Journal of Ambient Intelligence and Humanized Computing
https://doi.org/10.1007/s12652-020-02194-6

ORIGINAL RESEARCH

Integrating HSICBFO and FWSMOTE algorithm-prediction through risk factors in cervical cancer

S. Geeitha[1] · **M. Thangamani**[2]

Received: 5 February 2020 / Accepted: 6 June 2020

Abstract
The prominent objective of cervical carcinoma (CC) prediction lies in the optimal feature selection and balanced data. The problem of majority and minority class samples are solved in the proposed work. The objective of the work lies in solving imbalanced data distribution, and of risk factor validation in cervical cancer prediction. Feature Weighted Synthetic Minority Oversampling Technique (FWSMOTE) algorithm solves the minority class issues. The missing data imputation is performed by the Mode and Median Missing Data imputation. For optimal feature selection, Hilbert–Schmidt Independence Criterion with Bacteria Forage Optimization (HSICBFO) algorithm is implemented. Ensemble Support Vector Machine with Interpolation classifier is used for cancer prediction. Various measures are deployed to analyze the performance of the proposed classifier and produces 94.77%, 93.38%, 93.86%, 94.07%, 93.60% and 93.62% for precision, recall, specificity, F-Measure, accuracy and G-mean that helps in identifying the risk level of cervical carcinoma development and guidance for further diagnosis.

Keywords Cervical carcinoma (CC) · Minority class · Data imputation · Measures · Classification · Risk factor

1 Introduction

In worldwide, cervical carcinoma (CC) has huge impermanency among women. A leading mortality cancer among women is Cervical cancer (CC) with new cases of 500,000 and incidences is about 250,000 in 2012 (Wiiliam and Miller 2012). High risk factor of CC includes Human Papilloma-Virus (HPV) infection but not sufficient for malignancy initiation (Tjalma et al. 2005). HPV screening proves to be a greater protection against the malignant cervical carcinoma when compared to the pap smear testing (Geeitha and Thangamani 2020). Besides HPV infection, some factors increase the risk of carcinoma such as smoking, usage of birth control pills for contiguous five years, more sexual parity, etc., (Singh et.al 2018). Additionally, some information that include victim's past medical records reveals the fact that age group falling between 50 and 69 were found be highly infected with this carcinogenesis (Kour et al. 2010). Progression from precancerous disorder to invasive cancer can be avoided using genetic alterations (Martin et al. 2009). Cancer research uses the profile of gene expression. Numerous techniques are being implemented for identifying different factors of CC. Few polymorphisms have shown their consistency. Patterns of the gene expression are combined with statistical techniques to be used in various cancer types (Huang and Yu 2013; Zheng et.al 2011). Behavior of cancer cells can be understood by studying the individual genes in different carcinoma cells. Few studies used those methods in genomic scale (Rhodes et al. 2004; Segal et. al 2004).

Important information to the policy and decision makers are given by the quantification of CC rate. Population-based registries are used to get a CC's accurate measure. The estimation of the disease occurrence is defined in a clear-cut population. In cancer registries, the data collection quality and completeness, precise and consistent measures of inhabitants play a vital role (Sharma et. al 2013). The prediction models are affected by variations in the clinical outcome and it makes the process of identifying treatment protocols to CC as a complicated one (Hu et al. 2010). Target-gene

✉ S. Geeitha
geeitha.s@gmail.com

M. Thangamani
manithangamani2@gmail.com

[1] Department of Information Technology, Mahendra Engineering College for Women, Tiruchengode, India

[2] Department of Computer Science Engineering, Kongu Engineering College, Perundurai, India

Published online: 10 June 2020

treatments can be applied based on the developments in the analysis methods of gene expression to solve this issue and to make the survival of the people (Ithana et al. 2015). Several researchers have proved that metaheuristic algorithms contribute better to solve certain optimization paradigms (Khan et al. 2019). Especially bat optimization model is deployed for generating more appropriate features with objective function values (Geeitha and Thangamani 2018).

Normal cervical cancer data have some issues other than gene expression-associated analysis. Even distribution of normal data and gene expressions are assumed by most of the prediction techniques. In unbalanced prediction class, majority class corresponds to general data and minority class corresponds to training gene expression data. The prediction accuracy of unbalanced class is reduced by a large amount due to this minority class. Gene interactions are used in this proposed work instead of gene levels. Feature Weighted SMOTE (FWSMOTE) is used to solve the imbalanced issue. Gene co-expression network is constructed to find the changes induced by cancer and genes causing the disease. Feature selection contributes a major role in diagnosing the cancer diseases that helps in the insight of the gene expression. Selecting appropriate features relevant to carcinoma from gene dataset leads to proper disease analysis. Hilbert–Schmidt Independence Criterion (HSIC) method is used to perform feature selection. It also includes Bacterial Foraging Optimization (BFO) (HSICBFO). Ensemble Support Vector Machine with Interpolation (ESVMI) classifier is used to predict the cervical cancer disease.

2 Literature review

Tan et al. (2018) implemented an integrative analysis composing of analyzing cervical gene dataset, selecting gene features, analyzing using machine learning associated with meta-analysis of more than three datasets. An integrative survey using machine learning is used to identify 21 gene expressions. It includes one unsupervised and seven supervised methods. Among them, 21-gene expression set is selected, Gene Set Enrichment Analysis (GSEA) is performed and again nine probable gene expression markers show the noteworthy enhancement.

Kori and Arga (2018) performed meta-analysis on the cervical cancer-associated transcriptome data. Based on the analysis, biomolecules at protein, RNA (mRNA, miRNA) and metabolite levels are identified by integrating genome-scale biomolecular networks with gene expression profiles. The research explores a worthable dataset for future observation and clinical decision and proposed that the biomolecules possess a predominant capability when compared with existing biomarkers for testing or healing cervical cancer.

Deng et al. (2016) found the key genes of cervical cancer. The research used various bioinformatics tools to select differentially expressed genes (DEGs) that contains 143 genes. The cervical cancer development is based on these DEGs as shown by the results of bioinformatics analysis. A sub network is identified at two different conditions by comparing the two different Co-Expression Networks at two different states to find a common sub-network. It is also used to find hub genes in three sub-networks. Gene Ontology (GO) function enrichment, protein binding and pathway enrichment are used to analyze the gene hubs. The development and the guidance for treating cervical cancer can be identified by using this analysis.

Zhang and Zhao (2016) used DEGs between CC and normal controls. CC's Pathogenic network was constructed with known pathogenic genes. The sub-networks were identified for cluster analysis using Cluster ONE. Weight value is assigned to all the genes in the malignant network. Candidate genes are obtained according to the distribution of weights. The guidance for the future applications can be derived from theoretical results.

Chen et al. (2018) integrated Pearson's Correlation Coefficient (PCC) with a causal inference method ie., intervention effect during absence of DAG (IDA) with and Least Absolute Shrinkage and Selection Operator (LASSO) to form (PIL) methodby implementing Borda count election algorithm method and predicted microRNA (miRNA) targets and patients' pathway on miRNA based expression data.

