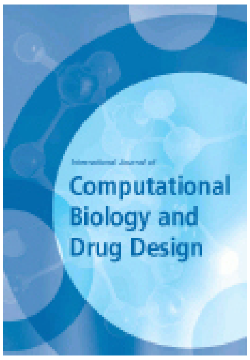


International Journal of Computational Biology and Drug Design

ISSN online: 1756-0764 - ISSN print: 1756-0756
https://www.inderscience.com/ijcbdd

Random forest with SMOTE and ensemble feature selection for cervical cancer diagnosis

Anjali Kuruvilla, B. Jayanthi

DOI: 10.1504/IJCBDD.2023.10053107

Article History:

Received:	06 August 2022
Accepted:	06 October 2022
Published online:	17 April 2023

Random forest with SMOTE and ensemble feature selection for cervical cancer diagnosis

Anjali Kuruvilla* and B. Jayanthi

School of Computer Studies,
Rathnavel Subramaniam College of Arts and Science,
Coimbatore, 641 402, Tamil Nadu, India
Email: anju.anjali2013@gmailcom
Email: jayanthi@rvsgroup.com
*Corresponding author

Abstract: Cervical tumours are a leading cause of death worldwide, although they can be prevented by removing afflicted tissues early on. Recognising population weaknesses is necessary for inclusive cervical screening programs. STDs and smoking cause cervical cancer. Creating a cancer classifier requires complex learning. FS decreases a prediction system's inputs. Reducing model parameters and time improves performance. The goal is to create a new ensemble feature selection (EFS) and classifier for cervical cancer diagnosis. EFS, several FSs used. EFS mixes the results of single FS approaches, including entropy elephant herding optimisation (EEHO), entropy elephant herding optimisation (EBFO), and recursive feature elimination (RFE), to improve results. Bootstrap aggregates EFS results. Classifier approach is Random Forest with SMOT (SMOTE). UCI's cancer database has 32 features and four classes. Classification performance is calculated using a confusion matrix and precision, recall, f-measure, and accuracy. The classification algorithms use MATLAB. The proposed algorithm gives an enhanced accuracy value of 94.7552%, 94.5221%, 94.8718%, and 94.2890% for the Hinselmann, Schiller, Citology, and Biopsy tests, respectively.

Keywords: cervical cancer; EFS; ensemble feature selection; entropy elephant herding optimisation; entropy elephant herding optimisation; EBFO; entropy butterfly optimisation algorithm; RFE; recursive feature elimination; RF; random forest; SMOTE; synthetic minority oversampling technique; classification.

Reference to this paper should be made as follows: Kuruvilla, A. and Jayanthi, B. (2023) 'Random forest with SMOTE and ensemble feature selection for cervical cancer diagnosis', *Int. J. Computational Biology and Drug Design*, Vol. 15, No. 4, pp.289–315.

Biographical notes: Anjali Kuruvilla is currently pursuing a PhD at RVS College of Arts and Science, Coimbatore and also working as Assistant Professor in the Department of Computer Applications, Alphonsa arts and science college, Affiliated to Calicut University, Kerala. She received an MPhil degree from RVS College of Arts and Science, Affiliated with Bharathiar University, Coimbatore. Her main research work focuses on data mining, pattern recognition, data visualisation and presentation and machine learning.

B. Jayanthi is currently working as HOD in the Department of Computer Studies, RVS College of Arts and Science, Affiliated with Bharathiar University, Coimbatore. She has more than 15 years of experience in teaching. She has many Scholars pursuing their PhD under her Guidance. Her main research work focuses on data mining and statistical analysis, operations-related data analytics, data warehousing and pattern recognition.

1 Introduction

One category of gynaecological tumour is cervical disease, and the problem of cervical cancer is frequently connected with human papilloma virus (HPV) infection. It is a typical crippling infection among women around the world. It is the third chiefly consistently analysed disease (approximately 485,000 occurrences) and the 4th largely powerful reason used for tumour-associated fatality (236,000) every year (Yang et al., 2018; Arbyn et al., 2020). The primary reason for this cancer is persevering illness via oncogenic HPV (Seo et al., 2016). Further issues like sexually transmitted diseases, oral defensive exploitation, smoking conditions, equality, and diet could contribute to cervical cancer improvement (Suehiro et al., 2019). Commonly, patients recognised with cervical cancer growth at introductory stages offer no perceptible hints or signs that can prompt misdiagnosis (Khan et al., 2019). Cervical cancer features could be extended by 2 to multiple periods if an HPV-infected long-suffering smokes (Brisson et al., 2020). In the event of different pregnancies, woman HPV-tainted patients with no pregnancies have lesser events of this disease when compared to more than single grown-up pregnancies.

Cervical cancer incidence is plentiful in lower and medium-developed countries (Bray et al., 2018). Screening is a significant task in a cervical tumour. The best diagnostic test is the least incursive, easy to accomplish, satisfactory to focus on, modest, and compelling in identifying the illness cycle in its initial incursive phase, while the simple treatment for the disease. Common screening strategies for cervical cancer are biopsy, Schiller, Hinslemann, and Cytology (Bedell et al., 2020). Cytology strategy is a minute analysis of cells smashed from the cervix, and it is utilised to distinguish carcinogenic states of the cervix (Bouvard et al., 2021). The biopsy technique is a careful interaction that incorporates the discovery of a livelihood tissue test to perform the analysis (Rerucha et al., 2018). The iodine solution is used for ocular investigation of the cervix, identified as Hinslemann analysis. Lugol's iodine is utilised for optical investigation of the cervix behind spreading Lugol's iodine recognition pace of suspicious areas over the cervix named as Schiller test (Ramaraju et al., 2016).

To resolve the restrictions and further develop the screening tests' quality, computer vision and computer-aided frameworks via data mining (DM) are utilised to check screening tests, building the cycle more precise and consistent. DM is assumed as one of the main demanding and significant exploration areas in clinical medicine because of the great significance of relevant medical problems. In the clinical region, DM methods are useful not just in discovering examples and connections amongst certain indications yet in addition in foreseeing different sicknesses (Pramanik et al., 2023). Carrying out various DM strategies is constantly examined, and clinical consideration could be recommended quickly to protect the lives, particularly the person who experiences cervical cancer.

Previous investigations utilised different machine learning (ML) strategies for cervical cancer detection and prediction (Wu and Zhou, 2017; Lu et al., 2020; Abdoh et al., 2018). In any case, ML procedures handle a few difficulties, consisting of issues of the best selection of attributes from the dataset, appropriating class, and achieving outcomes along with increased accuracy of classification. In this way, the current work focuses on tackling these difficulties. Earlier analyses are not consolidated EFS and balanced data for cervical disease diagnosis. FS turns into the fundamental process for numerous DM purposes. Choosing suitable attributes in the information is significant because unessential attributes could reduce the numerous classifiers' accuracy (Park et al., 2017). FS techniques are generally separated into filter, wrapper, and embedded strategies. Utilising a solitary attribute subset determination technique might create nearby optima. Other than these three notable FS drawing near, another gathering of techniques is built over the previous FS strategies: ensemble FS (Seijo-Pardo et al., 2017). EFS builds a group of attribute subsets and afterwards joins these subsets to create accumulated outcomes. EFS techniques are applied to join different FS strategies as opposed to utilising a single FS method (Brahim and Limam, 2018). Traditional soft computing algorithms have not proficiently worked in the EFS of high-dimensional dataset issues (Arora et al., 2020; Bansal and Jain, 2021). Then, various meta-heuristic approaches are adjusted for FS issues.

In the proposed work, a cervical cancer diagnosis model (CCDM) is introduced by using EFS, and RF is applied to the cervical cancer dataset. Thus, the key novelty of the current investigation is to join the FS techniques like entropy elephant herding optimisation (EEHO), entropy elephant herding optimisation (EBFO), and recursive feature elimination (RFE) to improve prediction accuracy. Data oversampling is performed via synthetic minority oversampling technique (SMOTE) method for adjusting the dataset. RF classifier is presented for the classification of cervical cancer based on selected features to increase the classification results. RF algorithm works better than the conventional classification methods. RF algorithm is a significant ML strategy because of its benefits of managing unequal datasets, speedy computation, and gives better performance.

The following organisation of the paper is described as follows: Section 2 provides a complete review of cervical cancer in terms of pre-processing, feature selection, and classification methods. The proposed approach for ensemble feature selection (EFS) and RF using SMOTE is defined in Section 3. Results evaluation of classification methods before and after SMOTE are discussed in Section 4. Finally, the overall work is concluded, and the scope of the work is included in Section 5.

2 Literature review

Wu and Zhou (2017) introduced the analysis of cervical cancer correctly using the support vector machine (SVM) algorithm. SVM techniques, SVM-RFE, and SVM-principal component analysis (SVM-PCA) are additionally introduced to analyse the harmful malignant tumour samples. Of all the classifiers, SVM-PCA outperforms than other two classifiers via the UCI dataset.

Abdoh et al. (2018) suggested feature variables of cervical cancer to the prediction model utilising the RF classifier method along with the SMOTE and two FS strategies, such as RFE and PCA. Many of the medicinal databases are often not balanced since the number of patients is significantly limited compared to the number of non-patients, which is solved by the SMOTE. Once evaluating the results, determining the hybrid of the RF classifier with SMOTE is developed to evaluate the performance of the system.

Jain et al. (2019) introduced the system performance, which is checked with base classifiers like RF, Kernel SVM (KSVM), decision tree (DT), and k-Nearest Neighbour (kNN), and afterwards assessed the outcomes with and without binary cuckoo optimisation (BCO). Cuckoo search optimisation (CSO) is introduced for the selection of optimal features from the dataset. The outcomes created presently chosen features to play a major vital role in cancer classification. Also, it shows that this classifier gives greater efficiency.

Nithya and Ilango (2019) intended to achieve further arrangement by utilising ML procedures in R to examine the feature variables of the dataset. Different categories of FS methods are developed to choose the essential features for cervical disease classification. Important attributes are recognised over different emphases of the training model via various FS techniques, and optimised FS methods are constructed. Moreover, this work intended to assemble some prediction models utilising C5.0, RF, kNN, and SVM methods. Most extreme prospects are investigated for training and execution assessment of the comparative methods. The classification accuracy of these methods is given in this work, dependent on the results achieved by them. In general, C5.0 and RF classifiers are implemented logically well with increased accuracy for distinguishing women displaying an experimental indication of cervical disease.

