Genes & Genomics
https://doi.org/10.1007/s13258-020-00916-w

Online ISSN 2092-9293
Print ISSN 1976-9571

RESEARCH ARTICLE

Detecting biomarkers from microarray data using distributed correlation based gene selection

Alok Kumar Shukla[1] · Diwakar Tripathi[2]

Received: 10 February 2019 / Accepted: 23 January 2020

Abstract

Background Over the past few decades, DNA microarray technology has emerged as a prevailing process for early identification of cancer subtypes. Several feature selection (FS) techniques have been widely applied for identifying cancer from microarray gene data but only very few studies have been conducted on distributing the feature selection process for detecting cancer subtypes.

Objective Not all the gene expressions are needed in prediction, this research article objective is to select discriminative biomarkers by using distributed FS method which helps in accurately diagnosis of cancer subtype. Traditional feature selection techniques have several drawbacks like unrelated features that could perform well in terms of classification accuracy with a suitable subset of genes will be left out of the selection.

Method To overcome the issue, in this paper a new filter-based method for gene selection is introduced which can select the highly relevant genes for distinguishing tissues from the gene expression dataset. In addition, it is used to compute the relation between gene–gene and gene–class and simultaneously identify subset of essential genes. Our method is tested on Diffuse Large B cell Lymphoma (DLBCL) dataset by using well-known classification techniques such as support vector machine, naïve Bayes, k-nearest neighbor, and decision tree.

Results Results on biological DLBCL dataset demonstrate that the proposed method provides promising tools for the prediction of cancer type, with the prediction accuracy of 97.62%, precision of 94.23%, sensitivity of 94.12%, F-measure of 90.12%, and ROC value of 99.75%.

Conclusion The experimental results reveal the fact that the proposed method is significantly improved classification accuracy and execution time, compared to existing standard algorithms when applied to the non-partitioned dataset. Furthermore, the extracted genes are biologically sound and agree with the outcome of relevant biomedical studies.

Keywords Feature selection · Information theory · Spearman's correlation · DLBCL

Introduction

In the field of computational biology, microarrays are used to measure the activity of thousands of genes simultaneously and generate a global portrait of cellular function (Pang et al. 2012). Generally, microarray data are images which have to be transmuted into gene expression matrices in which rows signify genes, columns characterize various samples like tissues, and numbers in each cell characterize the expression level of a certain gene in the particular sample. Commonly, DNA microarray gene expression data comprises thousands of gene expression profiles compared to small number of tissues.

It does not only diminish the generalization performance of the learning method but also increases the computational cost which is mainly produced the curse of dimensionality issue. Therefore, researchers have recently introduced several valuable filter techniques of feature selection that apply specific selection criteria and reduce the size of data sets which provide characteristics relevant to classification model to estimate tumors in the gene dataset. In this study, we address the distinguishing cancer tissue problem from benign tissues by gathering gene expressions from diffuse large-b cell lymphoma (DLBCL) microarray data.

✉ Alok Kumar Shukla
alokjestshukla@gmail.com

[1] Department of Computer Science and Engineering,
G L Bajaj Institute of Technology and Management,
Greater Noida, Uttar Pradesh, India

[2] SRM University, Amaravati, India

Published online: 10 February 2020

In general, humans comprise nearby 25,000 genes, only a fraction of these are keenly expressed as mRNAs at any once. If all the genes were expressed at once, this makes sense subsequently, we can get a confused stream of unstructured information. In the recent biological study, the popular mRNAs of non-hodgkin lymphoma (NHL) is diffuse b-cell lymphoma and explanations for 31% of NHL cases. In modern chemotherapy, 50% of patients complete durable and disease-free existence (Venkataramana et al. 2019). However, it is a heterogeneous disease, both morphologically and clinically. Before the launch of microarray technology, significant work focused on the find of significant genes that can help physicians measure the hostility of individual cases and give correct therapies.

Mining the valuable knowledge from large amount of data generally refers to data mining approach, in other way; data mining is the technique to ascertain many types of patterns that are innate in the data which are precise, different or/and valuable. It is an iterative process of creating a predictive and descriptive model by uncovering previously unknown patterns in huge amount of data in order to support decision making. Although, machine learning (ML) comprises an algorithm that enriches quality automatically by the experience based on data. The source for ML is data (technically says databases), fundamentally it involves two sets of data such as training data and test data. Commonly, ML uses data mining techniques to build models of what is trendy behind data so that it can predict future outcomes (Nguyen et al. 2019). Over the past few decades, due to the high dimensionality of the data sets, the selection of features has an active research area that consists in perceiving the relevant characteristics and discarding the irrelevant.