Guyon et al. (2002) suggested a backward feature model where elimination of redundant feature is carried out. The model recursively eliminates the unwanted or redundant features by assigning weights for the gene features with estimator. At the initial stage, certain gene features are exploited for training the estimator and then the absolute weights are assigned for every gene feature. In the second stage, these estimated weights are arranged in an descending order and finally the last features are discarded.

Chandra and Gupta (2011) introduced Effective range based gene selection (ERGS) that is based on theory of statistical inference. The weight of each feature is calculated and among them the larger weight value are allotted to the significant or relevant gene feature in order to differentiate the distinct classes. Here the feature selection is based on the filter method.

Fatlawi (2007) implemented classifiers which is highly based on cost. It evaluates three phases. Preparation of the original data for classification is done at the first stage. Cost selective decision tree classifier is deployed for classifying. Various metrics are used to evaluate the performance including cross validation. Accurate results in multi class and binary class classification can be achieved by using this model. It has a True Positive (TP) rate of 0.429.

Purnami et al. (2016) presented a model to predict cervical cancer survival. It uses Smooth Support Vector Machine (SSVM), Classification and Regression Tree (CART) and Three Order Spline SSVM (TSSVM) classification methods. Synthetic Minority Oversampling Technique (SMOTE) is deployed to handle the minority class based cervical cancer dataset. Sensitivity, Specificity and accuracy are evaluated to measure the performance of the method. The proposed model provides better outcome than the previous SMOTE-CART, SMOTE-TSSVM methods.

Nandagopal et al. (2019) proposed a fuzzy based logistic regression for predicting the gene expression. The work is mainly focuses on feature selection and solves the inefficiency problem by implementing LASSO Logistic Regression (LLR). The author also introduced Maximum Likelihood Estimation (MLE) a computational tractable method to predict the breast cancer cells.

Wu and Zhou (2017) diagnosed the cervical cancer with the help of Support Vector Machine (SVM) approach. To diagnose samples of the malignant cancer, a derivative methods of SVM such as SVM associated with Principal Component Analysis (SVM-PCA) and SVM associated with recursive feature elimination methods are proposed. By experimental comparison, the model concluded that the derivative with PCA model provides better results.

Anagaw et al. (2019) introduced a double learning method to enhance the performance of the classifier. The model mainly maximizes the computational activity and discards the minority data samples that leads to incorrect classification from minority samples. The model is dependent on noised samples. The work exploits the noised data for constructing the model using complement naive Bayesian (CNB) in order to provide better accuracy. Classifier's performance is measured by implementing the model on Wilt and Tic-Tac-Toe endgame datasets.

3 Proposed methodology

Existing prediction techniques considers the even distribution of normal data and DEGs related gene expressions. In classification, majority class corresponds to general data and minority class corresponds to training gene expression data that leads to unbalanced prediction where accuracy is reduced by a large amount due to imbalance class. To overcome this minority issue, Synthetic Minority Oversampling Technique (SMOTE) is most commonly used. The key genes and data sample features of cervical cancer are identified by the proposed method for the treatment of cervical cancer. Dataset of the selected cervical cancer is represented using 32 risk factors and 4 target variables by means of various data mining methods. The variables are Hinselmann, Cytology, Biopsy and Schiller. Major stages of the proposed work includes, preprocessing, imbalanced dataset conversion, network construction, Feature selection, and classification shown in the Fig. 1. Mode and Median Missing Data imputation (MMM-DI) is used to perform the missing data imputation. To balance samples of the cervical cancer, Feature Weighted SMOTE (FWSMOTE) algorithm is used. Hilbert–Schmidt Independence Criterion (HSIC) with Bacterial Foraging Optimization (BFO) known as HSICBFO is used to select the important features. This feature selection and sampling data into subsets lies the state of art in this work. Ensemble Support Vector Machine with Interpolation (ESVMI) classifier is used to predict the cervical cancer disease. The development of cervical cancer mainly depends on the DEGs and samples as shown by the data mining analysis.

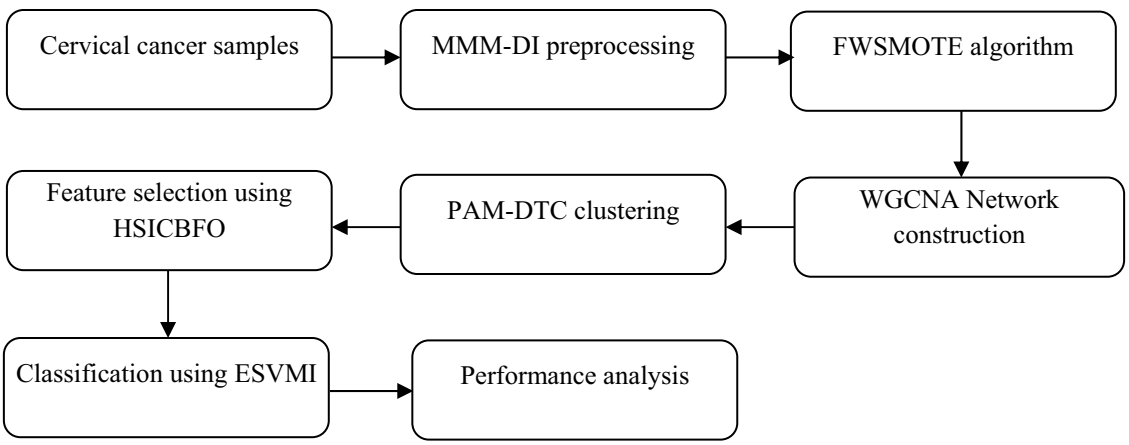

Fig. 1 Flow diagram of proposed work

3.1 Mode and Median Missing Data imputation (MMM-DI) algorithm

Missing values in the samples are computed by Mode and Median Missing Data imputation (MMM-DI) algorithm. The missing values of the residual genes are replaced by mode and median values of other samples under similar condition. The number with highest frequency represents mode of a feature set. The missing data of a variable is replaced by the mean of all known values of variable. But it does not depict the typical outcome.

The mean will be strongly affected, if there exists an outcome that is very far from the rest of the data and it is called as outlier. In this case, median is used as an alternative measure. In the proposed, attributes or features of cervical cancer dataset is considered as the mid value to find the median. The missed value in the samples is replaced by average value of mode and median.

$$MM_{DI} = \frac{mode + median}{2} \tag{1}$$

3.2 Feature Weighted SMOTE (FWSMOTE) algorithm

SMOTE is an oversampling technique (Han et al. 2005; Luengo et al. 2011). It synthesis minority gene samples in the class distribution. It is done by fabricating synthetic gene samples that operates in feature space. By taking all gene samples and by generating synthetic gene samples that falls on the line segments which join any or all minority class with k nearest neighbors, oversampling of minority class is done.

The proposed FWSMOTE algorithm gives importance to each feature to produce better results. The classification results are also increased. k nearest neighbors chosen initially in the algorithm. The difference that occurs between samples of feature weighted vectors is used to create synthetic samples. In this model, a random number that lies between 0 and 1 is multiplied and added with feature weighted vector for consideration. Thus, there occurs a random point between two specific features that falls on the line segment is selected. Finally, data region that belongs to minority class is broadened by SMOTE and the decision region is forced to be the common class (Maciejewski and Stefanawski 2011; Jeatrakul et al. 2010).