Ahishakiye et al. (2020) suggested an ensemble training method for cervical tumour classification utilising features. This method is chosen since it consolidates numerous ML methods into one model to reduce variance, bias, and development in execution. The algorithms are such as kNN, Classification and Regression Trees (CART), Naïve Bayes (NBs) algorithm, and SVM. Prediction techniques are chosen since the attention of this investigation is to tackle the prediction issue. Therefore, these algorithms might work well in the problem domain. The last classification model is trained and verified utilising these classifier models.

Nithya and llango (2020) aimed to diagnose the cervical tumour, and the given database consists of missing variables, repetitive attributes, and unbalanced objective labels. Subsequently, this work focused on dealing with these problems via the coordinated FS method to obtain an ideal attribute subset. The subsets achieved via this combined method could be employed to enhance prediction results. The perfect FS method could be preferred depending on the results and effectiveness of the classification methods in estimating the outcomes. For bio clinical and bioinformatics datasets, achieving the highest accuracy for data classification is difficult from this system. Thus, the point of this examination is to enhance a complete structure with combined FS algorithms to achieve optimal attribute subsets with prediction accuracy. Also, it provides lower computational complexities for the cervical cancer dataset.

Priya and Karthikeyan (2021) presented a short-term long memory with an artificial bee colony (LSTM-ABC) method for cervical tumour identification. The examination handles cervical tumour detection and utilises SMOTE to tackle the unbalanced class problem. From the pre-processed dataset, the FS is done utilising the ABC optimisation algorithm. The LSTM method is utilised for predicting cervical cancer growth dependent on the chosen attributes. The result demonstrated that the proposed LSTM-ABC classifier gives enhanced results than other classifiers for accuracy, specificity, and sensitivity.

Geeitha and Thangamani (2021) proposed a Feature Weighted SMOTE (FWSMOTE) for solving data imbalanced problems and risk variable tests in cervical tumour classification. The information imputation issue is solved using mode and median missing information attribution. For best FS, Hilbert–Schmidt independence criterion with bacteria forage optimisation (HSICBFO) method is introduced to increase classification results. Ensemble SVM with interpolation classification is utilised for cervical cancer. Different measures are conveyed to evaluate the classification performance and provide precision, recall, specificity, F-Measure, accuracy, and G-mean values by 94.77%, 93.38%, 93.86%, 94.07%, 93.60%, and 93.62%, respectively, aid in distinguishing the feature stage of cervical carcinoma improvement and direction for additional analysis.

Adem et al. (2019) presented an automatic analysis of the cervical tumour. For this reason, a dataset collection involving 668 examples, 30 features, and 4 classes from the UCI corpus is applied in the test and learning phases. Softmax classifier with stacked autoencoder, one of the deep learning (DL) algorithms, is utilised to predict the databases. Initially, a stacked autoencoder is introduced to the original dataset, a decreased-dimension dataset. This dataset is assigned to training via using the softmax layer. In this stage, 70% (468) of the samples are utilised for learning, and the residual 30% (200) of the samples are utilised for testing. In the investigation, methods are used independently for 4 classes of the dataset, and their diagnosis results are analysed with existing classifiers. Softmax classifier with stacked autoencoder model is utilised over cervical cancer dataset, and it provides results more than other ML strategies with increased classification rate. Given the best ML strategies in cancer research, novel techniques are introduced for patient analytic support schemes.

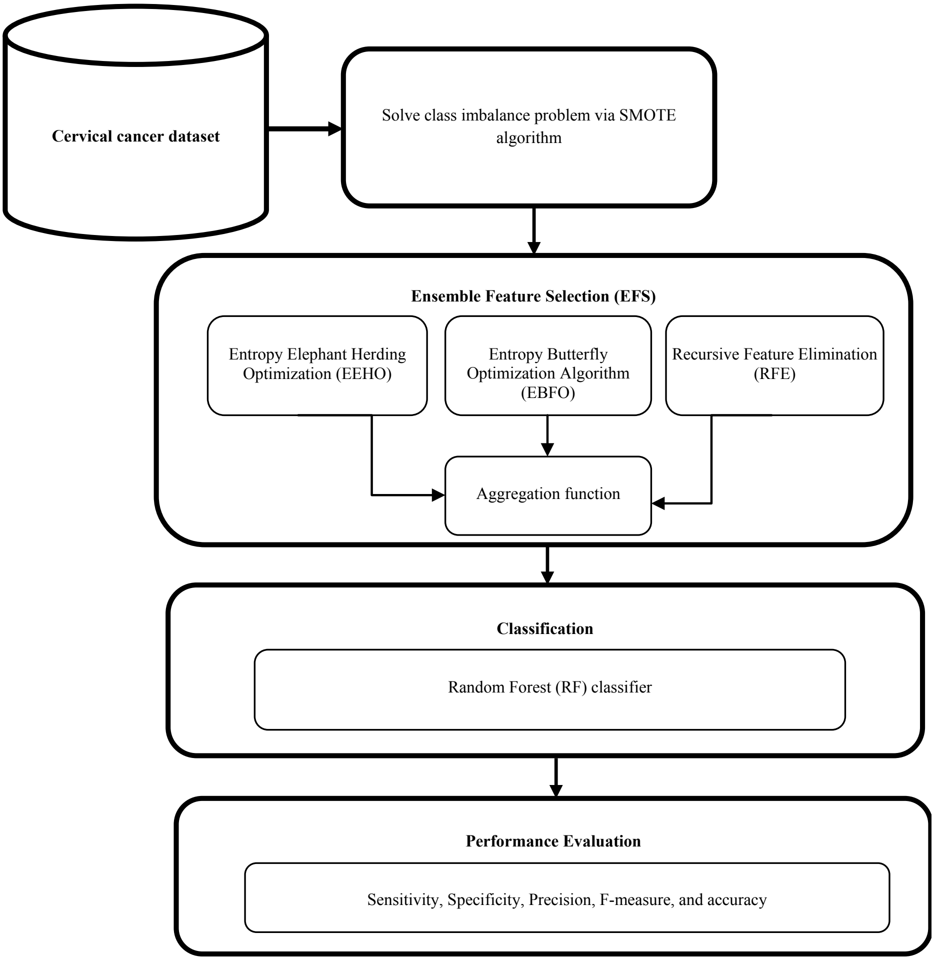
3 Proposed methodology

In this presented study, SMOTE method is applied to solve the class imbalance issue in the database by expanding the quantity of the minority class dependent on kNN to almost equivalent classes. Likewise, EEHO, EBFO, and RFE based FS methods to decrease the training period and discard the irrelevant attributes from the dataset in the classifier model. Then, the RF classifier strategy is utilised to classify the samples into positive and negative. The last phase is to evaluate the efficiency before and afterwards, employing the FS and SMOTE algorithms with improved results. At last, the performance of the model is estimated prior to and then afterwards SMOTE, then evaluated with other classifiers' results (Figure 1).

3.1 Dataset

The samples utilised in this work are collected from a repository of UCI (Fernandes et al., 2017). The database contained chronicled clinical archives, propensities, and demographic data for 858 patients, including 32 attributes for every patient.

Figure 1 Proposed cervical cancer diagnosis model (CCDM) framework


Cervical cancer dataset
Solve class imbalance problem via SMOTE algorithm
Ensemble Feature Selection (EFS)
Entropy Elephant Herding Optimization (EEHO)
Entropy Butterfly Optimization Algorithm (EBFO)
Recursive Feature Elimination (RFE)
Aggregation function
Classification
Random Forest (RF) classifier
Performance Evaluation
Sensitivity, Specificity, Precision, F-measure, and accuracy

3.2 Synthetic minority oversampling technique (SMOTE)

ML methods handle issues while one class influences the database, i.e., the amount of samples in that class goes beyond the number of previous classes. It's also termed an 'Imbalanced Database' that misleads the categorisation and changes the results as well. SMOTE is the technique that helps to solve the problem in terms of synthetically increasing the minority class based on kNN (Tarawneh et al., 2020) to dataset balance, and also it produces synthetic samples from the minority classes. The SMOTE tests strongly positively corresponded with the examples from the minority class used to produce them, and the SMOTE tests got utilising the same original samples. And it uses below equation (1) to increase the minority class,

$$x_{syn} = x_i + \left(x_{knn} - x_i\right) * t \tag{1}$$

The SMOTE procedure is described below:

1. Input attribute vector x_i, kNN x_{kNN} is applied to balance the dataset
2. Find the difference between the attribute vector and kNN
3. Multiplies the variance using an arbitrary value amongst 0 and 1
4. Includes the resultant value to attribute vector to classify a fresh sample on the row distribution
5. Repeat the steps from 1-4 for discovering new attribute vectors

3.3 Ensemble feature selection (EFS)

EFS schemes are defined to produce the best subgroup of attribute features by merging many FS depending on the EEHO. The EBFO and RFE are the intuition behind ensemble learning. The common design of EFS is to aggregate the decisions of FS methods to develop the representation capability (Ng et al., 2020). EFS techniques contain two main phases: generation of diverse feature selectors and aggregation of the decisions.

