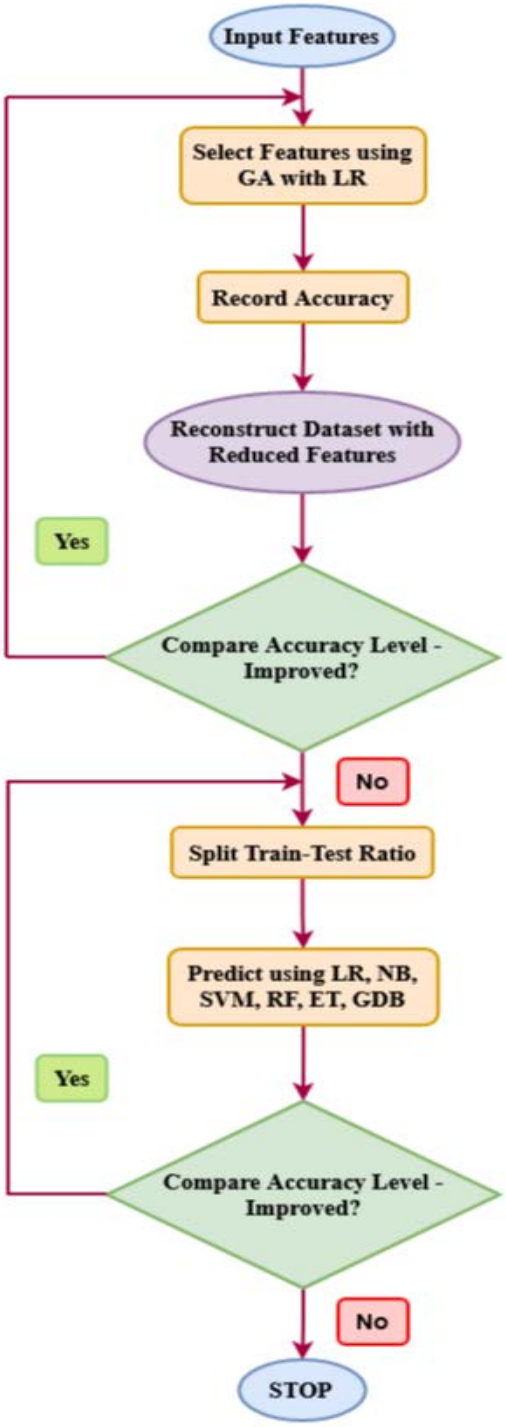

Fig.1. Hybrid Feature Selection and Stacked Generalization Model (HFSSGM) Algorithm

2. Relevant Literature

The cervix, the uterus, the vagina, and the ovaries make up the female reproductive system. The opening of the uterus from the vagina where cervical cancer progresses is the cervix [16]. A major source of cervical cancer is the sexually transmitted human papilloma virus (HPV) [7,17-19,82-85]. In low- and middle-income nations, the occurrence of cervical cancer is copious [20]. An important activity for cervical cancer is screening. The best screening procedure is the one that is least incursive, easy to achieve, appropriate to the patient, inexpensive, and efficient in diagnosing the disease progression at an early incursive point where it is uncomplicated to treat the disease. There are four diagnostic procedures, Cervical Cytology also known as Pap Smear Test, Biopsy, Schiller, and Hinslemann [21] and exhibited in Table 1.

Researchers have extensively utilized machine learning (ML) methods to predict diseases [28,29]. However, the

attempts have been skewed towards time to compute, and accuracy of prediction. Studies show a high accuracy of prognostication of breast cancer using ML methods. However, the researchers have not completed the analysis by validating the results and give an account of the incompleteness of the studies done so far. So, the question remains – can the accuracy levels be further improved meeting the desired performance levels?

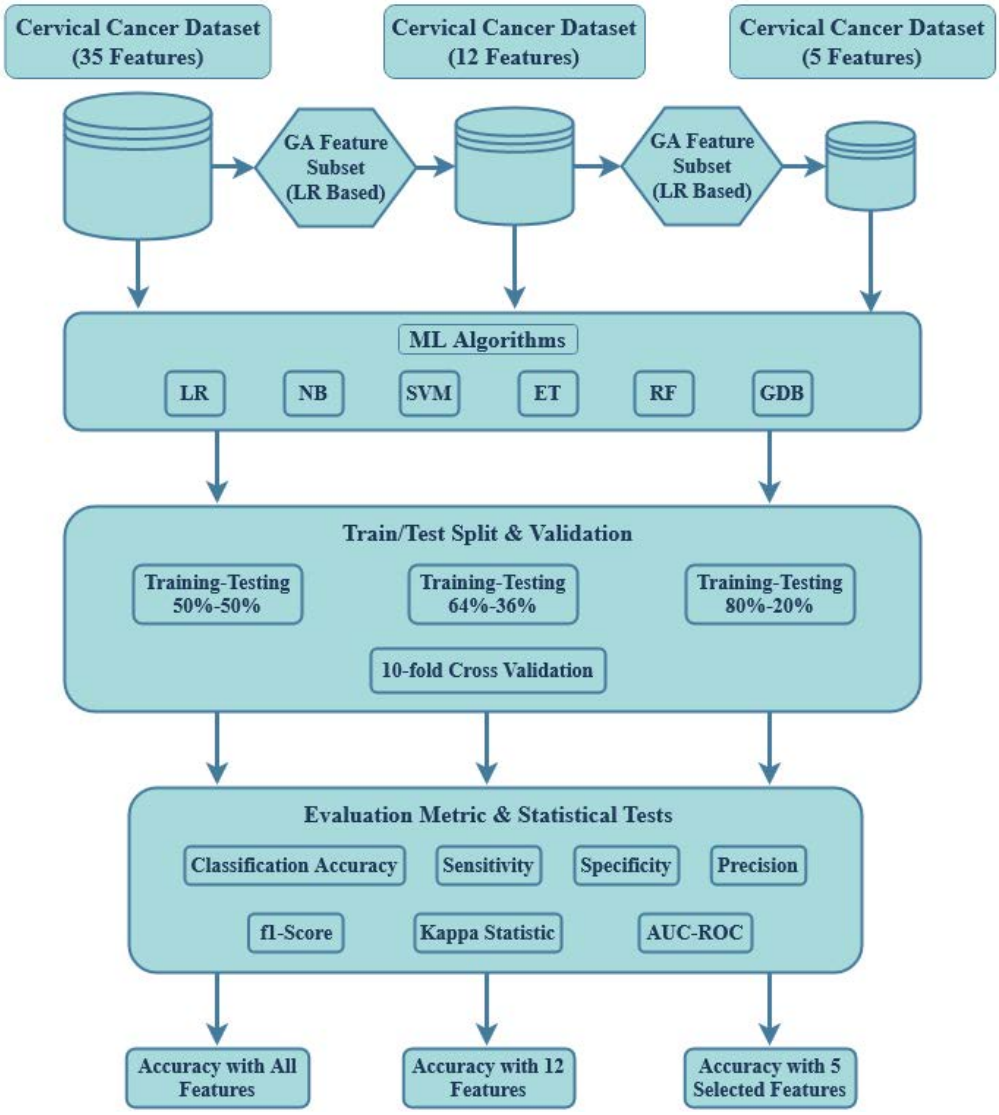

Fig.2. The Architecture for the Proposed System

Table 1. Cervical Cancer Screening Stratagem

Screening Methods	Description
Hinselmann	Hinselmann uses colposcopy examination by applying 5% acetic acid solution on the cervix tissues. A human expert can then classify the precancerous lesions in the cervical region from the change of appearance of cervix tissues.
Schiller	Schiller's method uses colposcopy examination with Lugol iodine that helps in identifying lesions that may be overlooked in the examination with acetic acid. Normal cervical regions become black or, mahogany brown stained while cervical polyps do not tinge with iodine [22,23].
Cytology	Cytology screening involves conventional and liquid-based cell examinations. Conventional cytology includes manual smearing and staining [24,25]. In liquid-based cytology, the cellular components from the cervix are submerged in a liquid [26].
Biopsy	A cervical biopsy of cells from the cervix can stipulate any abnormality, precancerous conditions or, cancer in the cervical region, unlike cytology and colposcopy screenings [27].

The database of diseases has many data items, starting from demographic to diagnostic test results to habits and similar. Researchers [30, 31] substantiated that prediction improves with the choice of the right features. As information and storage technology are making exponential progress, data sets are now omnipresent in pattern analysis, data processing, and machine learning (ML) systems, with a vast range of variables or, features. Therefore, there is a necessity for choosing a subset of many features that best suits the task. ML and training require a broad range of features that might take time to explore. Thus, FS marks the beginning of an ML analysis, followed by prediction using classifiers. There are numerous methods to select features, and these methods have shown varied outcomes when applied to different datasets. These results do not show the best method to use. Hence, researchers either apply one or, some of them or, all to carry out feature selections. These differences in results hint at the nature of the dataset. Thus, the question is how to determine a way to make dataset centric feature selection (FS) rather than the method-based FS?