The distance between the feature vectors provide solution to imbalanced problem by SMOTE algorithm. Each gene feature and gene vectors are assigned with specific weight values further for considering the distance from one feature vector to another. The weight vectors and feature vectors are used to compute the distance. Every feature and gene vectors are multiplied with weight values. For all of the k nearest neighbors of minority class, new distances are computed using newly generated genes and features vectors.

3.3 Network construction

Weighted correlation network analysis can be done by R software package "Weighted Correlation Network Analysis (WGCNA)" (Langfelder and Horvath 2008; DiLeo et al. 2011). It has functions for network construction, computations of topological properties, module detection, simulating data and articulate with external software. The data mining algorithms are deployed for various environments to develop gene expression data and data samples. Genes are represented as nodes in biological network and edges are represented using features.

A biological entity place in network analysis is taken as a confined neighborhood among the network. The correlation measure defines the notion where the cervical cancer gene expression data is constructed as a network graph. It depends on value of the similarity genes. To produce biological intuitions and to create gene networks, bioinformatics expertise in order to handle efficiently, integrate and effectively make use of data is required.

3.4 Cluster identification using PAM-DTC

Differentially expressed gene pairs are obtained using Pearson correlations associated with Partition Around Medoids (PAM) algorithm (Van der Laan et al. 2003).The correlation cutoff is fixed as 0.8 for the gene pairs. The gene pairs with greater correlation value when compared to cutoff value are linked to create differential co-expression network. Two DCNs are created for two different conditions. The conditions are cancer and normal state. Dynamic Tree Cut (DTC) (Langfelder et al. 2007) used to identify the right clusters. This identification is done based on the shape.

Network nodes are formed by these filtered genes. Co-expression network is constructed. Gene pairs are linked if they have similarity greater than a specified cutoff similarity. The software cytoscape envisions the construction of DCN. Gene expression data level correlation is defined by this. The correlation value is decreased between the variables of gene pair provided partial correlation not decreased towards zero.

Scrutinizing the collection of independent variables, computable measures of gene expressions or genetical marker sof Single Nucleotide Polymorphisms (SNPs) that delineate not less than the part of the correlation between two gene expressions provides an insight about gene regulation. Correlated gene expression contains similar functions that hold gene network reconstruction with highlighted feature.

3.5 Feature selection with HSICBFO

Informative features and gene selection is a challenging task. Best features of the samples are searched by Bacterial Foraging Optimization (BFO) with HSIC to solve this issue. It is an activity shown by bacterial foraging behaviors which is named as "chemotaxis". The gene and feature selection are done based on this behavior. *E. coli* and salmonella are motile bacteria boost among them by rotating flagella.

The features and genes are selected by rotating the flagella in counter clockwise direction. Due to this the organism "swims" which makes the bacterium to find most relevant cervical cancer risk factors in a random manner "tumble" takes a new path and it swim once more. Best features in positive directions are found by the bacterium by building a substitution amidst "swim" and "tumble". To increase prediction rate of cervical cancer the bacterium swings more frequently. Bacterium moves from one feature to other features in search of more relevant genes when the direction changes. It is called Tumbling. Complex combination of swimming and tumbling makes the Bacterial chemotaxis. Prediction rate of the cervical cancer can be increased by keeping those bacteria in highly concentrated places. The classical BFO system has chemotaxis, elimination-dispersal and reproduction mechanisms (Chen et. al 2009).

4 Chemotaxis

Tumble is defined as a unit walk with non linear direction and run is defined as a unit walk with optimal selection of the features in classical BFO. Bacterium with jth chemotactic, kth reproductive, and lth elimination-dispersal phase is represented by $\theta^i(j,k,l)$. Unit run length parameter is given by $C(i)$ and it is a chemotactic phase size at every run or tumble. The motion of ithbacterium in each computational chemotactic step is given by,

$$\theta^i(j+1,k,l) = \theta^i(j,k,l) + C(i)\frac{\Delta(i)}{\sqrt{\Delta^T(i)\Delta(i)}} \tag{2}$$

where jth chemotactic step's direction vector is given by $\Delta(i)$. Last chemotactic step equals the $\Delta(i)$ during the run of bacterial movement; otherwise, $\Delta(i)$, a random vector and its components ranges between [− 1, 1]. In the chemotaxis process, run activity for every step with its fitness is obtained by $cov(i,j,k,l)$ that acts as an evaluated function.

5 Reproduction

Step fitnesssum corresponds to the value of feature of a sample during its life,

$$\sum_{j=1}^{N_c} cov(i,j,k,l) \tag{3}$$

Here, upper limit step in chemotaxis procedure is given by N_c. Based on the fitness value features are sorted in reverse order. First half of the population are got survived in the reproduction process and they are divided into two identical parts. They are locatedat similar position. The specific process makes the standard feature population.

6 Elimination and dispersal

Local feature selection and convergence of reproduction process are based on the chemotaxis. For global optimum feature selection, reproduction and chemotaxis cannot produce optimum results. Bacteria are stuck around the positions of features. Due to this diversity of BFO may be changed and eliminate the accidents that is being trapped under the local optima. The reproduction process takes place for current feature on certain time and dispersion occurs in BFO. Based preset probability pr_{ed} few features are selected and killed. The algorithm moves to another feature position in the environment.

1. n, S, N_{ch}, N_{sw},, N_{re},N_{ed},Pr_{ed}, C(i) where (i = 1,2,….,S),θ^i are parameters that are initialized. n represents the of search space dimension, the number of bacterium is represented by S, chemotactic steps is given by N_{ch}, swim steps is given by N_{sw}, reproductive steps is given by N_{re}, elimination and difuse steps is given by N_{ed}, elimination probability is given by Pr_{ed}, the unit length of run is given by C(i)- (i.e., step of chemotactic size at every run or tumble).
2. Looping for elimination-difuseis given by l = l + 1.
3. Looping for Reproduction is given by k = k + 1.
4. Looping for Chemotaxis is given by loop j = j + 1.
   - 4.1. For i = 1 = 1, 2…,S, take a chemotactic step for bacteria as follows.
   - 4.2 Compute fitness function, cov(i, j, k, l).
   - 4.3 Let $cov_{last} = cov(i,j,k,l)$ to save this value since we may find better value throughrun.
   - 4.4 Next Tumble: a random vector is generated $\Delta(i) \epsilon \mathbb{R}^n$ for everycomponent,$\Delta_{m(i)}$, $m = 1,2,3 \ldots S$, that holds a random valuebetween [− 1, 1].

4.5 Proceed: Assume (2.1). This ends within a step of size C(i) that takes the tumble direction of bacteria i.
4.6 Calculate cov(i,j + 1,k,l,) along with θ^i(j + 1,k,l).
4.7 Swim:

(i) Considering m as 0 (taking as counter for length of swim).
(ii) While $m < N_{sw}$ (provided not to be climbed-forso long).

a. Considering m = m + 1.
b. If $cov(i, j+1, k, l) < cov_{last}$, assume $cov_{last} = cov(i, j+1, k, l)$, Further, next step of size $C(i)$ in taking similar direction can be considered as (2.1) and utilize the newly generated $\theta^i(j+1, k, l)$ to calculate new value for $cov(i, j+1, k, l)$.
c. Otherwiseconsider $m = N_s$.