3.3.1 Entropy elephant herding optimisation (EEHO)

The EEHO algorithm is a meta-heuristic intellectual method depending on the travelling behaviour of elephants. By the examination and analysis of the elephants, the elephant herd mostly contains the subsequent 2 features for attribute selection from cervical cancer analysis. The primary characteristic is that multiple clans are there in an elephant herd that contains its patriarch and associates who pursue the directions of the patriarch for the optimal set of attributes from cervical cancer disease diagnosis, and it has no adult male elephant. In the growing stage, young elephants live alone from the elephants. The main purpose of EEHO contains two functions Clan updating and Separating (Elhosseini et al., 2019). The elephant herd's initial attribute could be distracted by the clan upgrading function, which is described by equation (2),

$$x_{n,i,j} = x_{i,j} + r * a * \left(x_{b,i} - x_{i,j}\right) * ECE_W \tag{2}$$

Where the old and new attribute locations from the cervical dataset of elephant j in clan i are $x_{i.i}$ and $x_{n.i.i}$, correspondingly; $\alpha \in [0,1]$ indicates a scaling variable; $x_{b.i}$ Denotes the attribute location along with the optimal variable (accuracy) in clan i and r is an arbitrary number along by a typical dispersion with range [0, 1]. Equation (2) addresses the warn interaction of more entities (features), yet the matriarch in every faction has not been restructured (Li et al., 2020). To compute the weight worth of every attribute of the cervical disease dataset in the EEHO, expect that while specific attribute variables are noticed, it provides a specific measure of data to the target class. Then, this weight is refreshed to the EEHO scheme. Depending on the weight of Expected Cross-Entropy (ECE), the significance of attributes is chosen. ECE is performed based on Kullback-Leiber (KL) distance, and it computes the distance between the likelihood of the target class and the likelihood of the target class in the position of a particular attribute. The processing equation (3) could be described as follows (Shang et al., 2016),

$$\text{Cross Entropy(CE)(f)} = P(f)\sum_{i=1}^{|C|} P\left(f|c_i\right)\log\frac{P(f \mid c_i)}{P(c_i)} \tag{3}$$

where f is the feature, $P(f)$ is the probability of the sample enclosed in the learning collection, $P(c_i)$ denotes the probability of class c_i in the training set, $P\left(f|c_i\right)$ refers to the chance of a sample that has featured in the class c_i, and $|C|$ denotes the overall amount of classes in the learning collection. The information entropy is given by equation (4),

$$\text{Information Entropy}\left(\text{IE}\right)\left(f\right) = -\sum_{i=1}^{|C|} P\left(f|c_i\right)\log(P(f \mid c_i)) \tag{4}$$

In summary, combining equations (3) and (4); the ECE equation is as follows by equation (5),

$$\text{ECE}\left(f\right) = \frac{P(f)}{IE\left(f\right)+\varepsilon}\sum_{i=1}^{|C|}\left(P\left(f|c_i\right)+\varepsilon\right)\log\frac{P\left(f|c_i\right)+\varepsilon}{P(c_i)} \tag{5}$$

If a feature exists only in a single class, the range of information entropy is zero; that is, $\text{IE}\left(f\right) = 0$. Hence must establish a small factor in the denominator as a regulator. So, the upgrading task of the matriarch for FS for cervical cancer detection is shown in equations (6) and (7).

$$x_{n,i,j} = \beta * x_{c,i} \tag{6}$$

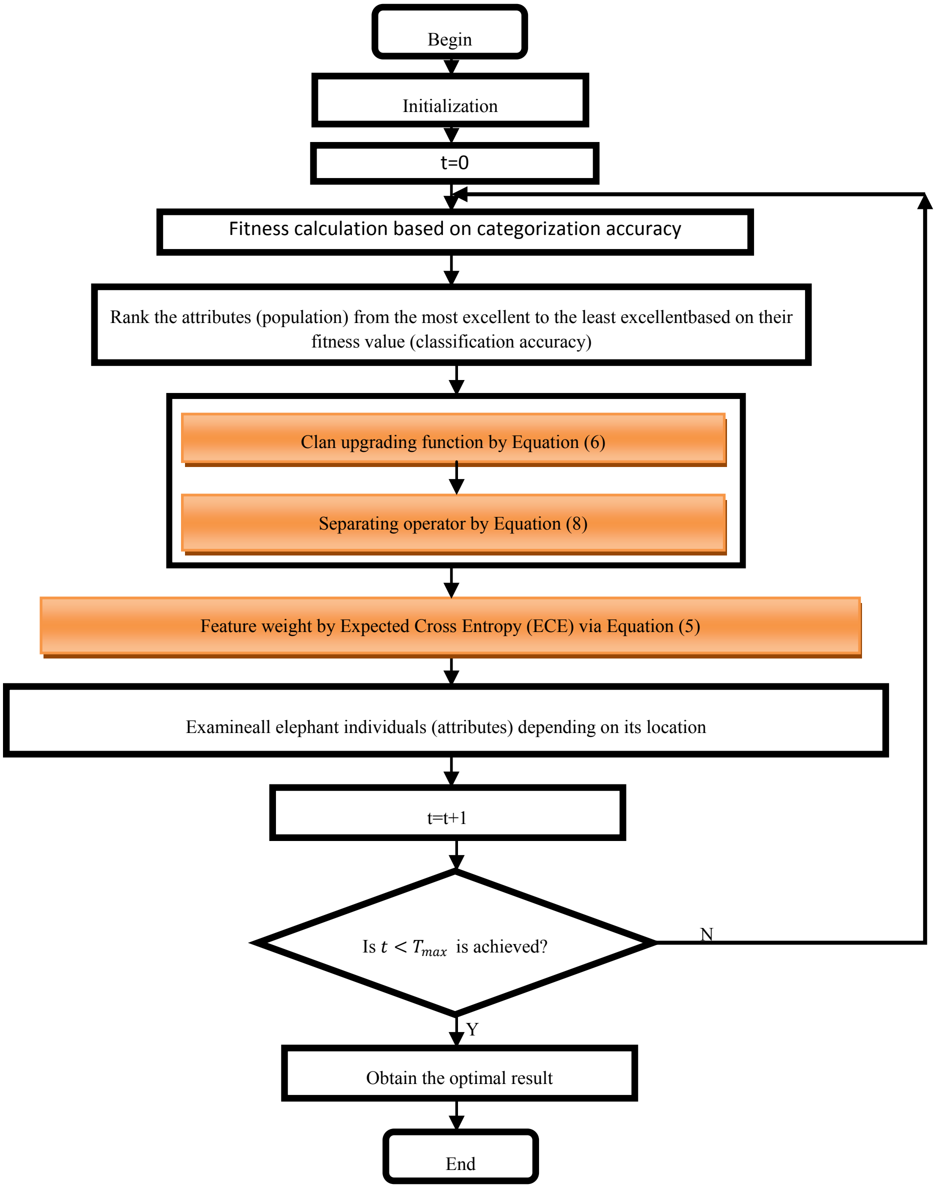
$$x_{c,i} = \frac{1}{n_i} \times \sum_{j=1}^{n_i} x_{i,j} \tag{7}$$

where $\beta \in [0, 1]$ indicates the scaling variable. The centre location (feature position) in a clan i is $x_{c.i}$ is able to be computed by equation (7). The elephant in a clan i is n_i. In equation (6), the revision of the matriarch location (feature position) is associated with the data of every member (features) in the clan. The separating function can be formed from the second attribute of the elephant herd. The splitting procedure is discussed in equation (8),

$$x_{w,i} = x_{min} + r * \left(x_{max} - x_{min}\right) \tag{8}$$

where $x_{w.i}$ is the location(feature position) through the lower fitness range (lesser categorisation efficiency) in clan i; x_{max} and x_{min} denote the top and bottom bound of the elephant's location (attribute location); correspondingly, $r \in [0, 1]$ denotes an arbitrary value generated via regular distribution. Algorithm 1 describes the functioning strategy of the EEHO method. It begins with initialising the populace through the number of attributes in the cervical disease dataset and afterwards assesses the fitness variable (prediction accuracy), depending on the discard of weak attributes (Cervical cancer attributes) in the clan; after that, at that point begin the technique with t emphasis to T_{max}. For every attribute, two activities, for example, clan upgrading and another splitting, are done via stages 6 to stage 8. When such tasks are done after that eliminates the most noticeable elephant from the clan through stage 11 to stage 13. Then, found the optimal attributes in stage 14. The workflow of the EEHO scheme is shown in Figure 2.

Figure 2 Workflow of entropy elephant herding optimisation (EEHO) algorithm (see online version for colours)


Begin
Initialization
t=0
Fitness calculation based on categorization accuracy
Rank the attributes (population) from the most excellent to the least excellentbased on their fitness value (classification accuracy)
Clan upgrading function by Equation (6)
Separating operator by Equation (8)
Feature weight by Expected Cross Entropy (ECE) via Equation (5)
Examineall elephant individuals (attributes) depending on its location
t=t+1
Is $t < T_{max}$ is achieved?
N
Y
Obtain the optimal result
End

Algorithm 1 Entropy elephant herding optimisation (EEHO) algorithm

```
Initialize the number of populations by the number of attributes and variables
Fitness calculation using categorization accuracy and their attribute location
While $t < T_{max}$
  For $i = 1 ton_c$
    For j=1 to $n_j$ (the number of elephants (attributes) in a single clan)
      Modify $x_{i,j}$ and produce $x_{n,i,j}$using equation (2), create attribute weight using ECE (f)inEquation (5)
      If $x_{i,j} = x_{b,i}$ then
        Modify $x_{i,j}$ and produce $x_{n,i,j}$using Equations (6&7)
      End if
    End for
  For $i = 1 ton_c$
    Swap the least excellent elephant (features) in a clan I using equation (8)
  End for
  Examine individuals(attributes) based on their fresh location
End while
```

3.3.2 Entropy butterfly optimisation algorithm (EBFO)

In this work, FS is performed by utilising the EBFO to choose the best attributes from the cervical disease database. EBFO is a novel approach that emulates food search (higher accuracy with chosen attributes) and mating butterfly activities for cervical cancer treatment. This EBFO scheme is primarily based on the searching technique of butterflies, which utilise their feeling of accuracy for ideal FS to decide the area of nectar accomplices (Arora and Singh, 2019; Tubishat et al., 2020). Depending on the technical perceptions, it is found that butterflies contain accurate intelligence of discovery on the basis of classification accuracy. A butterfly can create fragrance along with a little force that is associated with its fitness (accuracy); like a butterfly commencing with a specific region to another, its fitness can fluctuate consequently. In EBFO Algorithm, the entire idea of detecting and preparing the methodology depends on three significant terms, viz. tangible sense mode (c), stimulus strength (I), and energy exponent (a) for the best choice of attributes(Arora and Singh, 2019; Tubishat et al., 2020). In EBFO, I corresponds with the fitness (accuracy) for the determination of attributes from the cervical cancer dataset. Utilising these ideas, in the EBFO algorithm, the accuracy is formed as a component of I as subsequently via equation (9),

$$f = cI^{a} \tag{9}$$

where f denotes the actual fragrance strength, c represents the sense mode, which is computed based on categorisation accuracy, I indicates the stimulus strength, and a denotes the energy exponent is generated via fitness function. Hence, a and c are in the range [0,1]. Then again, if $a=0$, it implies that the fragrance produced through any butterfly cannot be detected via different butterflies. Hence, the boundary a organises the algorithm's behaviour. Another significant parameter is c the additionally vital parameter to discover the speed of convergence and how the EBFO method performs for the FS process. To show the above discussions as far as a hunting strategy, the above qualities of butterflies are idealised as the following:

1 All butterflies should produce a few fragrances, which empowers the butterflies (features) to draw in all others (features).

2 All butterflies can travel arbitrarily or toward the better butterfly emanating extra fragrance.

3 I of a butterfly is influenced through the scene of the objective function.

Three stages are there in EBFO, for example,

1 initialisation stage

2 iteration stage

3 last stage.