The classifiers evolved, and researchers eliminated the shortcomings of incomplete, incorrect, and non-standardized data, and binomial versus multi-domain attributes. A decision tree (DT), a supervised classifier, takes care of these issues as it predicts the dependent variable [32]. Outliers do not affect the outcomes of the decision trees as the classification is constructed on the proportion of samples with split ranges and not on absolute values. This approach needs no linearity in the correlation between the dependent attributes and the independent attributes. However, this method does not yield the desired accuracy for a large database. Bayes theorem, on comparatively small and simple datasets, is regarded as one of the most conventional classification techniques attributable to its simplicity, robustness, and prediction accuracy. The NB classifier's performance against large datasets and datasets with complex attribute dependencies is poor [33]. Support Vector Machine (SVM) model takes care of this issue [34,35] but does not show good results when large datasets contain noise. It needs combining with other ML techniques [36]. The concept of a combination of methods introduced the ensemble classifiers. Random forest (RF) gained importance as an ensemble classifier and demonstrated consistent results for different datasets [37,38]. In the process of evaluating these classifiers, authors [39] observe that DT produced equivalent or, better results than RF. This situation led to the use of all popular classifiers and compared the results. In the process, some ML techniques such as logistic regression (LR) continued to be applied over broad types of variables. This method is easy to interpret as it gives the linear combination variables that predict the occurrence (or, otherwise) of the disease. However, LR's problem is its propensity to generate over-fitted models [40]. So, the point to ponder is – which is the most preferred method?

In this paper, the authors proposed a multistage Genetic Algorithm (GA) feature selection approach, which in Stage-1 achieved an approximate solution with low computational effort. In Stage-2, once again utilizes GA to evaluate the features. The GA's computational effort in Stage-2 is reduced due to the reduced feature sub-space in Stage-1. In Stage-3, several state-of-the-art machine learning classifiers and an ensemble Random Forest algorithm are used in the reduced feature sub-space to identify patients with or, without cancer. Various train-test split and 10-fold cross-validations of the dataset as training and testing are used. The experimental results, which compare predictions with – (1) all features, (2) the feature subset obtained in Stage-1, and (3) the feature subset obtained in Stage-2 demonstrate the propositioned hybrid feature selection method's efficiency and effectiveness. It reveals an enhancement in performance when compared with previous works.

3. Choice of Data Mining Models

This paper compares data-mining models, namely, LR, NB, SVM, ET, RF, and GDB for analysis of the reason for cervical cancer and accurate prediction of the disease. The step-by-step HFSSGM algorithmic procedure represented in Fig.3 is explained below.

Step 1: (Dataset Acquisition) All records from the cervical cancer dataset are collected and read.

Step 2: (Classify cancer and no-cancer without feature selection) Classification algorithms, namely LR, NB, SVM, RF, ET, and GDB are used to measure the classification accuracy of the cancer patients.

Step 3: (Selection of relevant features) GA with LR is applied to get the relevant features.

Step 4: (Classify cancer and no-cancer) Classification of optimal feature subset using algorithms such as LR, NB, SVM, RF, ET, and GDB and accuracy measured.

Step 5: (Selection of more relevant features) GA with LR is applied once again to get more relevant features.

Step 6: (Classify cancer and no-cancer) Classification of optimal feature subset using algorithms such as LR, NB, SVM, RF, ET, and GDB and accuracy measured.

Step 7: (Validation) The classifiers are trained using the validation set. A train-test partitioning and 10-fold cross-validation technique are used for testing purposes.

Step 8: (Performance parameter computation) Computation of various correctness parameters -accuracy, sensitivity, specificity, precision, f1-score, ROC-AUC, and Kappa score.

The goal of the feature selection is to find combinations of features that produce the most predictive model for cervical cancer early diagnosis and progression. A GA is used to select one or, more sets of features, and LR algorithm is employed to build a prediction model. Fig. 3 shows the algorithmic system that combined GA and LR to predict cervical cancer status. The output features from GA are used as input for LR and the resultant different variable sets are used by the GA again to optimize and identify the best set of features. Earlier in the study of heart disease, an analogous method is used [41,42]. The following sections discuss the criteria for the choice of the DMTs.

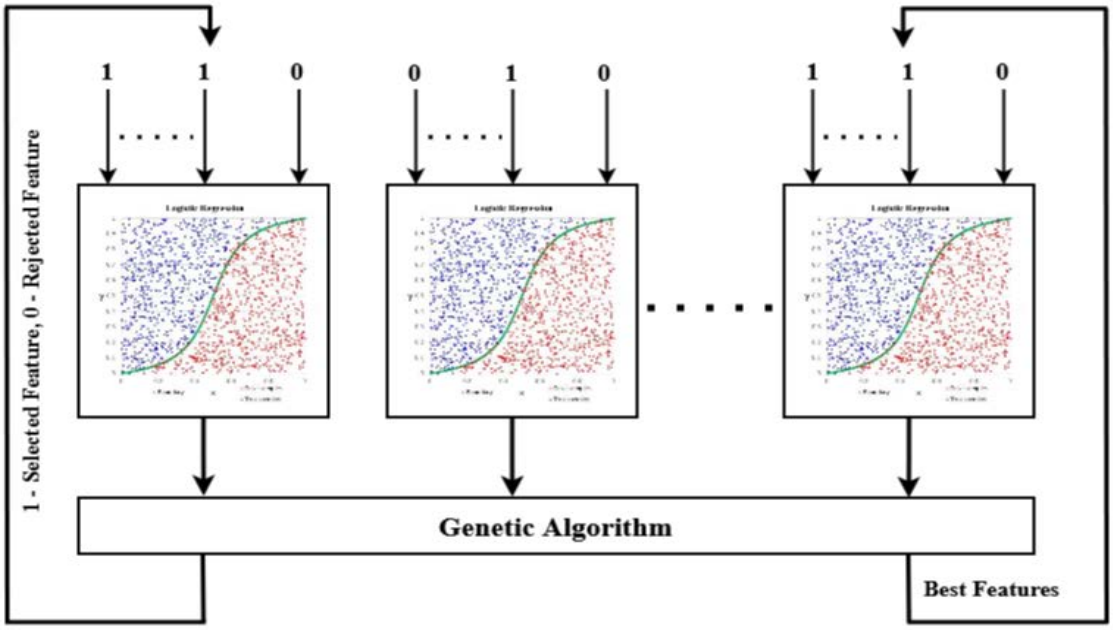

Fig.3. Genetic Algorithm and Logistic Regression Architecture for Feature Selection

3.1. Logistic Regression (LR)

LR is named after the logistic equation, the function applied at the core of the system. The logistic method, also recognized as the sigmoid function, was created by statisticians to explain the ecological features of population growth increase exponentially and optimize environmental performance. It's mainly an S-shaped curve that can obtain any number with a real value and map it to a value between 0 and 1, but on no account at those limits. A predictor or, collection of predictors of a dichotomous dependent variable, such as the presence of cancer in patients, the patient who survived or, died, or, the patient who reacted or, did not make recovery from treatment, is evaluated by Logistic Regression. Independent variables are interval, ratio level (or continuous), nominal or, ordinal (ranked) type data [43]. This can then be generalized to forecast the use of several predictors on a collection of dependent variables [44]. Each of the actions of self-care is a dichotomous variable; with either 'yes' or, 'no' responses (no answer reveals 'missing' data). Models are constructed using logistic regression from the statistics that better describe the relationships.

3.2. Naïve Bayes (NB)

Naïve Bayes Classifier is a simple type of Bayesian network classifier based on the application of the Bayes Theorem, with a strong presumption of the nonalignment of the attributes. Due to its plainness, stability, and good data set performance, the Naïve Bayes (NB) is treated as the most common classification type. NB is not doing so well for data sets where complicated attribute dependencies are present. For large data sets, the NB classification does not produce accurate results [45].

3.3. Support Vector Machine (SVM)

SVMs are one type of efficient ML technique with a high generalization capacity in practice. They are a group of margin classification models suggested by Vapnik and his group at AT&T Bell Laboratories in the 1990s [46]. In contrast to the methods of statistical learning based on practical risk minimization, the objective of SVM is to minimalize structural risk, which explains a strong ability to avoid over-fitting [47]. In the SVM model, a decision hyperplane is used for a separation gap that divides the maximum boundary between two classes. Compared to traditional ML approaches, SVMs have been utilized in many fields for their widespread generalization abilities. In particular, as a data-driven prediction technique, in recent years, SVM models have drawn the most attention to the diagnosis of diseases, such as diagnosis of cerebral palsy gait, detection of gastric lymph nodes, and diagnosis of prostate cancer [48-51].