4.8 Move to other bacterium (i + 1): move to (b),if $i \neq S$ in order to proceed with next bacteria.

5. If $j < N_{ch}$, select Step 3. Here,there occurs continuou-schemotaxis as the bacteria's life is not completed.
6. Reproduction,

6.1 Provided k and l, and for every$i = 1,2,\ldots,S$,consider

$$Cov^i_{health} = \sum_{j=1}^{N_{ch}+1} Cov(i, j, k, l) \tag{4}$$

Bacteria health. Bacterium is being sorted in ascending orders.
6.2 S, be the highest cov_{health} value of bacteria that die, best values of S features may split, and the replicas generated may be positioned at the similar location same as their parent.

7. For $k < N_{re}$ go to Step 2. In this criterion, particular reproduction steps are not yetarrived; in the chemotactic looping let the sequentially next generation begins.
8. Elimination-difusal: for i = 1,…S, possessing probability pr_{ed}, neglect and difuseevery features, that may result by maintaining the certain features in the population constant. To do this, if a featureis reduced, in the optimization area just difuse one to a random location. If $l < N_{ed}$, then take Step 2, else, end.

Let x and y be two random variables. They are said to be independent provided that bounded continuous function of those variable may be uncorrelated. This is the fact used by HSIC. Random variables x and y are considered as a multivariate random variables and their joint probability distribution function is given p_{xy}. The supports of the random variables x and y are given as X and Y. Here with separable RKHS, two variables F and G are taken for real valued functions that flows from X to R and Y to R along with the universal kernel respectively. The cross-covariance between F and G is given by,

$$cov(f(x), g(y)) = E_{x,y}\left[f(x)g(y)\right] - E_x\left[f(x)\right]E_y\left[g(y)\right] \tag{5}$$

whereas $f \in F$, $g \in G$, and E is considered as expectation function.

7 Topological properties

DEGs are screened to estimate and create linear model. For each gene, a linear model is regarded to for microarray experimentation. Variance between the samples and dissimilarities between the partial genes are grouped. Variance between the samples of genes is used for variance of gene approximation. The combination variance of neighborhood genes is also used for the same purpose. For a node *n* which has kn neighbors, topological coefficient T_n is computed as,

$$T_n = \text{avg}(J(n, m))/k_n \tag{6}$$

Here, the nodes sharing with minimum one neighbor along n is given by v J(n,m) and its value defines the number of neighbors sharing between node n and m, in addition to one when there occurs a direct communication between n and m. Topological coefficient measures the degree of a node sharing neighbors with other nodes. The topological coefficients distribution is more disseminated in cancer DCN and numbers of neighbors is more than 10.

Various modifications and strong linking of various modules are done in modularity of two DCNs. The genes with high eigenvectors and eigen values are contained by this. From these values ideal solution can be obtained in an easy way. Optimal selection genes and reduction in the iterations can be done by using the eigen genes.

7.1 Classification using ESVMI

The classification is done using supervised learning technique namely ESVMI. It is used in more than two classifications especially for multi class purpose. It uses hyperplane to divide the known pair that comprises {+ 1, − 1} cervical cancer training samples that are labeled. Maximum distance between positive and negative data samples of cervical cancer defines the hyperplane. A decision boundary which is non-linear is found in input data of the cervical cancer by separating the hyperplane

of the cervical cancer feature space (Claesen et al. 2014). Detailed explanation about ensemble SVMI can be taken from (Lo et al. 2015; Sorensen et al. 2018). By considering a set of N distinct cervical cancer instances (fe_i, y_i) with $fe_i \in \mathbb{R}^D$ and $y_i \in \mathbb{R}^d$. SVM can be designed as (Melgani and Bruzzone 2004)

$$\sum_i a_i K\left(fe, fe_i\right) we_i + b, i \in [1, N]. \quad (7)$$

Where kernel function is given by K(fe, fe_i), α is the parameter of SVMI and b is the threshold of ESVMI. Commonly used interpolation method includes distance based weighting method (Ly et al. 2013). Point similarity is used as a base in distance based weighting method. Adjacent points are weighted to predict the precipitation of unknown cervical cancer points. The distance between measured site and unknown cervical cancer point are used to compute the weight in this method. Semi-variance values are used to solve these weights that are measured in each station.

The classifiers' diversity brings the success for construction of ensemble system. Majority vote ensemble learning algorithm is used in this work. Figure 2 shows the architecture of ensembles using multiple SVMI models. Ensemble algorithms have different aggregation methods. The majority vote of the decisions taken by the classifiers are aggregated or combined.

Minority class is formed by dividing the majority samples of cervical cancer into subsets of similar sizes. The subsets will be made equal in size by randomly selecting and placing the samples of cervical cancer in the majority class until required CC sample size is met. The random selection process is done continuously until the creation of all subsets. The records in the subset are unique and across various subsets duplicated can be found.

Majority vote: The cervical cancer training sample construction made the difference between proposed method and random sampling with majority vote (Lo et al. 2015). The multiple training subsets are formed by combining the cervical cancer samples with dataset from all owners of the account. It is shown in Fig. 3. SVMI is trained individually with the same size cervical cancer training samples.

Majority votes are used for aggregation. This voting is based on classes or labels. It is similar to random sampling. The class that has highest vote is chosen as an ensemble (Hoi and Lyu 2004) and it is termed as predicted class. The 32 risk factors of 668 patients are considered along with the target variables for testing and the risk score is evaluated using the gene dataset GSM99077 and GSM99078 with the relevant risk factors.

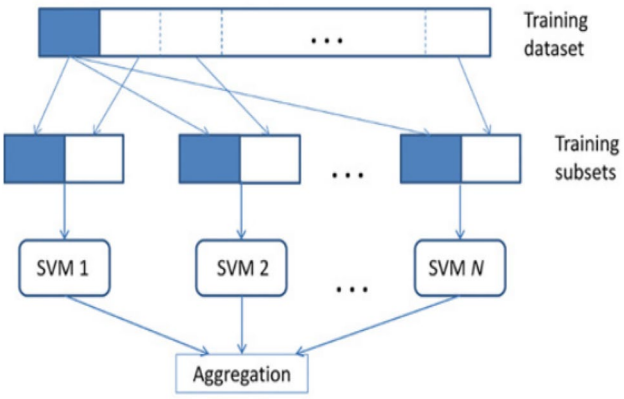

Fig. 2 Model of ensemble SVMI system with multiple SVMs

Fig. 3 The majority vote algorithm

Input:

Ds: Training samples of cervical cancer **with C classes of labels**

CLA: Classifier

La: cervical cancer sample labels

N: Number of La used

Do n= 1 to N

1. Call CLA with Ds_n and receive the classifier La_n
2. Compare W_n with C_n created from La_n update vote
3. Aggregate vote to the ensemble

End

8 Results and discussion

The performance of the proposed method is evaluated using synthetic data. With the intention of illustrating the performance of the feature selection algorithm, it provides the experimentation outcomes attained by means of several real-world datasets.The GEO data set for microarray data of cervical cancer (GEO accession: GSE4482 with GSM99077 and GSM99078) are takenfrom GEO website (https://www.ncbi.nlm.nih. gov/GEO/). The experimentation is done in MATLAB environment and results of observation are discussed in this phase. Hospitals in 'Universitario de Caracas' at Caracas are used to collect the dataset of cervical cancer (Fernandes et al. 2017) 32 risk factors are used to represent the dataset. The factors include patient's habits, historic medical records and demographic information and they are shown in Table 1. Hinselmann, Biopsy, Schiller and Cytology are target variables. Colposcopy using acetic acid is referred to as Hinselmanns test and using iodine are referred as Schillers test, cytology and biopsy. Few patients may not answer some questions due to their privacy. So dataset must be pre-handled in order to fill with values that are missing. Such a data is called imbalanced data. In the pre-treatment process, oversampling is applied. Due to the lack in values available, 27th and 28th risk factors are eliminated. So 668 patients' cervical cancer data are analyzed with 30 features.