In all EBFO iterations, the initialisation stage is conducted initially; subsequently, the seeking of best attributes is done in an iterative way, and in the last stage, the approach is ended while the optimal solution selection is determined. Classification accuracy is calculated in the initialisation stage using EBFO and its solution space. The variables for the boundaries utilised in EBFO are likewise appointed. The places of butterflies (features) are arbitrarily produced in the FS search space by means of their fragrance and objective function (Arora and Singh, 2019; Tubishat et al., 2020). Subsequently, completing the initialisation stage begins the iteration stage. In every iteration, every butterfly in the FS decision region travels to fresh locations, and afterward, their prediction accuracy is assessed. The initial fitness values have been calculated for each butterfly in various situations in the solution space. Subsequently, those butterflies can produce aroma in their locations utilising equation (10). In the global hunting stage, the butterfly moves toward the fittest range (g*)(optimal features) that could be addressed utilising equation (10),

$$x_i^{t+1} = x_i^t + \left(r^2 \times g^* - x_i^t\right) \times f_i * ECE_W \tag{10}$$

Here, x_i^t is the result vector x_i. For ith butterfly in iteration number t. Where, g^* Addresses the current best-chosen feature solution established amongst every one of the resolutions in the present iteration. The odour of with butterfly is addressed through f_i and $r \in [0,1]$ is a random number restricted local search stage is addressed via equation (11),

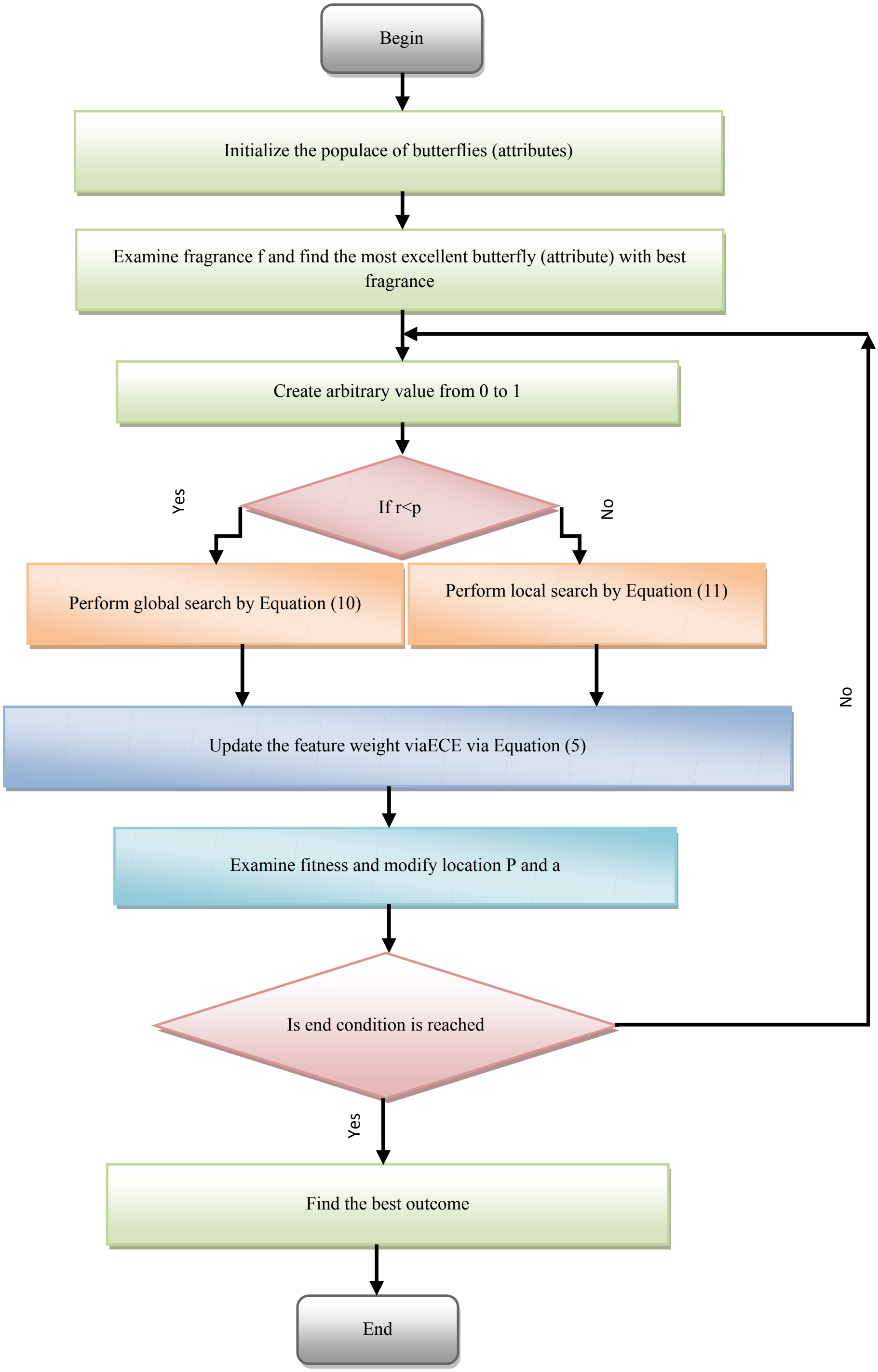
$$x_i^{t+1} = x_i^t + \left(r^2 \times x_j^t - x_k^t\right) \times f_i * ECE_W \tag{11}$$

Here x_i^t and x_k^t denote jth and kth butterflies from the FS solution region. When x_i^t and x_k^t include a location having an identical swarm, and $r \in [0,1]$ denotes the arbitrary value after that equation (11) returns into an arbitrary neighbourhood walk. Seeking for best features and mating partners through butterflies could appear on the global and local scale for optimal selection of attributes from the database. Switching probability p is utilised in EBFO to control among normal global results to exhaustive local results. Once the termination condition is not coordinated, the iteration stage is preceded. While the iteration stage is closed, the approach yields a better result established along with its best fitness. In equations (11 and 12), attribute weight is likewise added to EBFO to choose the best amount of attributes in the cervical cancer dataset. The general phases incorporated with the proposed EBFO scheme are explained in Algorithm 2. In Algorithm 2, primary populations are produced utilising the number of attributes in the cervical cancer database (Step 1), and afterward I_i at x_i (Step 2) is registered dependent on c and a (Step 3). Such variables are created through categorisation accuracy. Afterward, it commences with the termination condition (Step 4); for all butterflies in the database, the fragrance range is determined (Step 6). Then determine the better attribute over the populace (Step 8) and produce an arbitrary value r (Step 10). When $r < p$, the travel towards the most excellent butterfly using equation (11), else travel arbitrarily using equation (12). If $r < p$, travel in the direction of the better butterfly through equation (11), besides travel arbitrarily via equation (12). After that, modify a range (Step 17), and assess all the variables as per their fresh location (Step 18). At last, terminate the procedure (Step 19) (Figure 3).

Algorithm 2 Entropy butterfly optimisation algorithm (EBFO)

```
Input: Cervical cancer dataset

Objective function: Classifier accuracy,
$f(x), x = (x_1, x_2, …., x_{dim}) dim = no. of dimesnions$

Result: Best attributes

1. Create an early populace of $n$ butterflies $x_i = (i = 1,2, …, n)$using the attributes in
   the database
2. Intensity $I_i$ at $x_i$ is obtained by accuracy $f(x_i)$
3. Characterize $c$, $a$ and $p$
4. While termination condition is not achieved
5. For all butterflies$f$in the populace
6. Determine for $f$using equation (10) and generate weight via entropy by equation (5)
7. End for
8. Obtain the most excellent butterfly
9.   For all butterflies$f$ in the population
10. Produce arbitrary value r
11.     If$r < p$
12. Travel towards the most excellent butterfly (optimal features) by equation (10) and
    generate weight via entropy by equation (5)
13.      Else
14. Travel arbitrarily via Equation (11)
15.      End if
16.    End for
17. Modify the range of a
18.     Examine individuals(attributes) based on their fresh location
19.  End while
20. Obtain the most optimal attributes
```

Figure 3 Workflow of EBFO algorithm (see online version for colours)


Begin
Initialize the populace of butterflies (attributes)
Examine fragrance f and find the most excellent butterfly (attribute) with best fragrance
Create arbitrary value from 0 to 1
If r<p
Yes
No
Perform global search by Equation (10)
Perform local search by Equation (11)
Update the feature weight viaECE via Equation (5)
Examine fitness and modify location P and a
Is end condition is reached
No
Yes
Find the best outcome
End

3.3.3 Recursive feature elimination (RFE)

The RFE is a wrapper scheme, and it is a recursive strategy that grades attribute as indicated by a specific level of significance. It is a greedy optimisation method that is focused on discovering the standout performing attributes. This recursive approach of eliminating the weakest attributes proceeds until the necessary amount of attributes is attained (Huang et al., 2018; Chen et al., 2020). Features are sorted by moreover feature significance or coefficient feature of the classifier. Thus, the RFE performance deeply relies on the classifier utilised for ranking the attributes. RFE needs a predefined amount of attributes to stay, but it is typically not recognised at earlier how many attributes are substantial (Algorithm 3). To determine the best amount of attributes, classification accuracy is utilised with RFE to achieve diverse attribute subsets and choose the optimal scoring feature collection. Figure 4 depicts the flow of the RFE scheme.