3.4. Random Forest (RF)

RF is a well-known supervised classification method used in various classification fields. It is an ensemble learning technique [52] that operates on the concept of using a collection of weak learners to prepare a strong learner. RF uses the Classification and Regression Tree (CART) techniques [37] to create a mixture of multiple decision trees based on the bootstrap aggregation (bagging) technique [53]. The CART methodology correctly classified the dependent and independent variables and creates a relationship between them. In RF, each tree randomly chooses a subset of the dataset to build an independent decision tree. RF splits the selected random subset from the root node to the child node repeatedly until each tree reaches the leaf node without pruning. Each tree independently classifies the features and the target variable and votes for the final tree class. Depending on the margin of the votes cast, RF decides the final overall classification.

3.5. Extra Trees

The 'Extra Trees' Classifier (ETC), better discerned as the 'Extremely Randomized Trees' Classifier, is a random forest variant. Unlike a random forest, the whole sample is used at each step and the decision boundary is selected arbitrarily rather than the optimum. In real-life instances, the output is comparable to the regular, random forest, often a little better [54,55]. In specific, it is a set of decision trees and is compatible with other decision tree algorithms, such as bootstrap aggregation (bagging) and RF. The ET algorithm works by generating a huge number of un-pruned decision trees from the training dataset. In the event of regression, forecasts are made by combining the forecast of decision trees or, by using the majority vote in the case of classification.

3.6. Gradient Boosting (GDB)

GDB is a machine learning methodology for classification and regression techniques. In the form of an ensemble of weak classifiers, this generates a predictive model. There are two approaches to increase the accuracy of the classification techniques, either by the application of feature engineering or, by directly implementing boosting algorithms. The GDB algorithm was originally proposed by Friedman [56], and was widely applied in a range of clinical implementations [57-59]. GDB is a supervised and non-parametric machine learning algorithm. It approximates undefined functional mapping to subsequent output variables from input explanatory variables.

GDB requires three components – (a) The loss feature that needs to be optimized, (b) A weak learner for making predictions, (c) An additive model to abate the loss functions by adding weak learners.

GDB is a greedy algorithm that will easily overfit a training dataset. It may benefit from methods of regularization that penalize several portions of the algorithm and typically increase algorithm efficiency by reducing overfitting [60,61].

4. Performance Metrics

Performance assessment of the proposed work is accomplished using the following measures. Confusion Matrix is utilized to assess the performance of a learning model. Four terms, related to the confusion matrix are applied to establish the performance matrices. The number of cervical cancer patients classified as cancer patients is True Positive (TP). False Positive (FP) is the number of non-cancer patients classified as cervical cancer patients. True Negative (TN) is the number of patients that are classified as non-cancer patients without cervical cancer. False-negative (FN) is the number of patients classified as cancer patients without cervical cancer [62].

Accuracy: It is the ratio between numbers of correctly predicted instances to the total number of instances.

$$\text{ACCURACY} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \tag{1}$$

Precision: It evaluates the ratio of individuals predicted to be cancer patients and the total number of cancer patients.

$$\text{PRECISION} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2}$$

Recall/Sensitivity: It measures the proportion of individuals with cervical cancer and individuals predicted by an algorithm to be cancer patients.

$$\text{RECALL} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

Specificity: It evaluates the proportion of individuals without cervical cancer and individuals predicted by an algorithm to be non-cancer patients.

$$\text{SPECIFICITY} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{4}$$

F1 Score: It is the harmonic mean between precision and recall/sensitivity.

$$\text{F1SCORE} = \frac{2(\text{PRECISION} \times \text{RECALL})}{\text{PRECISION} + \text{RECALL}} \tag{5}$$

AUC-ROC Curve: AUC-ROC or, AUROC (Area Under the Receiver Operating Characteristics) is a probability curve denoting the ability of a model to differentiate between classes in the case of binary classification. ROC indicates an exchange between True Positive Rate (TPR) and False Positive Rate (FPR). AUC signifies degree or, a measure of separability where the value closer to 1 means an algorithm effectively classifies patients with and without cancer.

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{6}$$

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{7}$$

Kappa Statistics: A chance-corrected method for evaluating agreement (rather than association) between raters is Cohen's kappa () statistic [63]. Kappa is defined as follows:

$$K_{STAT} = \frac{A_{OBS} - A_{EXP}}{N - A_{EXP}} \tag{8}$$

Where, A_{EXP} is the number of agreements predicted by chance, N is the total number of observations, and A_{OBS} is the number of agreements observed between raters.

5. Dataset Description

The dataset of cervical cancer risk factors obtained from the UCI Machine Learning Repository is used for this research [26] and is exhibited in Table 2 below. The dataset, composed of demographics, custom, and medical records of 858 patients, with 32 attributes/features, and four target groups (Hinselmann, Schiller, Cytology and Biopsy), was collected at "Hospital Universitario de Caracas" in Caracas, Venezuela [26]. This analysis incorporates Hinselmann, Schiller and Cytology as attributes leaving Biopsy as the target class variable. The rationale behind this is that those variables are the result of tests conducted to establish the presence of abnormal cells. They are utilized as factors to the outcome of a biopsy result and the incidence of cervical cancer. However, the lack of values in the dataset signifies answers to a few questions were not attained due to concerns about the privacy of several patients [64]. The missing values of attributes are replaced by their mean values and normalized to remove data redundancy.

Table 2 presents all the features existing in the cervical cancer dataset and the data types of those features. Some patients decided not to answer some of the subjects based on their privacy concerns. The attributes represented by integer and boolean (0 or, 1) are regarded as data types. The missing values for integer type have been filled with the Boolean sample mean values.

Table 2. Description of the Cervical Cancer Dataset

Number	Attributes	Available Data	Missing Data	Data Type
F1	Age	858 (100%)	0 (0%)	Integer
F2	Number of sexual partners	832 (97%)	26 (3%)	Integer
F3	First Sexual intercourse (age)	851 (99%)	7 (1%)	Integer
F4	Number of Pregnancies	802 (93%)	56 (7%)	Integer
F5	Smokes	845 (98%)	13 (2%)	Boolean
F6	Smokes (years)	845 (98%)	13 (2%)	Boolean
F7	Smokes (packs/year)	845 (98%)	13 (2%)	Boolean
F8	Hormonal Contraceptives	750 (87%)	108 (13%)	Boolean
F9	Hormonal Contraceptives (years)	750 (87%)	108 (13%)	Integer
F10	Intrauterine Device (IUD)	741 (86%)	117 (14%)	Boolean
F11	IUD (years)	741 (86%)	117 (14%)	Integer
F12	Sexually Transmitted Disease (STD)	753 (88%)	105 (12%)	Boolean
F13	STDs (number)	753 (88%)	105 (12%)	Integer
F14	STDs: condylomatosis	753 (88%)	105 (12%)	Boolean
F15	STDs: cervical condylomatosis	753 (88%)	105 (12%)	Boolean
F16	STDs: vaginal condylomatosis	753 (88%)	105 (12%)	Boolean
F17	STDs: vulvo-perineal condylomatosis	753 (88%)	105 (12%)	Boolean
F18	STDs: syphilis	753 (88%)	105 (12%)	Boolean
F19	STDs: pelvic inflammatory disease	753 (88%)	105 (12%)	Boolean
F20	STDs: genital herpes	753 (88%)	105 (12%)	Boolean
F21	STDs: molluscum contagiosum	753 (88%)	105 (12%)	Boolean
F22	STDs: AIDS	753 (88%)	105 (12%)	Boolean
F23	STDs: HIV	753 (88%)	105 (12%)	Boolean
F24	STDs: Hepatitis B	753 (88%)	105 (12%)	Boolean
F25	STDs: HPV	753 (88%)	105 (12%)	Boolean
F26	STDs: Number of diagnosis	858 (100%)	0 (0%)	Integer
F27	STDs: Time since first diagnosis	71 (8%)	787 (92%)	Integer
F28	STDs: Time since last diagnosis	71 (8%)	787 (92%)	Integer
F29	Dx: Cancer	858 (100%)	0 (0%)	Boolean
F30	Dx: Cervical Intraepithelial Neoplasia (CIN)	858 (100%)	0 (0%)	Boolean
F31	Dx: Human Papillomavirus (HPV)	858 (100%)	0 (0%)	Boolean
F32	Dx (Diagnosis)	858 (100%)	0 (0%)	Boolean
Number	Target Variable	Patient	Non-Patient	Data Type
F33	Hinselmann	35 (4%)	823 (96%)	Boolean
F34	Schiller	74 (9%)	784 (91%)	Boolean
F35	Citology	44 (5%)	814 (95%)	Boolean
F36	Biopsy	55 (6%)	803 (94%)	Boolean

6. Results and Discussion

The following research questions are addressed in this research article: Which Data Mining Technique (DMT) is best for predicting diseases such as cervical cancer? and Which DMT framework can assist in meeting the three criteria-consistency, sensitivity, and specificity? To achieve the highest levels of consistency, sensitivity, and specificity, the author considers the most popular approaches and investigates their ensemble. Previous authors have focused solely on reducing variables to improve prediction. However, this method results in a loss of data. Thus, the author establishes a framework in this paper that proposes the use of data mining approaches, the measurement of consistency using kappa statistics, and the improvement of specificity and sensitivity parameters using an ensemble learning approach. As a result, the framework presented in this paper contributes to humanity's well-being by allowing for better disease prediction.