In biomedical data, accuracy as well as the correct diagnosis takes a major contribution in the evaluation of algorithms. Algorithms predicting all the cancer samples in small dataset will have high accuracy when compared to those predicting properly in large dataset. The parameters like Positive Predictive Accuracy (PPA), sensitivity, accuracy and specificity, p-value, G-mean are used as an evaluation parameters of the algorithm.

$$Total\,Accuracy = (Tp + Tn)/(Tp + Tn + Fp + Fn) \tag{8}$$

$$Sensitivity\,or\,recall = Tp/(Tp + Fn) \tag{9}$$

$$\Pr ecision\,or\,Positive\ \Pr edictive\,Accuracy = TP/(TP + FP) \tag{10}$$

Rate of True negative can be called as Specificity. It measures correct proportion of actual negatives that is accurately identified with negative values.

$$Specificity = TN/TN + FP \tag{11}$$

The weighted harmonic mean of PPA and *Sensitivity* is called F-measure

$$F = 2\frac{PPA.SEN}{PPA + SEN} \tag{12}$$

TP represents true positive. It refers to malignant cancer that is diagnosed correctly. TN represents true negative. It refers to negative predictions that predicts uninfected people equally those that are classified as malignant. FP represents false positive. It corresponds to the number of samples not affected with cervical cancer but recognized as CC group. FN can be grouped as the number of undetected malignant cancer instances. Thus the whole accuracy corresponds to precision parameter of algorithm. The proportion of accurately detected cancer samples among all samples are shown in the Table 2.

The ratio of correctly diagnosed malignant cancer samples to overall malignant samples is named as Sensitivity. It is also termed as recall. The precision of the model is reflected by this positive predictive accuracy. Positive accuracy is a ratio of samples which are malignant classified as malignant group. Negative prediction accuracy is computed as ratio of benign samples which are precisely classified to the provided benign samples.

P value is defined as the probability of without effect or without difference (null hypothesis), by obtaining a result which is equal to or maybe more extreme than

Table 1 Cervical cancer with risk factor attributes

Attributes name	Category	Attributes name	Category	Attributes name	Category
Age	Numerical	STDs	Numerical	STDs: Number of diagnosis	Numerical
Number of sexual partners	Numerical	STDs: Condylomatics	Numerical	STDs: Time since first diagnosis	Numerical
First sexual intercourse(age)	Numerical	STDs: Cervical Condylomatics	Numerical	STDs: Time since last diagnosis	Numerical
Number of pregnancies	Numerical	STDs: vaginal Condylomatics	Numerical	STDs(number)	Numerical
Smokes (years)	Numerical	STDs: vulvo-perinealCondylomatics	Numerical	STDs: molluscumcatagiosum	Numerical
Smokes (packs/years)	Numerical	STDs:Syphilis	Numerical	Dx:Cancer	Numerical
Hormonal contraceptives	Numerical	STDs:Pelvic inflammatory disease	Numerical	Dx:CIN	Numerical
Hormonal contraceptives(years)	Numerical	STDs: Gential herpes	Numerical	Dx:HPV	Numerical
IUD	Numerical	STDs:AIDS	Numerical	Dx	Numerical
IUD (years)	Numerical	STDs:HIV	Numerical		

Table 2 Performance comparison analysis

Methods	Samples	Precision (%)	Recall (%)/ sensitivity(%)	Specificity (%)	F-measure (%)	Accuracy (%)	*P* value	G-mean
DCA	50	80.95	77.27	85.71	79.07	82.00	84.81	81.38
	100	76.08	83.33	81.03	79.54	82.00	54.78	82.17
	150	81.42	81.42	83.75	81.42	82.66	100	82.58
	200	81.63	83.33	82.69	82.47	83.00	84.16	83.01
	250	83.20	83.84	81.66	83.52	82.80	93.14	82.74
DCN	50	77.77	91.30	77.77	84.00	84.00	35.03	84.27
	100	87.50	85.96	83.72	86.72	85.00	88.18	84.83
	150	90.12	83.90	87.30	86.90	85.33	44.72	85.58
	200	86.00	84.31	85.71	85.14	85.00	83.84	85.01
	250	84.39	88.80	81.03	86.54	85.20	55.32	84.83
EBO-SVM	50	95.45	77.77	95.65	85.71	86.00	30.21	86.25
	100	86.00	89.58	86.53	87.75	88.00	79.67	88.04
	150	88.09	90.24	85.29	89.15	88.00	81.88	87.73
	200	91.83	85.71	91.57	88.67	88.50	50.33	88.59
	250	88.13	87.39	89.31	87.76	88.40	92.44	88.34
ESVMI	50	96.42	90.00	95.00	93.10	92.00	65.66	92.46
	100	92.45	94.23	91.66	93.33	93.00	88.66	92.94
	150	94.80	92.40	94.36	93.59	93.33	82.55	93.38
	200	93.10	95.57	90.80	94.32	93.50	77.61	93.15
	250	94.77	93.38	93.86	94.07	93.60	86.59	93.62

Table 3 Risk value for cervical cancer and gene samples

Gene sample	GSM99077		GSM99078	
Target test(present/ absent)	0	1	0	1
Hinselmann	336.094	875.3217	133.3608	347.3244
Schiller	248.9974	1587.7929	98.8012	630.0302
Cytology	286.8702	1896.5304	113.829	752.5361
Biopsy	242.6265	2202.4224	96.2733	873.9129

wasreallyexamined. The probability is denoted by *P*. The following expressions are applied:

$$sv_j^2 = \sum_{i=1}^{N_j} (x_{ij} - s\mu_j)^2/(N_j - 1) \tag{13}$$

The principal mean of X predictor:

$$g\mu = \sum_{j=1}^{J} N_j s\mu_j / N \tag{14}$$

N_j, represents number of cases that equals Y and j.Mean of X (Predictor) for target class Y that equals j is given by $s\mu_j$, Variance ofX (Predictor) for target class Y that equals j is given by sv_j^2. All of this representations are grounded on non-missing pairs of (X, Y). Then, statistical F with p value is computed as,

$$p \text{ value } = \text{ Prob } \{F\,(J-1,\ N-J) > F\} \tag{15}$$

G-mean is defined as the Geometric mean of recall and precision. It is described as follows: And Table 3 depicts the performance of all the methods with respect to metrics.

$$G - \text{Mean} = \sqrt{\text{Recall*Precision}} \tag{16}$$

In the proposed work among 30 risk factor variables, feature selection is applied with four target labels: Hinselmann, Schiller, Cytology and Biopsy. Two test results (yes and no) are found to gene samples with four target labels. For those gene samples are integrated to four target labels (yes and no) with their risk values are discussed in the Table 3.