Algorithm 3 Recursive feature elimination (RFE)

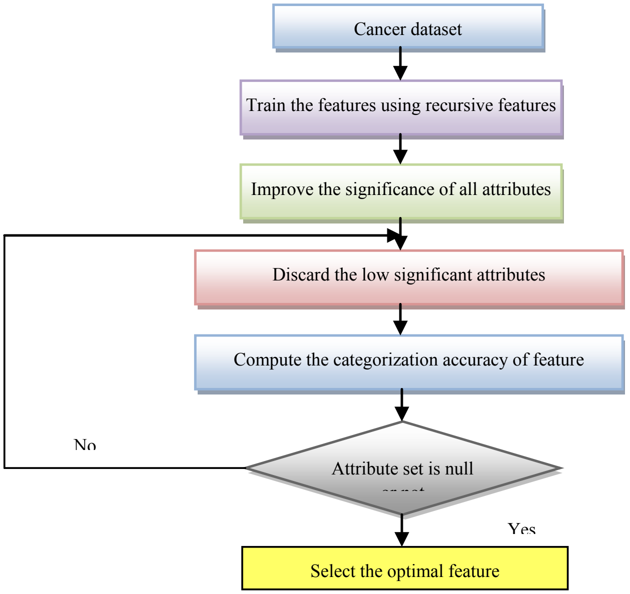
```
INPUT: Training data, set of $n$ features, ranking method $M(T, F)$

OUTPUT: The selected set of features $S_F$, a final ranking $R$

    Train the model
    Compute the model performance
    Compute the variable rankings
    For each subset size $S_i$, i=1,…,S do
    Repeat for $i$ in $\{1{:}p\}$
    Keep the $S_i$ most important variable
    Eliminate the least important feature
    Rank set $F$ using $M(T, F)$
    Train the model by $S_i$ features
    $f^*$ ←last ranked feature in F

    $R(p - i + 1) \leftarrow f^*$
    $F \leftarrow F - f^*$
    End for
    Re-compute the rankings for each feature
    End
    Compute the performance over the $S_i$
```

During all iterations, feature significance is calculated, and the least significant attribute is neglected. The other opportunity is to omit a subset of attributes in all iterations to accelerate the procedure.

3.3.4 Aggregation function

The various ranked results are consolidated into a single ensemble list by utilising an appropriate aggregation function which allocates every feature a 'general score' depending on the attribute's position (rank) in the first samples. Generally, allow L_k is denoted as the sorted catalogue results from the function of a specified FS method to the *k*th bootstrap test (*k* = 1, …, *B*). For every one of the actual attributes, $f_i (i = 1, \ldots, N)$ a general rank is subsequently determined via $score_i = score(f_i) = aggr(r_{i1}, r_{i2}, \ldots r_{iB})$. Here r_{ik} is denoted as ith attribute rank in the kth sorted, and aggregation function is specified as $aggr$. Depending on their general ranks, the attributes are arranged from the largely relevant to the most irrelevant features in the resultant ensemble result (Wald et al., 2012).

Figure 4 Flow diagram of RFE algorithm (see online version for colours)


Cancer dataset
Train the features using recursive features
Improve the significance of all attributes
Discard the low significant attributes
Compute the categorization accuracy of feature
No
Attribute set is null
Yes
Select the optimal feature

3.4 RF classification

In the RF, the CART method is used to expand many DTs according to the bootstrap merging (bagging) method (Galletta, 2016). The CART method is to discover the proper categorisation of many related attributes (*y*) and many non-related attributes (*x*) (Lee et al., 2020; Alam et al., 2019). In RF, all trees arbitrarily elect a division of the database to create a self-determining DT. It depends on the majority of trees attained by voting. The building of RF can be discussed below:

Step 1: Creates *N* amount of bootstrap examples from the given database.

Step 2: Every node gets an arbitrary example of features with *m* size, such that $m < M$ (*M* is denoted as the overall amount of attributes).

Step 3: Builds an opening via the m features chosen in Step 2 and computes the *k* node through the optimal division point. ('*k*' is denoted as the next node).

Step 4: A replicate splitting tree awaiting simply a single leaf node is attained, and the tree is terminated.

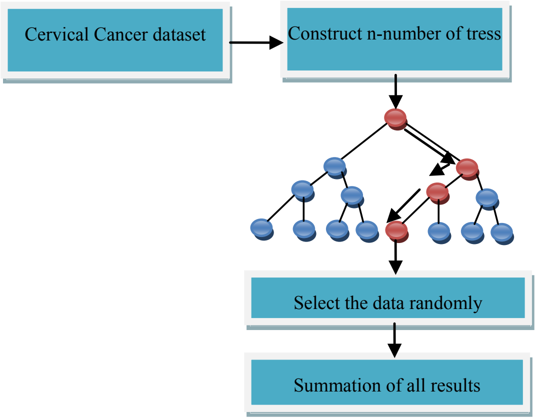
Step 5: This procedure is executed on all bootstrapped examples independently.

Step 6: Trees used a categorisation voting to acquire the categorisation information from the (n) trained trees.

Step 7: Maximum voted attributes are used to construct the last RF system.

Figure 5 refers to the structure of the RF classifier algorithm for a given cervical cancer dataset. RF helps to define the tree with many decisions at the training base and mode of class's output by tree. RF is basically a grouping of tree predictors that every tree relies on, and the random vector values are illustrated individually along with a similar level of distribution. Single DTs are suitable due to their high variance and bias (Alam et al., 2019). Instead of splitting a tree node using all features, it selects nodes from each tree, and it's only used as candidates to discover the optimal partition for the node. The inspiration following this two-phase randomisation is to de-connect trees, and thus, forest ensembles hold lower variance, a bagging scheme. Consequently, it is focused on reducing the variance. RF trees are demonstrated with the purpose of huge sample reliability needs terminal nodes along with huge sample sizes (Zhou et al., 2016); empirically, it is seen that the purpose of purity is frequently effectual, whereas the feature space is huge or the lesser number of samples. Deep trees are grown without pruning; in such cases, it promotes low bias when aggregation decreases variance and lower error (Zhou et al., 2016). Samples are chosen depending on the aggregation function, and their features are fed into the RF classifier. Lastly, the RF classifier is successfully classifying the cervical cancer output over the known dataset.

Figure 5 Structure of RF classifier algorithm (see online version for colours)



4 Results and discussion

This part discusses the outcomes of several classifiers experimented on the cervical tumour database from UCI (Fernandes et al., 2017), which contains past clinical samples, behaviour, and statistical data for 858 patients, along with 32 attributes for every patient. Missing values are replaced via the observation's values for neighbouring data samples. It recognises the neighbouring samples via distance computation, and the missing values can be found through the finished values of neighbouring observations. In the kNN algorithm, the main purpose is to distinguish 'k' samples in the given dataset which are related or nearby values over the space. Subsequently, '*k*' samples are used to calculate the missing values. Every sample's missing points can be imputed through the '*k*'-neighbour's mean value established in the given dataset. In the existence of missing data points, the Euclidean distance is computed by avoiding the missing data points and scaling up the weight of the non-missing data points, and it is calculated via equation (12),

$$d_{xy} = \sqrt{weight * squared\ distance\ from\ current\ coordinates} \tag{12}$$

$$weight = \frac{Overall\ amount\ of\ coordinates}{Amount\ of\ current\ coordinates} \tag{13}$$

4.1 Evaluation metrics

The results of all the classifiers are measured via the metrics like precision, sensitivity, specificity, f-measure, and accuracy, as shown by equations (14)–(18). Precision refers to the ratio of positive samples that have been accurately identified as positive. Recall or sensitivity described that the positive examples are selected to the complete amount of positive examples. Specificity is important to identify the suitable and unsuitable results done through the particular classifier. F-measure is described as the harmonic average of recall and precision. Accuracy is evaluated, which is assumed to be one of the extensively regarded measures used to study the performance of classifiers.

$$\text{Recall}(\text{R}) / \text{ Sensitivity}(\text{Sen}) = \frac{\text{TP}}{\text{TP} + \text{TN}} \tag{14}$$

$$\text{Specificity}(\text{Spe}) = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{15}$$

$$\text{Precision}(\text{P}) = \frac{TP}{FP + TP} \tag{16}$$

$$\text{F} - \text{measure} = \frac{2.\ \text{PR}}{P + R} \tag{17}$$

$$\text{Accuracy}(\text{Acc}) = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{18}$$

where TP is True Positive, TN is True Negative, FP is False Positive, and FN is False Negative.

4.2 Simulation experiment

The given dataset is imbalanced; the number of cancer files is smaller than the number of general files; thus SMOTE technique is applied to balance the amount of both classes. Classifications like SVM and RF algorithms are applied to examine the results of FS schemes and categorise the samples into either cancer patients or non-cancer patients. The experimentation is performed before and after SMOTE with the three FS algorithms.

Table 1 Results performance comparison of cervical cancer datasets vs. metrics

Imbalanced algorithm	FS+ Classifiers	*Hinselmann test results (%)*				
		Precision	*Recall*	*F-measure*	*Specificity*	*Accuracy*
Before SMOTE	SVM + RFE	61.0581	84.4107	70.9503	80.6707	86.8298
	SVM + PCA	61.1065	87.6457	71.6946	81.4620	87.6457
	SVM + EFS	61.1924	88.4615	72.4380	84.3458	88.4615
	RF + RFE	62.4831	89.2774	72.9300	84.4107	89.2774
	RF + PCA	63.7112	89.9767	74.8373	85.3066	89.9767
	RF + EFS	66.1292	91.9580	76.8447	89.0314	91.9580
After SMOTE	SVM + RFE	61.6768	85.9000	71.8004	84.0907	88.6946
	SVM + PCA	64.1042	87.1151	73.8589	85.9000	91.0256
	SVM + EFS	65.3877	90.3975	76.3316	87.1151	91.1246
	RF + RFE	66.0537	93.9542	77.1103	87.3992	92.0746
	RF + PCA	70.2417	94.2267	80.4853	87.6527	93.4581
	RF + EFS	71.8750	97.2661	82.6647	90.4801	94.7552
Imbalanced algorithm	FS+ Classifiers	*Schiller test results (%)*				
		Precision	*Recall*	*F-measure*	*Specificity*	*Accuracy*
Before SMOTE	SVM + RFE	61.1281	84.4241	70.9651	80.4734	86.1198
	SVM + PCA	61.1924	86.5089	72.4380	81.9325	87.1741
	SVM + EFS	69.1510	88.0774	76.8621	82.0511	88.7892
	RF + RFE	71.3677	88.2050	78.8470	84.4107	89.3939
	RF + PCA	71.6783	89.0314	79.0875	86.5343	89.6270
	RF + EFS	75.2701	93.0243	83.2107	89.1224	91.7249
After SMOTE	SVM + RFE	61.6768	85.9000	71.8004	85.2154	88.1147
	SVM + PCA	64.1042	87.1151	73.8589	87.0329	91.0171
	SVM + EFS	75.0350	93.5604	83.5011	87.1151	91.4917
	RF + RFE	77.5046	93.7259	84.8469	87.1869	93.0070
	RF + PCA	79.1498	94.1206	85.7539	87.5540	93.4182
	RF + EFS	80.6343	95.7788	87.5565	89.0965	94.5221
Imbalanced algorithm	FS+ Classifiers	*Citology test results (%)*				
		Precision	*Recall*	*F-measure*	*Specificity*	*Accuracy*
Before SMOTE	SVM + RFE	61.0581	84.4107	70.9503	79.7383	86.8298
	SVM + PCA	61.1924	85.7187	72.4380	81.6810	87.1795
	SVM + EFS	62.7268	87.8071	72.4422	82.7667	88.4615
	RF + RFE	65.1347	88.9742	75.2105	83.9381	89.2774
	RF + PCA	65.7271	89.0314	75.9971	84.4107	89.6841
	RF + EFS	66.9874	90.2334	76.0549	89.0314	91.1422