A 3-Stage Hybrid feature selection approach and a Stacked Classification model are evaluated on the cervical cancer dataset obtained from the UCI Machine Learning Repository with 35 features and one outcome variable as shown in Table 2. Stage-1 uses a Genetic Algorithm and Logistic Regression Architecture for Feature Selection and

selects twelve features well correlated with the class but not among themselves as depicted in Table 3. Stage-2 utilizes the same Genetic Algorithm and Logistic Regression Architecture for feature selection to select five features as shown in Table 4.

Table 3. Cervical Cancer Dataset with 12 Features and 1 Target Variable

Attributes
Sexually Transmitted Disease (STD), STDs (number), STDs: condylomatosis, STDs: syphilis, STDs: genital herpes, STDs: AIDS, STDs: HPV, STDs: Time since first diagnosis, Dx: Cancer, Dx: Cervical Intraepithelial Neoplasia (CIN), Dx: Human Papillomavirus (HPV), Schiller, Biopsy

Table 4. Cervical Cancer Dataset with 5 Features and 1 Target Variable

Attributes
STDs: syphilis, Dx: Cancer, Dx: Cervical Intraepithelial Neoplasia (CIN), Dx: Human Papillomavirus (HPV), Schiller, Biopsy

Table 5. Comparison of Accuracies with 35, 12 and 5 Features

Train – Test Split	Number of Features	LR	NB	SVM	ET	RF	GDB
50-50	35	0.95	0.86	0.96	0.95	0.94	0.95
	12	0.95	0.95	0.95	0.95	0.96	0.96
	5	0.94	0.94	0.96	0.96	0.96	0.96
66-34	35	0.96	0.32	0.96	0.96	0.95	0.97
	12	0.95	0.93	0.97	0.97	0.96	0.97
	5	0.97	0.95	0.97	0.97	0.97	0.97
80-20	35	0.95	0.83	0.95	0.94	0.95	0.95
	12	0.95	0.89	0.95	0.96	0.95	0.96
	5	0.96	0.93	0.95	0.96	0.96	0.96
10-fold Cross Validation	35	0.95	0.42	0.95	0.95	0.96	0.96
	12	0.96	0.83	0.96	0.96	0.96	0.96
	5	0.96	0.94	0.96	0.96	0.96	0.96

As exhibited in Table 5, different machine learning classifiers have yielded varying degrees of accuracy. 5 classifiers, namely LR, SVM, ET, RF, and GDB performed exceptionally well with an accuracy of over 93% for a distinct number of features, i.e., 35, 12, and 5. NB classifier quantified accuracy in ascending order with lowest using 35 features and highest using 5 features. This leads to highly optimistic results and does not reflect the actual predictive performance of the model. The exclusion of redundant variables ensured improvement in the classification accuracy of cervical cancer patients, but overall predicted accuracy might exhibit a shrinking effect. In this context, accuracy is not the ideal metric for assessing predictive performance, and other metrics like specificity, sensitivity, precision, f1-score, and kappa value are taken into consideration.

Table 6. Comparison of Specificity with 35, 12 and 5 Features

Train – Test Split	Number of Features	LR	NB	SVM	ET	RF	GDB
50-50	35	0.37	0.70	0.67	0.50	0.13	0.57
	12	0.27	0.87	0.70	0.70	0.80	0.73
	5	0.17	0.87	0.83	0.83	0.83	0.83
66-34	35	0.50	0.61	0.67	0.61	0.33	0.78
	12	0.22	0.89	0.89	0.89	0.78	0.83
	5	0.89	0.89	0.89	0.89	0.89	0.89
80-20	35	0.64	0.91	0.82	0.55	0.55	0.73
	12	0.82	0.91	0.91	0.91	0.82	0.91
	5	0.91	0.91	0.91	0.91	0.91	0.91
10-fold Cross Validation	35	0.98	0.42	0.94	0.97	0.99	0.98
	12	0.83	0.83	0.96	0.97	0.97	0.97
	5	0.95	0.94	0.96	0.97	0.97	0.97

Table 6, 7, 8, and 9 depict the performance analysis for specificity, sensitivity, precision, and f1-score respectively. ET, RF, and GDB attained specificity of 97% with 5 features utilizing 10-fold cross validation. RF with 35 features presented a significant specificity value rise from 13% in 50-50 data split to 99% in 10-fold cross validation. All

classifiers except NB showed superior sensitivity and f1-score for every combination of the number of features and training-testing data split. Sensitivity and f1-score increased gradually with the lessening of the number of features in the case of NB. All classifiers presented satisfactory precision rates ranging from 93% to 97% and with five features, the precision values were relatively indistinguishable.

Table 7. Comparison of Sensitivity with 35, 12 and 5 Features

Train – Test Split	Number of Features	LR	NB	SVM	ET	RF	GDB
50-50	35	0.99	0.87	0.98	0.98	1	0.98
	12	1	0.95	0.97	0.97	0.97	0.97
	5	1	0.95	0.97	0.97	0.97	0.97
66-34	35	0.99	0.30	0.98	0.98	0.99	0.98
	12	0.99	0.93	0.97	0.97	0.97	0.97
	5	0.97	0.95	0.97	0.97	0.97	0.97
80-20	35	0.98	0.82	0.96	0.96	0.98	0.96
	12	0.96	0.89	0.96	0.96	0.96	0.96
	5	0.96	0.94	0.96	0.96	0.96	0.96
10-fold Cross Validation	35	0.95	0.42	0.94	0.95	0.96	0.96
	12	0.96	0.83	0.96	0.96	0.96	0.96
	5	0.96	0.94	0.96	0.96	0.96	0.96

Table 8. Comparison of Precision with 35, 12 and 5 Features

Train – Test Split	Number of Features	LR	NB	SVM	ET	RF	GDB
50-50	35	0.94	0.93	0.96	0.94	0.94	0.95
	12	0.94	0.96	0.95	0.95	0.96	0.96
	5	0.93	0.96	0.97	0.97	0.97	0.97
66-34	35	0.96	0.87	0.96	0.96	0.94	0.97
	12	0.93	0.96	0.97	0.97	0.97	0.97
	5	0.97	0.97	0.97	0.97	0.97	0.97
80-20	35	0.95	0.95	0.96	0.94	0.95	0.95
	12	0.96	0.95	0.97	0.97	0.96	0.97
	5	0.97	0.96	0.97	0.97	0.97	0.97
10-fold Cross Validation	35	0.96	0.93	0.88	0.96	0.95	0.96
	12	0.97	0.95	0.96	0.97	0.97	0.97
	5	0.97	0.97	0.97	0.97	0.97	0.97

Table 9. Comparison of f1-score with 35, 12 and 5 Features

Train – Test Split	Number of Features	LR	NB	SVM	ET	RF	GDB
50-50	35	0.94	0.88	0.96	0.94	0.92	0.95
	12	0.93	0.95	0.95	0.95	0.96	0.96
	5	0.92	0.95	0.96	0.96	0.96	0.96
66-34	35	0.96	0.43	0.96	0.96	0.94	0.97
	12	0.93	0.94	0.97	0.97	0.96	0.97
	5	0.97	0.95	0.97	0.97	0.97	0.97
80-20	35	0.95	0.87	0.95	0.94	0.95	0.95
	12	0.96	0.91	0.96	0.96	0.96	0.96
	5	0.96	0.94	0.96	0.96	0.96	0.96
10-fold Cross Validation	35	0.95	0.56	0.90	0.95	0.95	0.96
	12	0.96	0.89	0.96	0.97	0.96	0.96
	5	0.97	0.95	0.97	0.97	0.97	0.97

The composite metric, ROC-AUC score given in Table 10, is used for comparing the performance of several classifiers and has provided clarity than accuracy, sensitivity, and specificity [65]. Kappa statistic gives the agreement rate between the expected and predicted outcome where values ranging from (1.0), (0.81-0.99), (0.61-0.80), (0.41-0.60), (0.21-0.40), (0.1-0.20) to (0) represent perfect, near-perfect, substantial, moderate, fair, slight and close to chance agreements respectively. All classifiers with five features and 10-fold cross validation confirmed the good agreement in terms of kappa value as exhibited in Table 11. Overall, GDB provided the best accuracy, precision, and specificity with five features, followed by ET, RF, and LR. The reduction in feature subspaces from 35 to 5, through feature selection, improved the performance of all classifiers as demonstrated in Table 12. Kappa statistic, precision, and specificity

values for all the six classifiers increased. Though, the AUC score of ET, RF, and GDB with 10-fold cross validation decreased for diminution in the number of features.