The risk score of the gene samples are computed via the use of the mean value of the selected features from gene set is multiplied by individual selected risk features from 32 risk factors. The risk value is computed as follows:

$$RiskScore = \sum_{i=1}^{n} \sum_{j=1}^{m} mean(x_i) * riskfactor_j \tag{17}$$

Figure 4 depicts the precision parameter comparison results with cc samples of four different methods such as Differential Co-expression Networks (DCNs), Differential

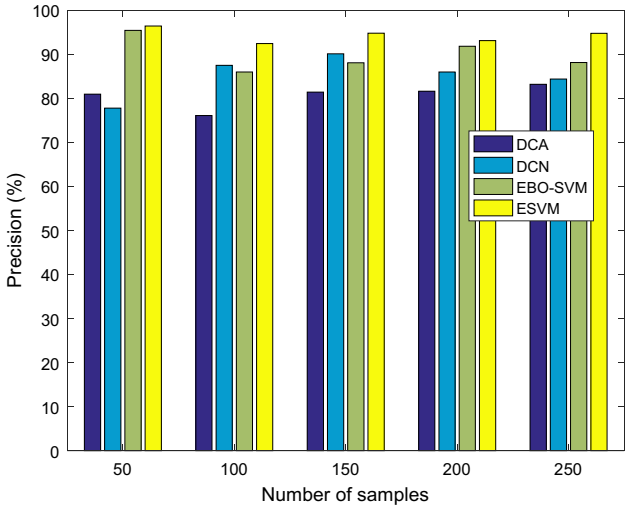

Fig. 4 Precision comparsion with samples vs. classifiers

Co-expression Analysis (DCA), Enhanced Bat Optimization Algorithm with Support Vector Machine (EBO-SVM) and proposed ESVMI classifier.The precision results are measured by changing the no. of samples from 50 to 250 with the interval of 50 no. of the samples. The proposed ESVMI technique provides higher precision results of 94.308% whereas existing DCA, DCN and EBO- SVM algorithms provides only 80.656%, 85.156% and 89.9% respectively for complete samples. The proposed ESVMI classifier has higher precision results since the proposed work introduces new iterpolation function for weight value computation in the SVM classifier. It increase the positive results for prediction.

Recall results of four different methods such as DCNs, DCA, EBO-SVM and proposed ESVMI classifier. Results are measured by changing the samples from 50 to 250 with 50 no. of the samples interval. Figure 5a shows the higher recall results of 93.116% is achieved by the proposed classifier, whereas other classifiers gives only 81.602%, 86.85% and 86.138% for existing DCA, DCN and EBO- SVM algorithms respectively.

Figure 5b presents the specificity results of four different methods that shows higher specificity results of 93.136% by the proposed classifier, whereas other classifiers gives only 82.968%, 85.86% and 86.76% of specificity values for existing DCA, DCN and EBO- SVM algorithms respectively.

Figure 6 shows F-measure results of four different methods such as DCNs, DCA, EBO-SVM and proposed ESVMI classifier.These results are measured by changing the samples from 50 to 250 with 50 no. of the samples interval and finally shows the higher F-measure results of 93.682% that is achieved by the proposed classifier, whereas other classifiers gives only 81.204%, 85.86% and 87.80% for existing DCA, DCN and EBO- SVM algorithms respectively.

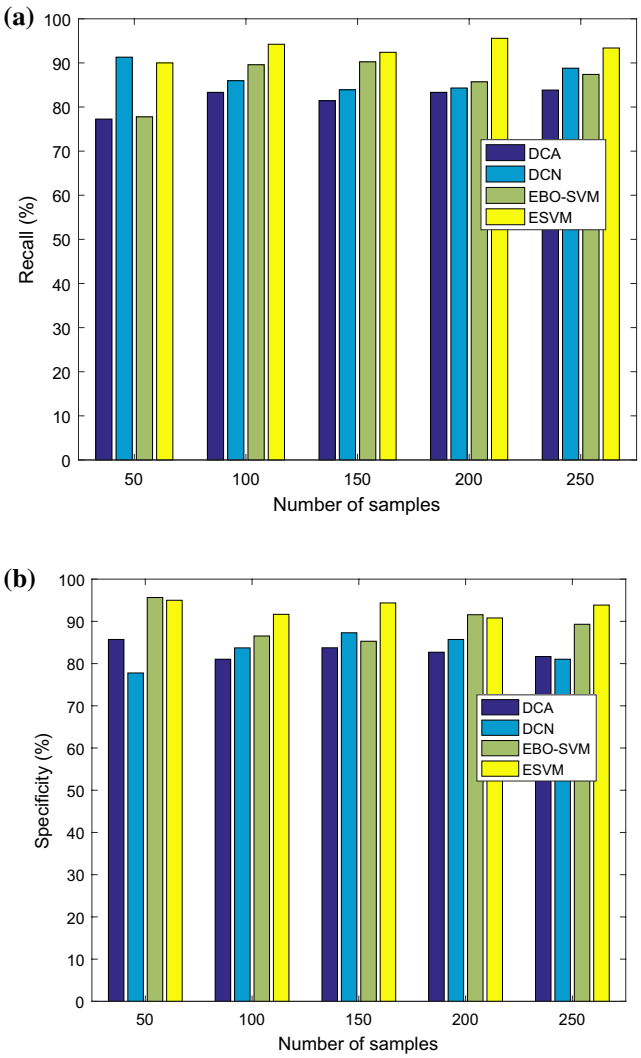

Fig. 5 **a** Recall (or) sensitivity results comparsion vs. classifiers. **b** Specificity results comparsion vs. classifiers