Earlier reported literature by applying various machine learning algorithms on the prediction of sub cancer has exposed poor performance, in addition, it elected insignificant subset of irrelevant genes in order to reduce the stability and precision of prediction (Ang et al. 2016). Therefore, Feature Selection (FS) methods become a need in this application. FS methods can be distributed into two categories: filter and wrapper (Shukla et al. 2019a, b, c). In order to maximize the performance, inspired us to relate with filter-based method in gene data that has intrinsically noisy, for the discrimination of tissues. However, we observed that classification performance of the single filter method is not satisfactory in the microarray dataset. Therefore, in this study, features are collected by two stages which can form the subset with lowest cardinality, as stated by the corresponding range of the function with minimum number of features. Thus, many different biomarkers are introduced that have better predictive accuracy. By applying the two-stage strategy, we need to apply spearman correlation (SC) with distributed filter feature selection method for identifying the cancer subtypes.

From the experimental study, we demonstrate that it is able to predict with great performance in terms of accuracy, precision, F-measure, ROC value, and sensitivity which outperform other FS techniques working on relevant problems and reduced the execution time (Etime). The most substantial statement of microarray in the analysis of gene expression is to classify the unfamiliar tissue according to their gene expression levels with the help of the expression levels of identified samples. We consider this unbiased in our work and select the high-quality characteristics of the available gene dataset that can be useful for the general practitioner to arrive at an accurate diagnosis. The rest of this paper is organized as follows. Section 2 briefly illustrates the filter-based feature selection methods, mutual information, as well as the relevant criteria for feature inclusion. In Sect. 3, we first introduce several definitions of filter methodology and then detail the proposed feature selection methods are illustrated. Experimental setting and results, in Sect. 4 analyzes the theoretical space and execution time. Finally, we conclude with a brief summary.

Related work

Recently, the fields of bioinformatics have made prodigious growth and have led to in-depth analytics that is demanded by generation, collection, and increase of substantial data. Meanwhile, we are entering a new period where innovative technologies are starting to analyze and explore knowledge from tremendous amount of gene data, bringing unlimited potential for information growth. From the reported literature (Guyon and Elisseeff 2003), some faults in analytical systems for tissue classification, such as interrelationships between resolve factors have not been considered and sprightly modifies in disease prediction direction after some time have been disregarded while effective finding strategy can cause better control of the circumstance and make better solutions over long time. Recently, many methods have studied which is applied to various applications including classification, is perilous to maximizing the classification performance to predict the disease accurately (Shukla et al. 2019a; Liu et al. 2017).

In Qu et al. (2019), author established a new hybrid technique for feature selection, called ensemble multi-population adaptive genetic algorithm that can manage the irrelevant genes and classify cancer accurately. The proposed hybrid algorithm comprised of two-phase. In the first phase, an ensemble gene selection method used to filter the noisy and redundant genes in high-dimensional datasets by combining multi-layer and F-score approaches. Then, an adaptive genetic algorithm based on multi-population strategy with support vector machine and naïve Bayes classifiers as a fitness function is applied for gene selection to select the extremely sensible genes from the reduced datasets. To find the small subset of genes in biological data with the most significant quantity of precision by new bacterial colony optimization method with multi-dimensional population, called BCO-MDP, was offered for gene selection for the purpose of classification (Wang et al. 2019). To address the combinational problem related to feature selection, population with multiple dimensionalities was represented with the help of subsets of different feature scopes. The population is grouped in terms of Tribes. The sizes of the feature subsets within a tribe were equal, while the dimensionalities differ when they belong to different tribes to achieve parallel solutions.

Due to the large capacity of data intricate in therapeutic contexts and disease diagnosis, the provision of the intended treatment method becomes almost impossible over a short period of time. This justified the use of pre-processing techniques and data reduction methods in such contexts. In this regard, clustering and metaheuristic algorithms maintain important roles. Alirezaei et al. (2019) have introduced a method based on k-means clustering was first utilized to detect and delete outliers. Then, in order to select significant and effective features, four bi-objective metaheuristic algorithms were employed to choose the least number of significant features with the highest classification accuracy using support vector machines (SVM). In Dara et al. (2010), author concerned with the analysis of training data distribution and its impact on the performance of multiple classifier systems. Several feature-based and class-based measures were proposed. These measures can be used to estimate statistical characteristics of the training partitions. To measure the effectiveness of different types of training partitions, author generated large number of disjoint training partitions with distinctive distributions. Then, they empirically evaluated these training partitions and their impact on the performance of the system by utilizing the proposed feature-based and class-based measures. As a comparison to other classifiers, support vector machine showed the extreme advantage for tumor classification due to its statistically proficient, computationally fast with a smaller number of parameters, easy to implement, and able to perform FS (Shukla et al. 2020).