Table 10. Comparison of AUC Value with 35, 12 and 5 Features

Train – Test Split	Number of Features	LR	NB	SVM	ET	RF	GDB
50-50	35	0.68	0.78	0.82	0.74	0.57	0.77
	12	0.63	0.91	0.84	0.84	0.89	0.85
	5	0.58	0.91	0.90	0.90	0.90	0.90
66-34	35	0.75	0.45	0.82	0.80	0.66	0.88
	12	0.61	0.91	0.93	0.93	0.88	0.90
	5	0.93	0.92	0.93	0.93	0.93	0.93
80-20	35	0.81	0.86	0.89	0.74	0.76	0.85
	12	0.89	0.90	0.93	0.94	0.89	0.94
	5	0.94	0.92	0.93	0.94	0.94	0.94
10-fold Cross Validation	35	0.95	0.89	0.84	0.93	0.93	0.93
	12	0.93	0.89	0.94	0.89	0.90	0.89
	5	0.94	0.91	0.90	0.89	0.89	0.89

Table 11. Comparison of Kappa Statistic with 35, 12 and 5 Features

Train – Test Split	Number of Features	LR	NB	SVM	ET	RF	GDB
50-50	35	0.48	0.34	0.65	0.54	0.22	0.61
	12	0.39	0.67	0.65	0.65	0.72	0.67
	5	0.26	0.65	0.73	0.74	0.74	0.74
66-34	35	0.60	-0.02	0.64	0.63	0.44	0.72
	12	0.31	0.57	0.74	0.76	0.70	0.73
	5	0.76	0.65	0.74	0.76	0.76	0.76
80-20	35	0.61	0.33	0.64	0.49	0.54	0.61
	12	0.67	0.48	0.69	0.72	0.67	0.72
	5	0.72	0.59	0.69	0.72	0.72	0.72
10-fold Cross Validation	35	0.60	0.12	0.00	0.58	0.51	0.65
	12	0.71	0.45	0.66	0.72	0.70	0.71
	5	0.74	0.65	0.73	0.74	0.73	0.74

6.1. Effect of Feature Reduction

The reduction in feature subspace from 35 to 5, through feature selection, improves the performance of all classifiers. The Kappa values of all the six methods increase. The precision and specificity values of all methods are non-decreasing. The AUC values of ET, RF, and GDB have decreased for 10-fold cross validation. The improvement in performance due to feature reduction is provided in Table 12.

Table 12. Improvement in Performance due to Feature Reduction from 35 to 5

Performance Metrics	35 Features		12 Features		5 Features	
	ML Techniques	Maximum Score	ML Techniques	Maximum Score	ML Techniques	Maximum Score
Accuracy	GDB	0.97	SVM, ET, GDB	0.97	LR, SVM, ET, RF, GDB	0.97
Sensitivity	RF	1	LR	1	LR	1
Precision	GDB	0.97	SVM, ET, RF, GDB, LR	0.97	LR, NB, SVM, ET, RF, GDB	0.97
Specificity	RF	0.99	ET, RF, GDB	0.97	ET, RF, GDB	0.97
f1-score	GDB	0.97	SVM, ET, GDB	0.97	LR, SVM, ET, RF, GDB	0.97
AUC	LR	0.95	SVM, ET, GDB	0.94	LR, ET, RF, GDB	0.94
Kappa	GDB	0.72	ET	0.76	LR, ET, RF, GDB	0.76

6.2. Comparison of Results with Previous Research Works

Table 13 compares our proposed HFSSGM model with other research works on the same dataset. The LR model accomplishes a slightly better outcome than other models in terms of sensitivity, but the additional features make it harder to interpret. HFSSGM outperforms the ensemble model of Ahishakiye et al. [66] with 5 features in terms of accuracy by 8%. Another ensemble methodology by Lu et al. [67] performs poorly against HFSSGM in the metrics of accuracy, sensitivity, precision, and f1-score. The proposed model bests accuracy level when compared to the same of

Nasution, Sitompul & Ramli [68] and Priya & Karthikeyan [69]. The accuracy, sensitivity, and specificity of HFSSGM are way better than those of Singh [70].

Table 13. Comparison of Performance with Previous Studies

Reference	Risk Factors Used	Machine Learning Technique	Accuracy (%)	Other Results
[66]	5	Ensemble of {KNN, CART, NB, SVM} with Voting Classifier	87.21	-
[77]	6	DT	97.52	Sensitivity – 100, Specificity – 95.03, Precision – 95.27, F-measure – 97.58
[67]	14	Ensemble of {LR, DT, SVM, MLP, KNN}	83.16	Recall – 28.35, Precision – 51.73, F1 score – 32.80
[68]	12	PCA + C4.5 DT	90.70	Particularity – 100, Precision - 100
[78]	14	C5.0	100	AUC – 0.91
		RF	100	AUC – 0.91
		RPART	97	AUC – 0.81
		SVM	93	AUC – 0.8
		KNN	89	AUC – 0.5
[69]	10	DT	91.03	
		Rotation Tree	88.52	
		RF	92.63	
		SVM – Linear	93.82	
		Backpropagation	97.25	
[79]	4.6	Binary Firefly Algorithm (BFA) + RF	97.36	
[70]	7	SVM – Linear	65.47	Sensitivity – 55.56, Specificity – 66.15,
		RF	70	Sensitivity – 44.4, Specificity – 71.53,
		GBM	40.31	Sensitivity – 77.8, Specificity – 41.8,
[80]	15	Chicken Swarm Optimization (CSO) + KNN	97.82	
		CSO + RF	99.53	
Present Work	**12**	**LR, NB, SVM, ET, RF, GDB** (Values are given for GDB method.)	**96**	**Sensitivity – 96 Specificity – 97 f1-score – 96 Precision – 97 Kappa – 0.71 AUC – 89**
	5	**LR, NB, SVM, ET, RF, GDB** (Values are given for LR method.)	**96**	**Sensitivity – 96 Specificity – 95 f1-score – 97 Precision – 97 Kappa – 0.74 AUC – 94**

A comprehensive comparison is also tabulated in Table 14 where HFSSGM is judged with previous studies considering Biopsy as the lone target variable. HFSSGM's accuracy and specificity levels surpass or, are at par with the same metrics of Wu & Zhou [71], Sagala [72], and Sharma [73]. Both Abdoh, Rizka & Maghraby [74] and Geetha et al. [75] used SMOTE-RF-RFE and SMOTE-RF-PCA algorithms for identifying risk factors. The proposed model outperforms their estimations in terms of accuracy and sensitivity. HFSSGM delivers better accuracy than Yang [76], both working with 5 features.

Cancer prediction by HFSSGM with reduced features of five has resulted in slightly lesser accuracy than some learning models presented by [78] using 14 features, [80] handling 15 features, and [81] operating on 10 features. In Table 15, the results of feature selection using GA and LR are compared with the feature extraction results of existing studies. From 3-Stage Hybrid feature selection, 35 attributes originally present in the cervical cancer dataset are reduced to a selection of 12 feature set and further trimmed down to a set of 5 best features.

Acquired set of 12 features includes 'Sexually Transmitted Disease (STD)', 'STDs (number)', 'STDs: condylomatosis', 'STDs: syphilis', 'STDs: genital herpes', 'STDs: AIDS', 'STDs: HPV', 'STDs: Time since first diagnosis', 'Dx: Cancer', 'Dx: Cervical Intraepithelial Neoplasia (CIN)', 'Dx: Human Papillomavirus (HPV)', and 'Schiller'. The final set of 5 foremost features consists of 'STDs: syphilis', 'Dx: Cancer', 'Dx: Cervical Intraepithelial

Neoplasia (CIN)', 'Dx: Human Papillomavirus (HPV)', 'Schiller'. A thorough analysis of selected features determined from pertinent studies confirms the role of some prevalent risk factors as the main activators of cervical cancer. The significant factors are 'age of first sexual intercourse' [66,68-70,72,73,78], 'age of patient' [66,68-70,73,78], 'duration of hormonal contraceptive usage in years' [68,70,72,73,78,81], 'number of sexual partners' [66,68-70,78], 'number of pregnancies' [68-70,72,78], 'duration of smoking in years' [68,69,78,81], and 'number of contracted STDs' [66,68,70,81]. Other substantial risk factors are 'number of packets of smoking substance in a year' [68,70,73], 'usage of hormonal contraceptives' [66,68,69], 'duration of IUD usage' [66,68,72], and 'number of diagnosis with STDs' [69,78,81].

Interestingly, the foremost 5 features in this study are not present in the above-mentioned risk factors. These 5 features are taken into consideration independently with different combinations of risk elements by some existing analyses. This novel set of risk factors includes 'presence of syphilis (STD)' [73], diagnosis of 'Dx: cancer' [78,81], presence of 'Dx: Cervical Intraepithelial Neoplasia (CIN)' [81], diagnosis of 'Dx: Human Papillomavirus (HPV)' [78,81], and the result of 'Schiller' test [78]. Based on the evaluation of performance measures of HFSSGM, it is evident that the recommended set of risk attributes is a recent addition to the prevailing collection of high-risk factors.