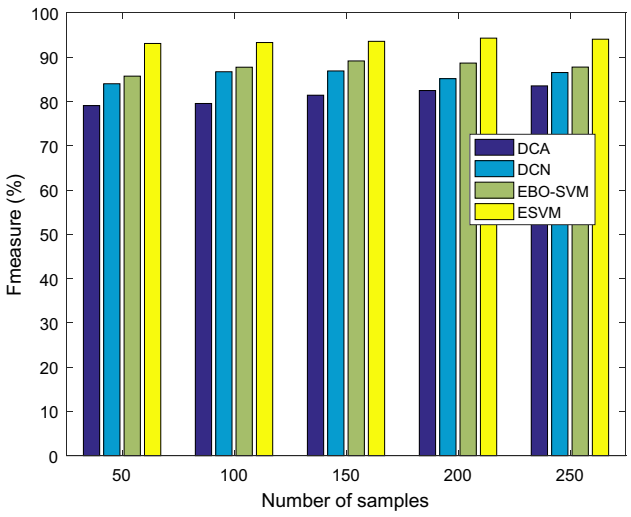

Fig. 6 F-measure results comparsion vs. classifiers

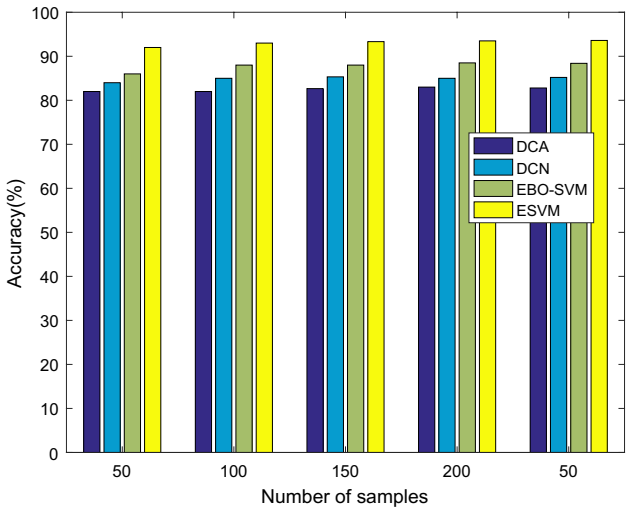

Fig. 7 Accuracy results comparsion vs. classifiers

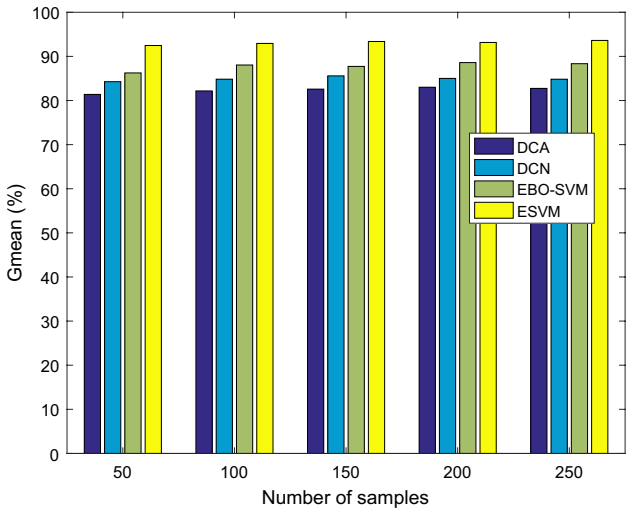

Fig. 9 G-Mean comparison vs. classifiers

The overall accuracy results of the four classifiers are shown in the Fig. 7. The proposed ESVMI gives higher accuracy results of 93.08%, the other methods such as DCA, DCN and EBO-SVM gives of 82.49%, 84.90% and 87.78% for overall samples. With the above results it concludes the proposed work possess higher rate of accuracy since the proposed work ensemble classification is performed by detection stage. So it improves the prediction results than the other methods.

The P-value statistical measure with respect to four classifiers is shown in the Fig. 8. From the result, it concludes that the proposed ESVMI classifier gives higher p-value results of 80.21, other classifiers are DCA, DCN and EBO-SVM gives only lesser value of 83.78%, 61.41% and 66.906% for overall samples.

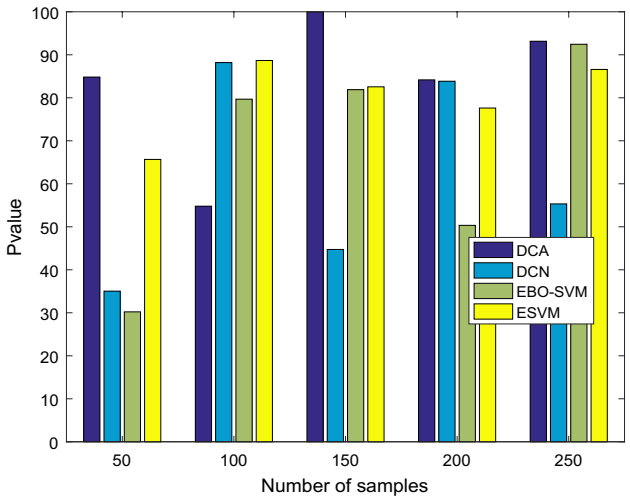

Fig. 8 P-value comparison vs. classifiers

Four different classifiers results measured with respect to G-mean are shown in the Fig. 9. It concludes that the proposed ESVMI classifier provides higher average G-mean value of 93.11%, other classifiers are DCA, DCN and EBO-SVM gives only lesser value of 82.376%, 84.904% and 87.79% for overall samples.

9 Conclusion and future work

Among women, cervical cancer causes death and it cannot be identified in early stages. At present various methods are used to diagnose the cervical cancer. Various methods are developed to detect it using data-driven methods are implemented in recent years. These techniques are to make diagnosis at proper time. Selection of features and sample balancing are more important in cervical cancer diagnosis. Preprocessing, imbalanced dataset conversion, network construction, Feature selection, and classification are the important four stages of the proposed work. The performance measures are computed and compared with the other classifiers. Mode and Median Missing Data imputation (MMM-DI) is used to perform missing data imputation in preprocessing stage. The samples of cervical cancer are balanced by Weighted SMOTE (FWSMOTE) algorithm. Hilbert–Schmidt Independence Criterion (HSIC) is used to select the important features from the samples. This process includes the Bacterial Foraging Optimization (BFO). The entire technique is known as HSICBFO. Diagnosis of cervical cancer is done by the proposed Ensemble Support Vector Machine with Interpolation (ESVMI) classifier. From University of California at Irvine (UCI) repository, dataset of cervical cancer is gathered. The risk factors measures and computation overhead are reduced by using combination of

ESVMI algorithm and HSICBFO. In this work, Risk factors are evaluated using some test values. But still it remains a challenge in handling all features associated with risk factors In future, specific risk issue relevant to gene expression associated with patient's habitat need to be considered for evaluating the epidemic target factors. The insight of the risk factors should be taken in to consideration for further diagnosis of the cervical cancer. Deep learning algorithms can be utilized in future to enhance the results.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflicts of interest.

References

Anagaw A, Chang Y (2019) A new complement naïve Bayesian approach for biomedical data classification. J Ambient Intell Hum Comput 10:3889–3897. https://doi.org/10.1007/s12652-018-1160-1

Chandra B, Gupta M (2011) An efficient statistical feature selection approach for classification of gene expression data. J Biomed Inform 44(4):529–535

Chen H, Zhu Y, Hu K (2009) Cooperative bacterial foraging optimization. Discret Dyn Nat Soc 815247:1–17

Chen R, Shi YH, Zhang H, Hu JY, Luo Y (2018) Systematic prediction of target genes and pathways in cervical cancer from microRNA expression data. Oncol Lett 15(6):9994–10000

Claesen M, Smet FD, Suykens JA, Moor BD (2014) Ensemble SVM: a library for ensemble learning using support vector machines. J Mach Learn Res 15:141–145

Deng SP, Zhu L, Huang DS (2016) Predicting hub genes associated with cervical cancer through gene co-expression networks. IEEE/ACM Trans Comput Biol Bioinf 13(1):27–35

DiLeo MV, Strahan GD, Den Bakker M, Hoekenga OA (2011) Weighted correlation network analysis (WGCNA) applied to the tomato fruit metabolome. PLoS ONE 6(10):e26683

Fatlawi HK (2007) Enhanced classification model for cervical cancer dataset based on cost sensitive classifier. Int J Comput Tech 4(4):115–120