Table 1 Results performance comparison of cervical cancer datasets vs. metrics (continued)

Imbalanced algorithm	FS+ Classifiers	*Citology test results (%)*				
		Precision	*Recall*	*F-measure*	*Specificity*	*Accuracy*
After SMOTE	SVM + RFE	61.6768	85.9000	71.8004	83.5667	88.6946
	SVM + PCA	64.1042	87.1151	73.8589	84.5952	90.9091
	SVM + EFS	67.0756	89.8342	76.8044	85.9000	91.0256
	RF + RFE	70.4175	90.9398	79.3736	86.5093	93.1170
	RF + PCA	73.1006	92.9975	82.1299	87.0148	93.7647
	RF + EFS	74.7429	93.7039	82.8769	87.6564	94.8718
Imbalanced algorithm	FS+ Classifiers	*Biopsy test results (%)*				
		Precision	*Recall*	*F-measure*	*Specificity*	*Accuracy*
Before SMOTE	SVM + RFE	61.4814	84.4107	72.9460	70.9503	86.1198
	SVM + PCA	61.9712	89.0314	75.5013	72.4380	87.7622
	SVM + EFS	66.6700	90.2491	78.4595	76.9298	88.2615
	RF + RFE	68.2305	90.9215	79.5760	78.2545	89.1214
	RF + PCA	69.0640	91.7310	80.3975	79.0713	89.8601
	RF + EFS	70.3573	92.8892	81.6232	79.2241	91.2587
After SMOTE	SVM + RFE	61.6768	85.9232	73.8000	85.9000	88.1231
	SVM + PCA	64.1042	87.1151	75.6096	86.1073	91.0256
	SVM + EFS	71.2841	92.5654	81.9247	86.4896	91.6084
	RF + RFE	72.4068	92.9721	82.6894	86.8372	92.3077
	RF + PCA	72.4998	93.3452	82.9225	87.1151	92.4242
	RF + EFS	76.3681	95.2553	85.8117	88.6096	94.2890

4.2.1 Hinselmann test

In this test, the RF previous to SMOTE attained an overall accuracy of 91.9580% for the EFS algorithm. RF with SMOTE algorithm attained overall accuracy of 94.7552% with the EFS algorithm. The SMOTE improved the accuracy by 2.7952%, sensitivity was improved from 91.9580% to 97.2661%, specificity was improved from 89.0314% to 90.4801%, precision was improved from 66.1292% to 71.8750%, and F-measure was improved by 76.8447% to 82.6647% as shown in Table 1.

Figures 6 and 7 show the metrics results of the comparison of classifiers before and after SMOTE for the Hinselmann test. In Figure 6, the proposed algorithm gives an enhanced accuracy value of 91.9580%, the other classifiers like SVM + RFE, SVM + PCA, SVM + EFS, RF + RFE, and RF + PCA give the accuracy value of 86.8298%, 87.6457%, 88.4615%, 89.2774%, and 89.9767% respectively(See Table 1). The proposed system has a 5.1282%, 4.3123%, 3.4965%, 2.6806%, and 1.9813% increased accuracy value when compared to SVM with RFE, PCA and EFS, RF with RFE, and RF with PCA methods respectively(See Table 1).

Figure 6 Hinselmann test results before SMOTE vs. classifiers (see online version for colours)

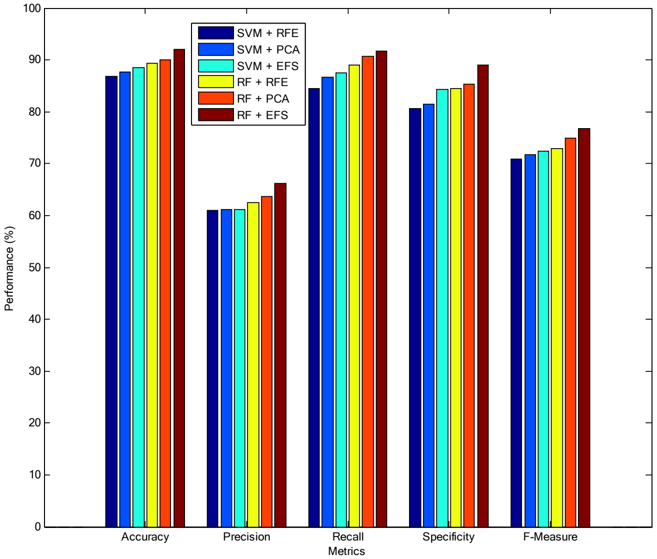
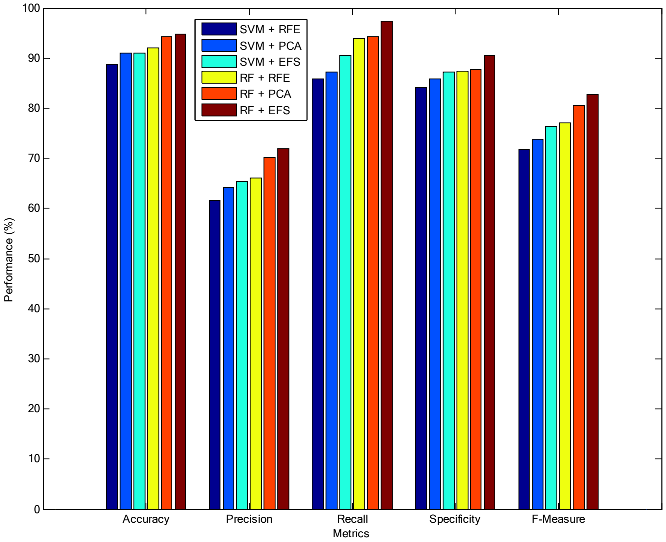


Figure 7 Hinselmann test results after SMOTE vs. classifiers (see online version for colours)



Figures show the overall metrics results from a comparison among SMOTE and non-SMOTE methods against classifiers. Figure 7 results show that the proposed algorithm gives an enhanced accuracy value of 94.7552%, and the other classifiers like SVM + RFE, SVM + PCA, SVM + EFS, RF + RFE, and RF + PCA give the accuracy value of 88.6946%, 91.0256%, 91.1246%, 92.0746%, and 93.4581% respectively (See Table 1) (See Table 1). The proposed system has a 6.0606%, 3.7296%, 3.6306%, 2.6806%, and 1.2971% increased accuracy value when compared to SVM with RFE, PCA and EFS, RF with RFE, and RF with PCA methods respectively(See Table 1).

4.2.2 Schiller test

In this test, the RF before SMOTE attained overall accuracy of 91.7249% for the EFS algorithm. RF with SMOTE algorithm attained overall accuracy of 94.5221% with the EFS algorithm. The SMOTE improved the accuracy by 2.7972%, sensitivity was improved from 93.0243% to 95.7788%, specificity was improved from 89.1224% to 89.0965%, precision was improved from 75.2701% to 80.6343%, and F-measure was improved by 83.2107% to 87.5565% as shown in Table 1.

Important metrics results from a comparison of classifiers before and after SMOTE for the Schiller test are illustrated in Figures 8 and 9. The proposed algorithm gives an enhanced accuracy value of 91.7249%, and the other classifiers like SVM + RFE, SVM + PCA, SVM + EFS, RF + RFE, and RF + PCA give the accuracy value of 86.1198%, 87.1741%, 88.7892%, 89.3939%, and 89.6270% respectively(See Table 1). The proposed system has a 5.6051%, 4.5508%, 2.9357%, 2.3310%, and 2.0979% increased accuracy value when compared to SVM with RFE, PCA and EFS, RF with RFE, and RF with PCA methods respectively(See Table 1).

Figure 8 Schiller test results before SMOTE vs. Classifiers (see online version for colours)

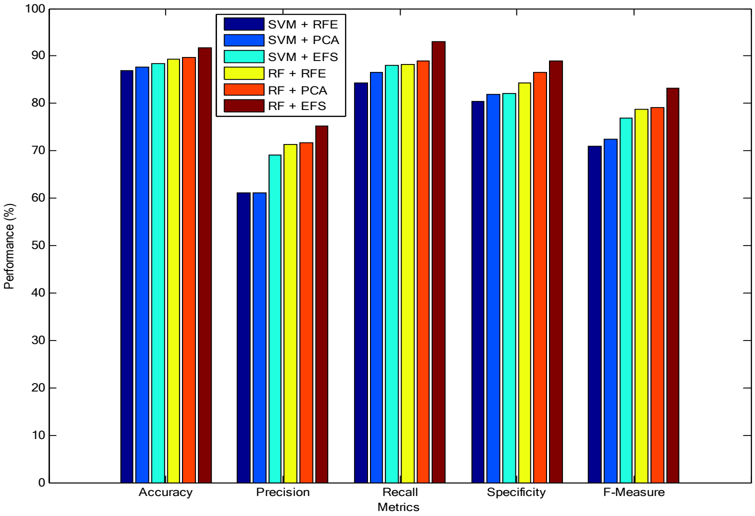
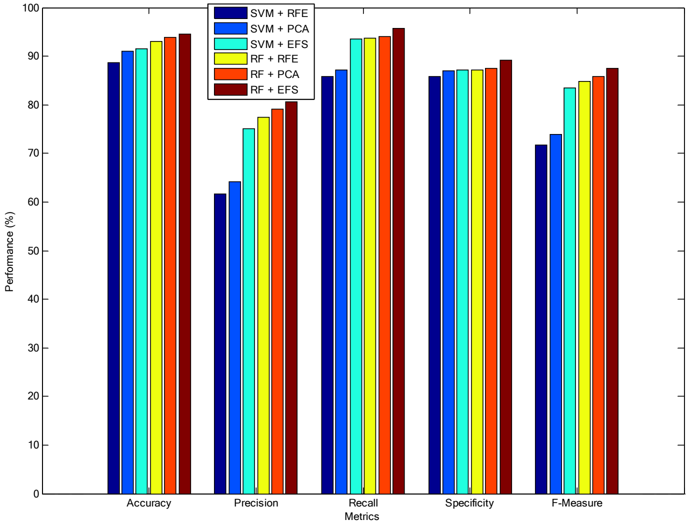


Figure 9 shows that the proposed algorithm gives an enhanced accuracy value of 94.5221%, the other classifiers like SVM + RFE, SVM + PCA, SVM + EFS, RF + RFE, and RF + PCA give the accuracy value of 88.1147%, 91.0171%, 91.4917%, 93.0070%, and 93.4182% respectively (See Table 1). The proposed system has a 6.4074%, 3.5050%, 3.0304%, 1.5151%, and 1.1039% increased accuracy value when compared to SVM + RFE, SVM + PCA, SVM + EFS, RF + RFE, and RF + PCA methods respectively (See Table 1).