Table 14. Comparison of Performance for Biopsy as Target Variable

Reference	ML Technique	Risk Factors	Accuracy (%)	Sensitivity (%)	Specificity (%)
[71]	SVM-RFE	6	92.39	100	87.32
		18	94.03	100	90.05
	SVM-PCA	8	93.45	100	89.09
		11	94.03	100	90.05
[74]	SMOTE-RF-RFE	6	95.23	94.94	95.52
		18	95.87	94.42	97.26
	SMOTE-RF-PCA	8	95.55	93.77	97.26
		11	95.74	94.16	97.76
[75]	SMOTE-RF-RFE	6	93.35	94.00	99.52
		18	95.99	93.66	99.86
	SMOTE-RF-PCA	8	96.06	92.84	99.10
		11	95.96	96.51	100
[72]	CFS + NB	4	96.24	100	92.50
	CFS + KNN		96.24	100	92.50
	CFS + SVM		93.24	100	89.45
[73]	SVM-RBF + IGAAB	7	94.45	97.10	92.40
	SVM Linear + IGAAB	7	94.33	96.70	95.20
	DT + IGAAB	7	95.97	98.90	93.50
[76]	MLP	5	96.2		
	RF		88.7		
[81]	DBSCAN + SMOTETomek + RF	10	96.71	97.41	96.04
	DBSCAN + SMOTE+ RF		97.01	98.04	95.94
	iForest + SMOTETomek + RF		98.92	98.92	98.91
	iForest + SMOTE + RF		98.93	98.94	98.13
Present Work	**LR**	**5**	**96**	**96**	**95**

Table 15. Comparison of Selected Features from Existing Literature

Reference	Selected Features/Risk Factors
Proposed Feature Selection using GA & LR	F18, F29, F30, F31, F34
[72]	F3, F4, F9, F11
[66]	F1, F2, F3, F8, F11, F13
[70]	F1, F2, F3, F4, F7, F9, F13
[73]	F1, F3, F7, F9, F10, F18, F27
[69]	F1, F2, F3, F4, F5, F6, F8, F10, F12, F26
[81]	F6, F9, F13, F20, F23, F26, F29, F30, F31, F32
[68]	F1, F2, F3, F4, F6, F7, F8, F9, F11, F13, F27, F28
[78]	F1, F2, F3, F4, F6, F9, F26, F29, F31, F32, F33, F34, F35, F36

7. Conclusion

Varied supervised learning techniques showed different results in terms of accuracy, sensitivity, and specificity. The significance of the features also changed with the methods, making treatment of serious diseases, such as cervical cancer, suffer either due to incomplete testing or, over examination or, misinterpretation due to the test parameters asynchronous values. This paper established an algorithm to search for the right features without impacting the precision levels. The large dataset was mined using GA-based iterations with a probabilistic prediction of the disease (using Logistic Regression - LR), thus recommending the set of features that enhanced the accuracy. Like any statistical method, LR results are also biased as the apparent error rates may underestimate the true value as the model tends to concentrate near the observed points that fit most. Thus, these points may wrongly illustrate an optimistic picture of the model's true precision. The reduced data set suggested from the first stage of iteration was subjected to the second round to overcome this issue. The features reduced more than half this time as well, and the accuracy remained unchanged. Proven data mining techniques employed on the second (reduced) dataset led to the highest accuracy level compared to the previous studies.

This paper pioneers a multi-stage-multi-step iterative approach for the correct diagnosis of cervical cancer and can be extended to diagnose any disease. The stages and steps are carefully chosen based on literature, their limitations, and their usefulness. However, being an iterative approach, the time-bound stopping criteria may yield local optima unless the accuracy is found to be close to cent percent. In such cases, the algorithm needs to be fine-tuned with different train-test splits.

The important features of the proposed ensemble model include its ability to determine the best possible ratio on the training dataset versus the testing dataset. The proposed model simultaneously finding the optimum combination of both sets based on the defined ratio as well as its experimentation to find an accurate rule using the ensemble method. When compared to its foundation learner and other independent learners specified in the literature, test results show that the proposed ensemble learning classification model is effective in improving performance metrics and classification accuracy. Some of the potential implications of the current research work and the proposed classifier are listed below.

1. Detection of cervical cancer is frequently accompanied by several routine medical tests performed in laboratories in the presence of experts or doctors or during hospital admission. However, this is usually a very expensive and time-consuming procedure. To predict disease with greater accuracy this proposed model uses some features extracted from regular lifestyles and a few medical test reports from laboratories in text or number format.

2. This proposed model is intended for medical service providers or doctors to provide accurate classification of cervical cancer based on fewer precise and explanatory test data from patients. As a result, the primary consumer (i.e., medical practitioners or doctors) can predict cervical cancer more quickly and accurately (including in cases of clinical assumption) and efficiently indicate the disease's risk level with the help of this model. This proposed model can be used as an electronic doctor, which means that the disease can be diagnosed even if medical practitioners are not present. As a result, it has the potential to save lives while also significantly reducing medical costs.

References

[1] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: Cancer J. Clin.*, vol. 71, pp. 209-249, 2021.

[2] N. Kamil and S. Kamil, "Global cancer incidences, causes and future predictions for subcontinent region," *Syst. Rev. Pharm.*, vol. 6, pp. 13, 2015.

[3] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2016," *CA: Cancer J. Clin.*, vol. 66, pp. 7-30, 2016.

[4] G. A. Mishra, S. A. Pimple, and S. S. Shastri, "An overview of prevention and early detection of cervical cancers," *Indian J. Med. Paediatr. Oncol.*, vol. 32, pp. 125, 2011.

[5] M. Schiffman, P. E. Castle, J. Jeronimo, A. C. Rodriguez, and S. Wacholder, "Human papillomavirus and cervical cancer," *Lancet*, vol. 370, pp. 890-907, 2007.

[6] A. C. Rodríguez, M. Schiffman, R. Herrero, A. Hildesheim, C. Bratti, M. E. Sherman, and R. D. Burk, "Longitudinal study of human papillomavirus persistence and cervical intraepithelial neoplasia grade 2/3: critical role of duration of infection," *J. Natl. Canc. Inst.*, vol. 102, pp. 315-324, 2010.

[7] World Health Organization (WHO), https://www.who.int/news-room/fact-sheets/detail/human-papillomavirus-(hpv)-and-cervical-cancer. (Accessed 28 March 2021).

[8] M. E. Plissiti and C. Nikou, "A review of automated techniques for cervical cell image analysis and classification," in *Biomedical Imaging and Computational Modeling in Biomechanics*, U. Andreaus and D. Iacoviello, Eds. Dordrecht: Springer, 2013, pp. 1-18.

[9] A. Jemal, M. M. Center, C. DeSantis, and E. M. Ward, "Global patterns of cancer incidence and mortality rates and trends," *Cancer Epidemiol. Biomarkers Prev.*, vol. 19, pp. 1893-1907, 2010.

[10] S. Bobdey, J. Sathwara, A. Jain, and G. Balasubramaniam, "Burden of cervical cancer and role of screening in India," *Indian J. Med. Paediatr. Oncol.*, vol. 37, pp. 278, 2016.

[11] L. Kjellberg, G. Hallmans, A. M. Åhren, R. Johansson, F. Bergman, G. Wadell, and J. Dillner, "Smoking, diet, pregnancy and