Traditional algorithms designed for executing on single machine lack scalability to deal with the increasing amount of data that has become available in the current Big Data era. In Daniel and Luis (2019), author presented a completely redesigned distributed version of the popular ReliefF algorithm based on the novel Spark cluster computing model, called DiReliefF. Similarly, Zhao et al. (2016) introduced a new method for useful analysis on big data based on distributed FS. Explicitly, to expose the unseen patterns for economic development, designed framework combined the efficient FS methods and econometric model building. For partial least squares-based gene microarray analysis, from Wu et al. (2017) proposed the novel gene selection method, called distributed GPU selection. It was used to identify inclusive genes associated with the primary disease to quantify, consistency and uniqueness of the differential gene expressions across many cross-validations or randomization methods regarding existence and randomization p values. It was used to accelerate the gene selection problem and about 8–11 times speed-up the process on the microarray datasets (Palma-Mendoza et al. 2018) called Distributed CFS.

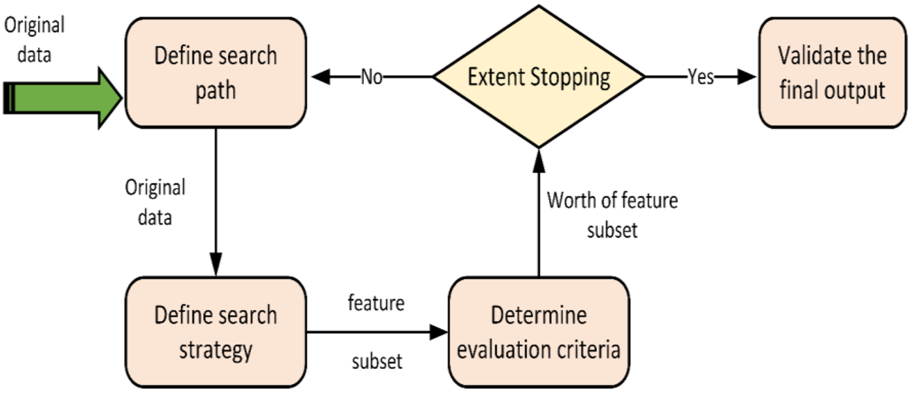
Multi-label learning generalizes outdated learning by allowing an instance to belong to multiple labels simultaneously. This causes multi-label data to be characterized by its large label space dimensionality and the dependencies among labels. These challenges have addressed by feature selection techniques which improve the ultimate model accuracy. In Gonzalez-lopez et al. (2019), author proposed a distributed model to compute a score that measured the quality of each feature with respect to multiple labels on Apache Spark. In addition, author proposed two different approaches that study how to aggregate the mutual information of multiple labels such as Euclidean norm maximization and geometric mean maximization. In the previous section, positively several filter method have employed to classify microarrays, to best our knowledge, there has been no determination in the available literature to tackle distributed problem by two-stage feature selection.

Methodology

The calculations of the classification model are deeply influenced by the excellence of the input features. Generally, it can be observed by reduction of dimensionality method which aims to choose a small subset of the relevant characteristics of the original attributes by eliminating irrelevant, redundant or noisy features can see in Fig. 1. There are two main approaches that modify the input space: feature transformations (e.g. principal component analysis) and feature selection (e.g. mutual information, t test). Feature selection identifies a subset of relevant features in the original space whereas feature transformation generates new feature space of smaller dimensionality also can generally lead to better classification performance, i.e., higher learning precision, lower calculation costs, and better model interpretation. Recently, researchers in bioinformatics, data mining proposed a variety of distinct FS algorithms regarding theory and experiment and shown the effectiveness of their work (Fabris et al. 2016).

The most existing methods are used in the proposed approach in this paper for gene selection which are information theory, similarity-based, and statistical based concept (Hu et al. 2018). To clarify the differences between proposed

Fig. 1 The overall process of feature selection



approach and the other methods, we present more detailed description of the three methods as follows.

Filter methods

Recently, increasing awareness has focused on the study of filter-based gene selection methods, and so far, numerous filter gene selection methods have introduced in the previous studied. Since to our awareness, there is no broad discussion of the filter-based gene selection methods for gene selection. In Ang et al. (2016), author has presented more extensive range of surveys and provided the straightforward organization of gene selection and reviews filter, wrapper, and hybrid methods in the high dimensional datasets. We observed that introduced investigation has not systematically studied the filter gene selection methods on microarray and also less contribution to the description of data characteristics.

In general, filter methods have been employed the ranking measure instead of the accuracy to qualify subset of characteristics. This measure is chosen to be rapid to calculate, at the same time as it captures the utility of the set of features. The most popular measures include mutual information, Pearson correlation coefficient, and distance inter or intra-class scores for each class/combination of characteristics (Gutkin et al. 2009). Filters method to be less expensive regarding computational burden widespread than wrappers but yield a set of genes that are not adapted to an explicit type of learning model (Ruiz et al. 2006). However, filter FS does not understand the expectations of a prediction model, so it is more useful to expose the relationships between the features. Various filters provide a classification of characteristics instead of a subset of the best specific feature and the limit point in the ranking.