Figure 9 Schiller test results after SMOTE vs. Classifiers (see online version for colours)



4.2.3 Citology test

In this test, the RF before SMOTE attained overall accuracy of 91.1422% for the EFS algorithm. RF with SMOTE algorithm attained overall accuracy of 94.8718% with the EFS algorithm. The SMOTE improved the accuracy by 3.7296%; sensitivity was improved from 90.2334% to 93.7039%, specificity was improved from 89.0314% to 87.6564%, precision was improved from 66.9874% to 74.7429%, and F-measure was improved by 76.0549% to 82.8769% as shown in Table 1.

Overall metrics results from classifiers comparison before and after SMOTE for Cytology test are illustrated in Figures 10 and 11. The proposed algorithm gives an enhanced accuracy value of 91.1422%, and the other classifiers like SVM + RFE, SVM + PCA, SVM + EFS, RF + RFE, and RF + PCA give the accuracy value of 86.8298%, 87.1795%, 88.4615%, 89.2774%, and 89.6841% respectively(See Table 1). The proposed system has a 4.3124%, 3.9627%, 2.6807%, 1.8648%, and 1.4581% increased accuracy value when compared to SVM with RFE, PCA and EFS, RF with RFE, and RF with PCA methods respectively(See Table 1).

Figure 11 shows that the proposed algorithm gives an enhanced accuracy value of 94.8718%, the other classifiers like SVM + RFE, SVM + PCA, SVM + EFS, RF + RFE, and RF + PCA give the accuracy value of 88.6946%, 90.9091%, 91.0256%, 93.1170%, and 93.7647% respectively(See Table 1). The proposed system has a 6.1772%, 3.9627%, 3.8462%, 1.7548%, and 1.1071% increased accuracy value when compared to SVM + RFE, SVM + PCA, SVM + EFS, RF + RFE, and RF + PCA methods respectively (See Table 1).

4.2.4 Biopsy test

In this test, the RF before SMOTE attained overall accuracy of 91.2587% for the EFS algorithm. Following SMOTE algorithm, the RF classifier attained overall accuracy of 94.2890% with the EFS algorithm. The SMOTE improved the accuracy by 3.0303%;

sensitivity was improved from 92.8892% to 95.2553%, specificity was improved from 79.2241% to 88.6096%, precision was improved from 70.3573% to 76.3681%, and F-measure was improved by 81.6232% to 85.8117% as shown in Table 1.

Figure 10 Citology test results before SMOTE vs. Classifiers (see online version for colours)

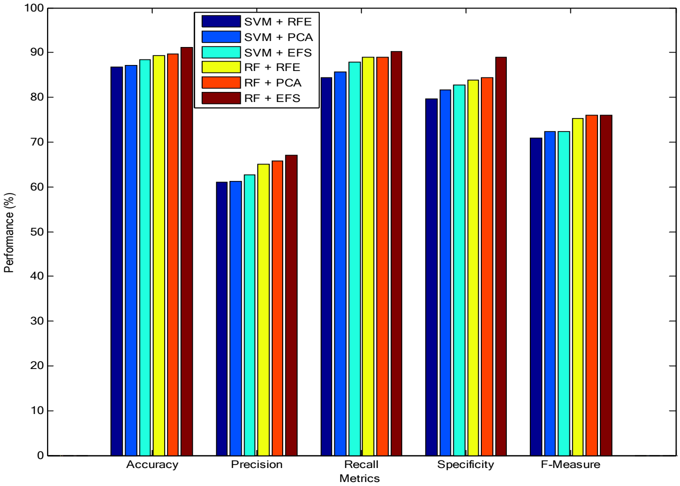
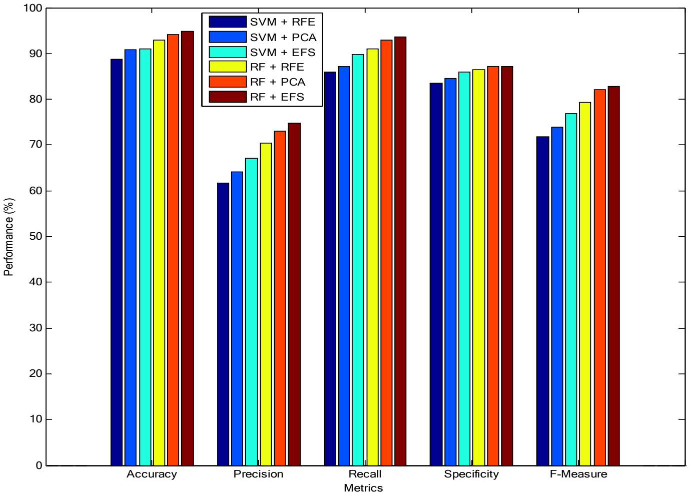


Figure 11 Citology test results after SMOTE vs. Classifiers (see online version for colours)



Overall metrics results from classifiers comparison before and after SMOTE for biopsy test are illustrated in Figures 12 and 13. The proposed algorithm gives an enhanced accuracy value of 91.2587%, and the other classifiers like SVM + RFE, SVM + PCA, SVM + EFS, RF + RFE, and RF + PCA give the accuracy value of 86.1198%, 87.7622%, 88.2615%, 89.1214%, and 89.8601% respectively (See Table 1). The proposed system has a 5.1389%, 3.4965%, 2.9972%, 2.1373%, and 1.3986% increased accuracy value when compared to SVM with RFE, PCA and EFS, RF with RFE, and RF with PCA methods respectively(See Table 1).

Figure 12 Biopsy test results before SMOTE vs. Classifiers (see online version for colours)

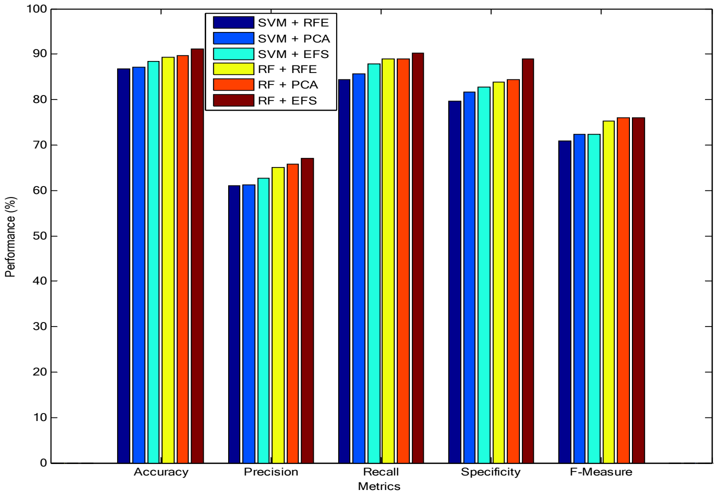


Figure 13 Biopsy test results after SMOTE vs. Classifiers (see online version for colours)

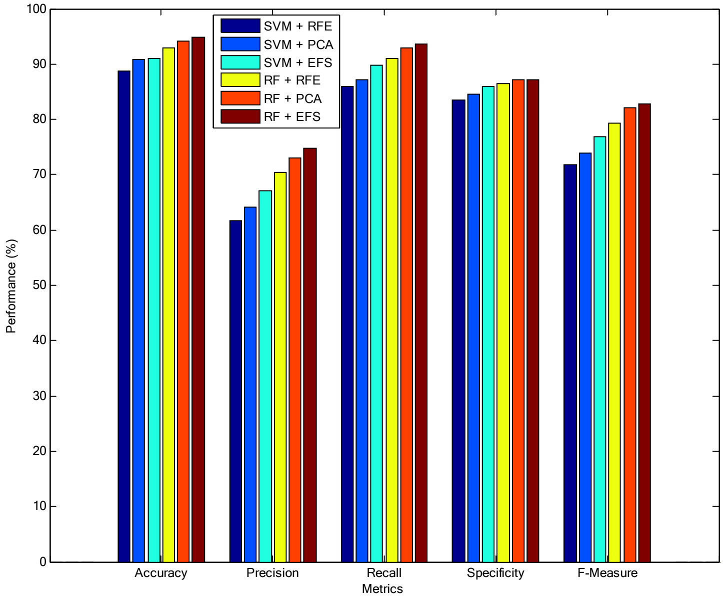


Figure 13 shows that the proposed algorithm gives an enhanced accuracy value of 94.2890%, and the other classifiers such as SVM with RFE, PCA and EFS, RF with RFE, and RF with PCA give the accuracy value of 88.1231%, 91.0256%, 91.6084%, 92.3077%, and 92.4242% respectively(See Table 1). The proposed system has a 6.1659%, 3.2634%, 2.6806%, 1.9813%, and 1.8648% increased accuracy value when compared to SVM with RFE, PCA and EFS, RF with RFE, and RF with PCA methods respectively(See Table 1). From the results, all the tests after SMOTE with the proposed algorithm give higher results for all the metrics when compared to before SMOTE algorithm.

5 Conclusion and future work

This study focuses on diagnosing cervical cancer, hence the missing values, irrelevant attributes, and imbalanced samples. SMOTE is proposed to fix unbalanced data. EFS recognise cervical cancer patients' most important characteristics. EFS combine EEHO, EBFO, and RFE to get better results than a single FS approach. ECE determines to attribute weight in BFO and EHO. ECE uses KL distance to duplicate the distance between the target class possibility and the target class on a certain attribute condition. RFE used an FS method that used RF to order qualities and eliminate the weakest ones. The suggested classification method uses feature subsets produced using an aggregation function. The optimal attribute subset depends on classification performance and diagnosis efficiency. RF classifier is recommended for classification and RF methods; CART is applied to many DTs depending on the bagging method. SMOTE improves the proposed classification's overall efficacy compared to other categories. The SMOTE-RF-EFS approach improves cervical cancer test precision, recall, specificity, f-measure, and accuracy. Future work could use numerous approaches to solve the imbalanced issue and various classification methods, especially ensemble schemes, to boost model efficiency.