oral contraceptive use as risk factors for cervical intra-epithelial neoplasia in relation to human papillomavirus infection," *Br. J. Canc.*, vol. 82, pp. 1332-1338, 2000.
[12] M. Plummer, R. Herrero, S. Franceschi, C. J. Meijer, P. Snijders, F. X. Bosch, and N. Muñoz, "Smoking and cervical cancer: pooled analysis of the IARC multi-centric case–control study," *Canc. Causes Contr.*, vol. 14, pp. 805-814, 2003.
[13] P. Luhn, J. Walker, M. Schiffman, R. E. Zuna, S. T. Dunn, M. A. Gold, and N. Wentzensen, "The role of co-factors in the progression from human papillomavirus infection to cervical cancer," *Gynecol. Oncol.*, vol. 128, pp. 265-270, 2013.
[14] V. Moreno, F. X. Bosch, N. Muñoz, C. J. Meijer, K. V. Shah, J. M. Walboomers, and International Agency for Research on Cancer (IARC) Multicentric Cervical Cancer Study Group, "Effect of oral contraceptives on risk of cervical cancer in women with human papillomavirus infection: the IARC multicentric case-control study," *Lancet*, vol. 359, pp. 1085-1092, 2002.
[15] S. R. Pradhan, S. Mahata, D. Ghosh, P. K. Sahoo, S. Sarkar, R. Pal, and V. D. Nasare, "Human Papillomavirus Infections in Pregnant Women and Its Impact on Pregnancy Outcomes: Possible Mechanism of Self-Clearance," in *Human Papillomavirus*, R. Rajkumar, Eds. IntechOpen, 2020, pp. 1-27.
[16] S. Subramanian, R. Sankaranarayanan, P. O. Esmy, J. V. Thulaseedharan, R. Swaminathan, and S. Thomas, "Clinical trial to implementation: Cost and effectiveness considerations for scaling up cervical cancer screening in low-and middle-income countries," *J. Cancer Policy*, vol. 7, pp. 4-11, 2016.
[17] K. U. Petry, "HPV and cervical cancer," *Scand. J. Clin. Lab. Invest.*, vol. 74, pp. 59-62, 2014.
[18] G. Ronco, J. Dillner, K. M. Elfström, S. Tunesi, P. J. Snijders, M. Arbyn, and International HPV Screening Working Group, "Efficacy of HPV-based screening for prevention of invasive cervical cancer: follow-up of four European randomised controlled trials," *Lancet*, vol. 383, pp. 524-532, 2014.
[19] Y. Hiraku, S. Kawanishi, and H. Ohshima, Eds. *Cancer and inflammation mechanisms: chemical, biological, and clinical aspects*. New Jersey: John Wiley & Sons, 2014.
[20] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: Cancer J. Clin.*, vol. 68, pp. 394-424, 2018.
[21] R. A. Kerkar and Y. V. Kulkarni, "Screening for cervical cancer: an overview," *J. Obstet. Gynecol. India*, vol. 56, pp. 115-122, 2006.
[22] J. W. Sellors and R. Sankaranarayanan, *Colposcopy and treatment of cervical intraepithelial neoplasia: a beginner's manual*. International Agency for Research on Cancer, 2003.
[23] H. Ramaraju, Y. Nagaveni, and A. Khazi, "Use of Schiller's test versus Pap smear to increase the detection rate of cervical dysplasias," *Int. J. Reprod. Contracept. Obstet. Gynecol.*, vol. 5, pp. 1446-1450, 2017.
[24] E. Bengtsson and P. Malm, "Screening for cervical cancer using automated analysis of PAP-smears," *Comput. Math. Meth. Med.*, 2014.
[25] G. Guvenc, A. Akyuz, and C. H. Açikel, "Health belief model scale for cervical cancer and Pap smear test: psychometric testing," *J. Adv. Nurs.*, vol. 67, pp. 428-437, 2011.
[26] K. Fernandes, J. S. Cardoso, and J. Fernandes, "Transfer learning with partial observability applied to cervical cancer screening," in *Iberian conference on pattern recognition and image analysis*, Cham: Springer, June 2017, pp. 243-250.
[27] M. T. Galgano, P. E. Castle, K. A. Atkins, W. K. Brix, S. R. Nassau, and M. H. Stoler, "Using biomarkers as objective standards in the diagnosis of cervical biopsies," *Am. J. Surg. Pathol.*, vol. 34, pp. 1077, 2010.
[28] M. U. Sarwar, M. K. Hanif, R. Talib, A. Mobeen, and M. Aslam, "A survey of big data analytics in healthcare," *Int. J. Adv. Comput. Sci. Appl.*, vol. 8, pp. 355-359, 2017.
[29] E. W. Steyerberg, *Clinical prediction models*. Cham: Springer International Publishing, 2019.
[30] M. Tubishat, N. Idris, L. Shuib, M. A. Abushariah, and S. Mirjalili, "Improved Salp Swarm Algorithm based on opposition based learning and novel local search algorithm for feature selection," *Expert Syst. Appl.*, vol. 145, 2020.
[31] S. Maldonado, J. López, A. Jimenez-Molina, and H. Lira, "Simultaneous feature selection and heterogeneity control for SVM classification: An application to mental workload assessment," *Expert Syst. Appl.*, vol. 143, 2020.
[32] M. Shouman, T. Turner, and R. Stocker, "Applying k-nearest neighbour in diagnosing heart disease patients," *Int. J. Inf. Educ. Technol.*, vol. 2, pp. 220-223, 2012.
[33] Y. Ji, S. Yu, and Y. Zhang, "A novel naive bayes model: Packaged hidden naive bayes," in *2011 6th IEEE Joint International Information Technology and Artificial Intelligence Conference*, vol. 2, IEEE, August 2011, pp. 484-487.
[34] B. Schölkopf, A. J. Smola, R. C. Williamson, and P. L. Bartlett, "New support vector algorithms," *Neural Comput.*, vol. 12, pp. 1207-1245, 2000.
[35] G. Cavallaro, M. Riedel, M. Richerzhagen, J. A. Benediktsson, and A. Plaza, "On understanding big data impacts in remotely sensed image classification using support vector machine methods," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 8, pp. 4634-4646, 2015.
[36] Y. Tang and J. Zhou, "The performance of PSO-SVM in inflation forecasting," in *2015 12th International Conference on Service Systems and Service Management (ICSSSM)*, IEEE, June 2015, pp. 1-4.
[37] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, pp. 5-32, 2001.
[38] X. Chen and H. Ishwaran, "Random forests for genomic data analysis," *Genomics*, vol. 99, pp. 323-329, 2012.
[39] T. Shaikhina, D. Lowe, S. Daga, D. Briggs, R. Higgins, and N. Khovanova, "Decision tree and random forest models for outcome prediction in antibody incompatible kidney transplantation," *Biomed. Signal Process. Contr.*, vol. 52, pp. 456-462, 2019.
[40] M. A. Babyak, "What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models," *Psychosom. Med.*, vol. 66, pp. 411-421, 2004.
[41] O. Gayou, S. K. Das, S. M. Zhou, L. B. Marks, D. S. Parda, and M. Miften, "A genetic algorithm for variable selection in logistic regression analysis of radiotherapy treatment outcomes," *Med. Phys.*, vol. 35, pp. 5426-5433, 2008.
[42] A. K. Chaudhuri and A. Das, "Variable Selection in Genetic Algorithm Model with Logistic Regression for Prediction of Progression to Diseases," in *2020 IEEE International Conference for Innovation in Technology (INOCON)*, IEEE, November