Fernandes K, Cardoso JS, Fernandes J (2017) Transfer learning with partial observability applied to cervical cancer screening. Proc Iberian Conf Pattern Recognit Image Anal 10255:243–250 **(Springer International Publishing AG LNCS)**

Geeitha S, Thangamani M (2018) Incorporating EBO-HSIC with SVM for gene selection associated with cervical cancer classification. J Med Syst Springer 42(11):225

Geeitha S, Thangamani M (2020) A cognizant study of machine learning in predicting cervical cancer at various levels—a data mining concept. Int J Emerg Technol 11(1):23–28

Guyon I, Weston J, Barnhill S, Vapnik V (2002) Gene selection for cancer classification using support vector machines. Mach Learn 46(1–3):389–422

Han H, Wang WY, Mao BH (2005) Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In: International Conference on Intelligent Computing (ICIC) Springer Berlin Heidelberg Part I, LNCS, Vol. 3644, pp. 878–887

Hoi CH, Lyu MR (2004) Group-based relevance feedback with support vector machine ensembles. Proceedings of the 17th International Conference on Pattern Recognition Cambridge UK. Vol. 3, pp. 874–877

Hu X, Schwarz JK, Lewis JS Jr, Huettner PC, Rader JS, Deasy JO, Grigsby PW, Wang X (2010) A microRNA expression signature for cervical cancer prognosis. Cancer Res 70(4):1441–1448

Huang DS, Yu HJ (2013) Normalized feature vectors: a novel alignment-free sequence comparison method based on the numbers of adjacent amino acids IEEE/ACM Trans. Comput Biol Bioinformat 10(2):457–467

Itahana Y, Han R, Barbier S, Lei Z, Rozen S, Itahana K (2015) The uric acid transporter SLC2A9 is a direct target gene of the tumor suppressor p53 contributing to antioxidant defense. Oncogene 34(14):1799–1810

Jeatrakul P, Wong KW, Fung CC (2010) Classification of imbalanced data by combining the complementary neural network and SMOTE algorithm. In International Conference on Neural Information Processing (ICONIP), Springer, Berlin, Heidelberg, part II, LNCS Vol. 6444, pp. 152–159

Khan A, Shah R, Imran M et al (2019) An alternative approach to neural network training based on hybrid bio meta-heuristic algorithm. J Ambient Intell Human Comput 10:3821–3830

Kori Arga M (2018) Potential biomarkers and therapeutic targets in cervical cancer: Insights from the meta-analysis of transcriptomics data within network biomedicine perspective. PLoS ONE 13(7):e0200717

Kour P, Lal M, Panjaliya R, Dogra V, Gupta S (2010) Study of the risk factors associated with cervical caner. Biomed Pharmacol J 3(1):179–182

Langfelder P, Horvath S (2008) WGCNA: an R package for weighted correlation network analysis. BMC Bioinf 9(1):1–13

Langfelder P, Zhang B, Horvath S (2007) Defining clusters from a hierarchical cluster tree: the dynamic tree cut package for R. Bioinformatics 24(5):719–720

Lo SL, Chiong R, Cornforth D (2015) Using support vector machine ensembles for target audience classification on Twitter. PLoS ONE 10(4):e0122855

Luengo J, Fernández A, Garcia S, Herrera F (2011) Addressing data complexity for imbalanced data sets: analysis of SMOTE-based oversampling and evolutionary undersampling. Soft Comput 15(10):1909–1936

Ly S, Charles C, Degre A (2013) Different methods for spatial interpolation of rainfall data for operational hydrology and hydrological modeling at watershed scale: a review. Biotechnol Agron Soc Environ 17(2):392–406

Maciejewski T, Stefanowski J (2011) Local neighbourhood extension of SMOTE for mining imbalanced data. IEEE Symposium on Computational Intelligence and Data Mining (CIDM), pp. 104–111

Martin CM, Astbury K, McEvoy L, O'Toole S, Sheils O, O'Leary JJ (2009) Gene expression profiling in cervical cancer: Identification of novel markers for disease diagnosis and therapy. Methods Mol Biol 511:333–359

Melgani F, Bruzzone L (2004) Classification of hyper spectral remote sensing images with support vector machines. IEEE Trans Geo Sci Remote Sens 42(8):1778–1790

Nandagopal V, Geeitha S, Vinoth Kumar K, Anbarasi J (2019) Feasible analysis of gene expression—a computational based classification for breast cancer. Measurement (Elsevier) 140:120–125

Purnami SW, Khasanah PM, Sumartini SH, Chosuvivatwong V, Sriplung H (2016) Cervical cancer survival prediction using hybrid of SMOTE, CART and smooth support vector machine. In: AIP conference proceedings, AIP Publishing 1723(1)

Rhodes DR, Yu J, Shanker K, Deshpande N, Varambally R, Ghosh D, Barrette T, Pandey A, Chinnaiyan AM (2004) Large-scale meta-analysis of cancer microarray data identifies common

transcriptional profiles of neoplastic transformation and progression. Proc Natl Acad Sci 101(25):9309–9314

Segal E, Friedman N, Koller D, Regev A (2004) A module map showing conditional activity of expression modules in cancer. Nat Genet 36(10):1090–1098

Sharma M, Bruni L, Diaz M, Castellsague X, de Sanjose S, Bosch FX, Kim JJ (2013) Using HPV prevalence to predict cervical cancer incidence. Int J Cancer 132(8):1895–1900

Singh S, Narayan N, Sinha R, Sinha P, Sinha VP, Upadhye JJ (2018) Awareness about cervical cancer risk factos and symptoms. Int J Reprod Contracept Obstet Gynecol 7(12):4987–4991

Sorensen L, Nielsen M, Alzheimer's Disease Neuro imaging Initiative (2018) Ensemble support vector machine classification of dementia using structural MRI and mini-mental state examination. J Neurosci Methods 302:66–74

Tan MS, Chang SW, Cheah PL, Yap HJ (2018) Integrative machine learning analysis of multiple gene expression profiles in cervical cancer. PeerJ 6:e5285

Tjalma WA, Van Waes TR, Van den Eeden LE, Bogers JJ (2005) Role of human papillomavirus in the carcinogenesis of squamous cell carcinoma and adenocarcinoma of the cervix. Best Pract Res Clin Obstetr Gynaecol 19(4):469–483

Van der Laan M, Pollard K, Bryan J (2003) A new partitioning around medoids algorithm. J Stat Comput Simul 73(8):575–584

William TC, DS Miller (2012) Adenocarcinoma of the uterine corpus. Clin Gynecol Oncol, Eight Edition, Elsevier, Philadelphia, PA, ISBN No. 9780323074193, pp. 141–174

Wu W, Zhou H (2017) Data-driven diagnosis of cervical cancer with support vector machine-based approaches. IEEE Access. ISSN:2169–3536. Vol. 5, pp.25189–25195

Zhang YX, Zhao YL (2016) Pathogenic network analysis predicts candidate genes for cervical cancer. Comput Math Methods Med 3186051:1–8

Zheng CH, Zhang L, Ng VTY, Shiu SCK, Huang DS (2011) Molecular pattern discovery based on penalized matrix decomposition. IEEE/ACM Trans Comput Biol Bioinformat 8(6):1592–1603