References

Abdoh, S.F., Rizka, M.A. and Maghraby, F.A. (2018) 'Cervical cancer diagnosis using random forest classifier with SMOTE and feature reduction techniques', *IEEE Access*, Vol. 6, pp.59475–59485.

Adem, K., Kiliçarslan, S. and Cömert, O. (2019) 'Classification and diagnosis of cervical cancer with stacked autoencoder and softmax classification', *Expert Systems with Applications*, Vol. 115, pp.557–564.

Ahishakiye, E., Wario, R., Mwangi, W. and Taremwa, D. (2020) 'Prediction of cervical cancer basing on feature factors using ensemble learning', *2020 IST-Africa Conference (IST-Africa)*, Africa, pp.1–12.

Alam, M.Z., Rahman, M.S. and Rahman, M.S. (2019) 'A random forest based predictor for medical data classification using feature ranking', *Informatics in Medicine Unlocked*, Vol. 15, pp.1–12.

Arbyn, M., Weiderpass, E., Bruni, L., de Sanjosé, S., Saraiya, M., Ferlay, J. and Bray, F. (2020) 'Estimates of incidence and mortality of cervical cancer in 2018: a worldwide analysis', *The Lancet Global Health*, Vol. 8, No. 2, e191–e203.

Arora, J., Agrawal, U., Tiwari, P., Gupta, D. and Khanna, A. (2020) 'Ensemble feature selection method based on recently developed nature-inspired algorithms', *International Conference on Innovative Computing and Communications*, Springer, Singapore, pp.457–470.

Arora, S. and Singh, S. (2019) 'Butterfly optimization algorithm: a novel approach for global optimization', *Soft Computing*, Vol. 23, No. 3, pp.715–734.

Bansal, A. and Jain, A. (2021) 'Comparison of meta-heuristic with evolutionary and local search methods for feature selection', *Metaheuristic and Evolutionary Computation: Algorithms and* Applications, Springer, Singapore, pp.529–554.

Bedell, S.L., Goldstein, L.S., Goldstein, A.R. and Goldstein, A.T. (2020) 'Cervical cancer screening: past, present, and future', *Sexual Medicine Reviews*, Vol. 8, No. 1, pp.28–37.

Bouvard, V., Wentzensen, N., Mackie, A., Berkhof, J., Brotherton, J., Giorgi-Rossi, P., Kupets, R., Smith, R., Arrossi, S., Bendahhou, K. and Canfell, K. (2021) 'The IARC perspective on cervical cancer screening', *New England Journal of Medicine*, Vol. 385, No. 20, pp.1908–1918.

Brahim, A.B. and Limam, M. (2018) 'Ensemble feature selection for high dimensional data: a new method and a comparative study', *Advances in Data Analysis and Classification*, Vol. 12, No. 4, pp.937–952.

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R.L., Torre, L.A. and Jemal, A. (2018) 'Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries', *CA: A Cancer Journal for Clinicians*, Vol. 68, No. 6, pp.394–424.

Brisson, M., Kim, J.J., Canfell, K., Drolet, M., Gingras, G., Burger, E.A., and Hutubessy, R. (2020) 'Impact of HPV vaccination and cervical screening on cervical cancer elimination: a comparative modelling analysis in 78 low-income and lower-middle-income countries', *The Lancet*, Vol. 395, No. 10224, pp.575–590.

Chen, Q., Meng, Z. and Su, R. (2020) 'WERFE: a gene selection algorithm based on recursive feature elimination and ensemble strategy', *Frontiers in Bioengineering and Biotechnology*, Vol. 8, p.496.

Elhosseini, M.A., El Sehiemy, R.A., Rashwan, Y.I. and Gao, X.Z. (2019) 'On the performance improvement of elephant herding optimization algorithm', *Knowledge-Based Systems*, Vol. 166, pp.58–70.

Fernandes, K., Cardoso, J.S. and Fernandes, J. (2017) 'Transfer learning with partial observability applied to cervical cancer screening', *Iberian Conference on Pattern Recognition and Image Analysis*, Springer, Cham, pp.243–250.

Galletta, S. (2016) 'On the determinants of happiness: a classification and regression tree (CART) approach', *Applied Economics Letters*, Vol. 23, No. 2, pp.121–125.

Geeitha, S. and Thangamani, M. (2021) 'Integrating HSICBFO and FWSMOTE algorithm-prediction through risk factors in cervical cancer', *Journal of Ambient Intelligence and Humanized Computing*, Vol. 12, No. 3, pp.3213–3225.

Huang, X., Zhang, L., Wang, B., Li, F. and Zhang, Z. (2018) 'Feature clustering based support vector machine recursive feature elimination for gene selection', *Applied Intelligence*, Vol. 48, No. 3, pp.594–607.

Jain, R., Sangwan, S.R., Bachhety, S., Garg, S. and Upadhyay, Y. (2019) 'Optimized model for cervical cancer detection using binary cuckoo search', *Recent Patents on Computer Science*, Vol. 12, No. 4, pp.293–303.

Khan, I., Nam, M., Kwon, M., Seo, S.S., Jung, S., Han, J.S., Hwang, G.S. and Kim, M.K. (2019) 'LC/MS-based polar metabolite profiling identified unique biomarker signatures for cervical cancer and cervical intraepithelial neoplasia using global and targeted metabolomics', *Cancers*, Vol. 11, No. 4, pp.1–20.

Lee, T.H., Ullah, A. and Wang, R. (2020) 'Bootstrap aggregating and random forest', *Macroeconomic Forecasting in the Era of Big Data*, Springer, Cham, pp.389–429.

Li, J., Lei, H., Alavi, A.H. and Wang, G.G. (2020) 'Elephant herding optimization: variants, hybrids, and applications', *Mathematics*, Vol. 8, No. 9, pp.1–25.

Lu, J., Song, E., Ghoneim, A. and Alrashoud, M. (2020) 'Machine learning for assisting cervical cancer diagnosis: an ensemble approach', *Future Generation Computer Systems*, Vol. 106, pp.199–205.

Ng, W.W., Tuo, Y., Zhang, J. and Kwong, S. (2020) 'Training error and sensitivity-based ensemble feature selection', *International Journal of Machine Learning and Cybernetics*, Vol. 11, No. 10, pp.2313–2326.

Nithya, B. and Ilango, V. (2019) 'Evaluation of machine learning based optimized feature selection approaches and classification methods for cervical cancer prediction', *SN Applied* Sciences, Vol. 1, No. 6, pp.1–16.

Nithya, B. and Ilango, V. (2020) 'Machine learning aided fused feature selection based classification framework for diagnosing cervical cancer', *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*, Erode, India, pp.61–66.

Park, H.W., Li, D., Piao, Y. and Ryu, K.H. (2017) 'A hybrid feature selection method to classification and its application in hypertension diagnosis', *International Conference on Information Technology in Bio-and Medical Informatics*, Springer, Cham, pp.11–19.

Pramanik, S., Galety, M.G., Samanta, D. and Joseph, N.P. (2023) 'Data mining approaches for healthcare decision support systems', *Emerging Technologies in Data Mining and Information Security*, Springer, Singapore, pp.721–733.

Priya, S. and Karthikeyan, N.K. (2021) 'Deep learning classification to improve diagnosis of cervical cancer through swarm intelligence-based feature selection approach', *Intelligent Systems, Technologies and Applications*, Springer, Singapore, pp.247–264.

Ramaraju, H.E., Nagaveni, Y.C. and Khazi, A.A. (2016) 'Use of schille*r's Test Vs. Pap Smear to Increase Detection Rate of Cervical Dysplasias. International Journal of Reproduction, Contraception, Obstetrics and Gynecology*, Vol. 5, No. 5, pp.1446–1451.

Rerucha, C.M., Caro, R. and Wheeler, V. (2018) 'Cervical cancer screening', *American Family Physician*, Vol. 97, No. 7, pp.441–448.

Seijo-Pardo, B., Porto-Díaz, I., Bolón-Canedo, V. and Alonso-Betanzos, A. (2017) 'Ensemble feature selection: homogeneous and heterogeneous approaches', *Knowledge-Based Systems*, Vol. 118, pp.124–139.

Seo, S.S., Oh, H.Y., Lee, J.K., Kong, J.S., Lee, D.O. and Kim, M.K. (2016) 'Combined effect of diet and cervical microbiome on the risk of cervical intraepithelial neoplasia', *Clinical Nutrition*, Vol. 35, No. 6, pp.1434–1441.

Shang, S., Shi, M., Shang, W. and Hong, Z. (2016) 'Improved feature weight algorithm and its application to text classification', *Mathematical Problems in Engineering*, Vol. 2016, No. 7819626, pp.1–12.

Suehiro, T.T., Malaguti, N., Damke, E., Uchimura, N.S., Gimenes, F., Souza, R.P., and Consolaro, M.E, L. (2019) 'Association of human papillomavirus and bacterial vaginosis with increased risk of high-grade squamous intraepithelial cervical lesions', *International Journal of Gynecologic Cancer*, Vol. 29, No. 2, pp.242–249.

Tarawneh, A.S., Hassanat, A.B., Almohammadi, K., Chetverikov, D. and Bellinger, C. (2020) 'Smotefuna: synthetic minority over-sampling technique based on furthest neighbour algorithm', *IEEE Access*, Vol. 8, pp.59069–59082.

Tubishat, M., Alswaitti, M., Mirjalili, S., Al-Garadi, M.A. and Rana, T.A. (2020) 'Dynamic butterfly optimization algorithm for feature selection', *IEEE Access*, Vol. 8, pp.194303–194314.

Wald, R., Khoshgoftaar, T.M. and Dittman, D. (2012) 'Mean aggregation vs. robust rank aggregation for ensemble gene selection', *2012 11th International Conference on Machine Learning and Applications*, USA, pp.63–69.

Wu, W. and Zhou, H. (2017) 'Data-driven diagnosis of cervical cancer with support vector machine-based approaches', *IEEE Access*, Vol. 5, pp.25189–25195.

Yang, X., Da, M., Zhang, W., Qi, Q., Zhang, C. and Han, S. (2018) 'Role of lactobacillus in cervical cancer', *Cancer Management and Research*, Vol. 10, pp.1219–1229.

Zhou, Q., Zhou, H. and Li, T. (2016) 'Cost-sensitive feature selection using random forest: selecting low-cost subsets of informative features', *Knowledge-Based Systems*, Vol. 95, pp.1–11.