2020, pp. 1-6.
[43] L. Connelly, "Logistic regression," *Medsurg Nurs.*, vol. 29, pp. 353-354, 2020.
[44] E. Tzeng, C. Devin, J. Hoffman, C. Finn, P. Abbeel, S. Levine, and T. Darrell, "Adapting deep visuomotor representations with weak pairwise constraints," in *Algorithmic Foundations of Robotics XII*, Cham: Springer, December 2016, pp. 688-703.
[45] M. A. Jabbar and S. Samreen, "Heart disease prediction system based on hidden naïve bayes classifier," in *2016 International Conference on Circuits, Controls, Communications and Computing (I4C)*, IEEE, October 2016, pp. 1-5.
[46] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, pp. 273-297, 1995.
[47] N. E. Ayat, M. Cheriet, and C. Y. Suen, "Automatic model selection for the optimization of SVM kernels," *Pattern Recogn.*, vol. 38, pp. 1733-1745, 2005.
[48] J. Kamruzzaman and R. K. Begg, "Support vector machines and other pattern recognition approaches to the diagnosis of cerebral palsy gait," *IEEE Trans. Biomed. Eng.*, vol. 53, pp. 2479-2490, 2006.
[49] Y. J. Son, H. G. Kim, E. H. Kim, S. Choi, and S. K. Lee, "Application of support vector machine for prediction of medication adherence in heart failure patients," *Healthc. Inform. Res.*, vol. 16, pp. 253-259, 2010.
[50] T. Ishikawa, J. Takahashi, H. Takemura, H. Mizoguchi, and T. Kuwata, "Gastric lymph node cancer detection using multiple features support vector machine for pathology diagnosis support system," in *The 15th International Conference on Biomedical Engineering*, Cham: Springer, 2014, pp. 120-123.
[51] V. Shah, B. Turkbey, H. Mani, Y. Pang, T. Pohida, M. J. Merino, and M. Bernardo, "Decision support system for localizing prostate cancer based on multiparametric magnetic resonance imaging," *Med. Phys.*, vol. 39, pp. 4093-4103, 2012.
[52] W. Y. Loh, "Classification and regression trees," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 1, pp. 14-23, 2011.
[53] A. Liaw and M. Wiener, "Classification and regression by randomForest," *R news*, vol. 2, pp. 18-22, 2002.
[54] A. K. Verma, S. Pal, and S. Kumar, "Prediction of skin disease using ensemble data mining techniques and feature selection method—a comparative study," *Appl. Biochem. Biotechnol.*, vol. 190, pp. 341-359, 2020.
[55] O. Maier, M. Wilms, J. von der Gablentz, U. M. Krämer, T. F. Münte, and H. Handels, "Extra tree forests for sub-acute ischemic stroke lesion segmentation in MR sequences," *J. Neurosci. Meth.*, vol. 240, pp. 89-100, 2015.
[56] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Ann. Stat.*, pp. 1189-1232, 2001.
[57] S. Dodd, M. Berk, K. Kelin, Q. Zhang, E. Eriksson, W. Deberdt, and J. C. Nelson, "Application of the Gradient Boosted method in randomised clinical trials: Participant variables that contribute to depression treatment efficacy of duloxetine, SSRIs or placebo," *J. Affect. Disord.*, vol. 168, pp. 284-293, 2014.
[58] J. Xie and S. Coggeshall, "Prediction of transfers to tertiary care and hospital mortality: A gradient boosting decision tree approach," *Stat. Anal. Data Min.*, vol. 3, pp. 253-258, 2010.
[59] Y. Chen, Z. Jia, D. Mercola, and X. Xie, "A gradient boosting algorithm for survival analysis via direct optimization of concordance index," *Comput. Math. Meth. Med.*, 2013.
[60] J. C. Weiss, D. Page, P. L. Peissig, S. Natarajan, and C. McCarty, "Statistical relational learning to predict primary myocardial infarction from electronic health records," *Proc. Innov. Appl. Artif. Intell. Conf.*, vol. 2012, pp. 2341-2347, 2012.
[61] R. Martin, D. Rose, K. Yu, and S. Barros, "Toxicogenomics Strategies for Predicting Drug Toxicity," *Pharmacogenomics*, vol. 7, pp. 1003-1016, 2006.
[62] A. Ray and A. K. Chaudhuri, "Smart healthcare disease diagnosis and patient management: Innovation, improvement and skill development," *Machine Learning with Applications*, vol. 3, 2021.
[63] L. F. Chalak, L. Pavageau, B. Huet, and L. Hynan, "Statistical rigor and kappa considerations: which, when and clinical context matters," *Pediatr. Res.*, vol. 88, pp. 5, 2020.
[64] N. Razali, S. A. Mostafa, A. Mustapha, M. H. Abd Wahab, and N. A. Ibrahim, "Risk Factors of Cervical Cancer using Classification in Data Mining," *J. Phys. Conf.*, vol. 1529, pp. 022102, April 2020.
[65] G. Vandewiele, I. Dehaene, G. Kovács, L. Sterckx, O. Janssens, F. Ongenae, and T. Demeester, "Overly optimistic prediction results on imbalanced data: a case study of flaws and benefits when applying over-sampling," *Artif. Intell. Med.*, vol. 111, 2021.
[66] E. Ahishakiye, R. Wario, W. Mwangi, and D. Taremwa, "Prediction of Cervical Cancer Basing on Risk Factors using Ensemble Learning," in *2020 IST-Africa Conference (IST-Africa)*, IEEE, May 2020, pp. 1-12.
[67] J. Lu, E. Song, A. Ghoneim, and M. Alrashoud, "Machine learning for assisting cervical cancer diagnosis: An ensemble approach," *Future Generat. Comput. Syst.*, vol. 106, pp. 199-205, 2020.
[68] M. Z. F. Nasution, O. S. Sitompul, and M. Ramli, "PCA based feature reduction to improve the accuracy of decision tree c4.5 classification," *J. Phys. Conf.*, vol. 978, pp. 012058, 2018.
[69] S. Priya and N. K. Karthikeyan, "A Heuristic and ANN based Classification Model for Early Screening of Cervical Cancer," *Int. J. Comput. Intell. Syst.*, vol. 13, pp. 1092-1100, 2020.
[70] H. D. Singh, *Diagnosis of Cervical Cancer using Hybrid Machine Learning Models*. Doctoral dissertation, Dublin, National College of Ireland, 2018.
[71] W. Wu and H. Zhou, "Data-driven diagnosis of cervical cancer with support vector machine-based approaches," *IEEE Access*, vol. 5, pp. 25189-25195, 2017.
[72] N. T. Sagala, "A Comparative Study of Data Mining Methods to Diagnose Cervical Cancer," *J. Phys. Conf.*, vol. 1255, pp. 012022, 2019.
[73] M. Sharma, "Cervical cancer prognosis using genetic algorithm and adaptive boosting approach," *Health Technol. (Berl)*, vol. 9, pp. 877-886, 2019.
[74] S. F. Abdoh, M. A. Rizka, and F. A. Maghraby, "Cervical cancer diagnosis using random forest classifier with SMOTE and feature reduction techniques," *IEEE Access*, vol. 6, pp. 59475-59485, 2018.
[75] R. Geetha, S. Sivasubramanian, M. Kaliappan, S. Vimal, and S. Annamalai, "Cervical cancer identification with synthetic minority oversampling technique and PCA analysis using random forest classifier," *J. Med. Syst.*, vol. 43, pp. 1-19, 2019.
[76] W. Yang, X. Gou, T. Xu, X. Yi, and M. Jiang, "Cervical Cancer Risk Prediction Model and Analysis of Risk Factors based on Machine Learning," in *Proceedings of the 2019 11th International Conference on Bioinformatics and Biomedical Technology*, New York: Association for Computing Machinery, May 2019, pp. 50-54.

[77] Y. M. S. Al-Wesabi, A. Choudhury, and D. Won, "Classification of cervical cancer dataset," in *Proceedings of the 2018 IISE Annual Conference*, IISE, December 2018, pp. 1456-1461.
[78] B. Nithya and V. Ilango, "Evaluation of machine learning based optimized feature selection approaches and classification methods for cervical cancer prediction," *SN Applied Sciences*, vol. 1, pp. 1-6, 2019.
[79] R. Sawhney, P. Mathur, and R. Shankar, "A firefly algorithm based wrapper-penalty feature selection method for cancer diagnosis," in *International Conference on Computational Science and Its Applications*, Cham: Springer, July 2018, pp. 438-449.
[80] A. K. Tripathi, P. Garg, A. Tripathy, N. Vats, D. Gupta, and A. Khanna, "Prediction of Cervical Cancer Using Chicken Swarm Optimization," in *International Conference on Innovative Computing and Communications*, Singapore: Springer, 2020, pp. 591-604.
[81] M. F. Ijaz, M. Attique, and Y. Son, "Data-driven cervical cancer prediction model with outlier detection and over-sampling methods," *Sensors*, vol. 20, pp. 2809, 2020.
[82] Prabhjot Kaur, Yashita Pruthi, Vidushi Bhatia, Janmjay Singh,"Empirical Analysis of Cervical and Breast Cancer Prediction Systems using Classification", International Journal of Education and Management Engineering, Vol.9, No.3, pp.1-15, 2019.
[83] Dhwaani Parikh, Vineet Menon,"Machine Learning Applied to Cervical Cancer Data", International Journal of Mathematical Sciences and Computing, Vol.5, No.1, pp.53-64, 2019.
[84] Kemal Akyol," A Study on Test Variable Selection and Balanced Data for Cervical Cancer Disease", International Journal of Information Engineering and Electronic Business, Vol.10, No.5, pp. 1-7, 2018.
[85] G. Anna Lakshmi, S. Ravi," A Double Layered Segmentation Algorithm for Cervical Cell Images based on GHFCM and ABC", International Journal of Image, Graphics and Signal Processing, Vol.9, No.11, pp. 39-47, 2017.

Authors' Profiles

Avijit Kumar Chaudhuri received his B.SC degree from Calcutta University, A.M.I.E degree in computer engineering from The Institution of Engineers (India). He has completed MTech and MBA degrees in computer science and E-Business from the Sam Higginbottom University of Agriculture, Technology & Sciences – SHUATS, formerly known as AAIDU, India and Annamalai University, India in 2007 and 2014, respectively. Since July 2010, he has been with the Department of Computer Science and Engineering, Techno Engineering College Banipur, where he is an Assistant Professor. Prior to that, he was the Academic In-Charge, Sikkim Manipal University Learning Centre, Kolkata. Mr. Chaudhuri started his career as a Centre Academic Head in Aptech Computer Education and worked as Visiting Faculty in University Learning Centers of Vidyasagar University and IGNOU. His current research interests include Artificial Intelligence (AI), Data Mining, Machine Learning and Hybrid Learning.

Arkadip Ray received his B. Tech degree in Computer Science & Engineering from Government College of Engineering & Ceramic Technology (GCECT), Kolkata, India and completed M. Tech degree in Information Technology from GCECT with GATE scholarship. His current research interests include Digital Watermarking, Network Security, Data Mining, Machine Learning. He has 8 research publications in esteemed International and National journals.

How to cite this paper: Avijit Kumar Chaudhuri, Arkadip Ray, Dilip K. Banerjee, Anirban Das, "A Multi-Stage Approach Combining Feature Selection with Machine Learning Techniques for Higher Prediction Reliability and Accuracy in Cervical Cancer Diagnosis", International Journal of Intelligent Systems and Applications(IJISA), Vol.13, No.5, pp.46-63, 2021. DOI: 10.5815/ijisa.2021.05.05