Information theory concept

In this section, we explain main concepts of Shannon's theory for FS, which is mainly employed to choose a group of more helpful genes that can be used for the application of machine learning and bioinformatics. It addresses information sighting and information analysis but not so much information synthesis and measure of the uncertainty of random feature. The level of uncertainty is connected to the probability of the fundamentals composing the feature. The concept of uncertainty can understand as the measure of how much information is needed to describe the element. Intuitively, high entropy indicates that the elements in the feature have about the same probability of occurrence, while low entropy means larger differences in the probabilities of occurrences. Thus, entropy is related to the probabilities of the feature rather than the actual values.

Most of the existing FS methods used mutual information methods that are fruitless for detecting higher-order feature interactions. To fill related shortcomings like the relevance of the features defined without distinguishing the significance of the candidate characteristics and the importance of the selected features by two-way interactions for feature selection. We first identify comfortable assumption to decompose the mutual information-based feature selection problem into low-order interactions. A direct calculation of the disintegrated interaction terms is computationally expensive. Therefore, we used joint mutual information to estimate the interaction terms with computationally efficient measure as shown in the selection approach Eq. (1).

Joint mutual information To address interdependability difficulty, researchers have been introduced a nonlinear feature selection methods, one of them is joint mutual information (JMI); it is also testified on several publically high dimensional data sets and compared to state-of-arts competing methods. It is determined by mutual information as demonstrated in Eq. (1). To be able to measure the MI between attributes such as class label y and features X is precise as:

$$I(y;X) = H(y) - H\left(\frac{y}{X}\right) \tag{1}$$

The entropy and conditional entropy of the attributes is defined as $H(y)$ and $H\left(\frac{y}{x}\right)$, respectively. The significance of attributes is shown in Eq. (2).

$$M_{JMI}(X) = \sum_{x_j \in S} I\left(x_k; x_j; y\right) \propto \sum_{x_j \in S} I\left(y; x_k / x_j\right) \tag{2}$$

where $I\left(x_k; x_j; y\right)$ shows the MI between inventive the attribute set x_k and selected attribute x_j with respect to the class y.

Minimal-redundancy-maximal-relevance A filter feature selection technique has been widely used to excavation biological data. Recently, in the classical filter method called minimal-redundancy-maximal-relevance (mRMR), is available in information theory that revealed that a specific part of the redundancy, called irrelevant redundancy, may be involved in the minimal-redundancy component of this method (Peng et al. 2005). To find the maximum dependency with respect to target class y, new approach is introduced called max-dependency described in Eq. (3).

$$maxw(X, y) = I\left(y; x_1, x_2, \dots, x_N\right) = H(y) - H\left(\frac{y}{x_1, x_2, \dots, x_N}\right) \tag{3}$$

Equation (3) shows that dependency between X and y estimated value can be large. The relationship between redundancies between features is expressed by Eqs. (4) and (5).

$$\min Z\ (\mathrm{X}, \mathrm{y}\ ) = 1/|s^2| \sum_{x_j \in s} I\left(x_j; x_k\right) \tag{4}$$

$$\mathrm{Max}\, \emptyset(w, Z) = \mathrm{w}{-}\mathrm{Z} \tag{5}$$

The integration of Eqs. (4) and (5) known as minimal-redundancy-maximal-relevance which describes according to Eq. (6).

$$j_{mRMR}\left(\emptyset\right) = I(y; X) - 1/|s^2| \sum_{x_j \in s} I\left(x_j; x_k\right) \tag{6}$$

where x_j is selected subset of features and x_k is original features set.

Information gain Information gain (IG) measures the quantity of information about the label likelihood, if the only information available in the presence of a feature and the resultant class distribution. Concretely, it measures the expected reduction in entropy (uncertainty associated with a random feature). As shown in Eqs. (7)–(9), measures the amount of each gene according to information gain concerning the class.

$$\mathrm{IG} = H(X) - H\left(\frac{y}{X}\right) \tag{7}$$

$$H(X) = -\sum_{x}^{X} P(x) \log(P(x)) \tag{8}$$

$$H\left(\frac{y}{X}\right) = -\sum_{x}^{X} P(x) \sum_{y}^{Y} P\left(\frac{y}{x}\right) \log P\left(\frac{y}{x}\right) \tag{9}$$

where X and Y feature. $P(x)$ and $P\left(\frac{y}{x}\right)$ is the probabilities distribution of x and y.

Similarity-based feature selection

Relief-F Feature selection is a valuable technique to deal with dimensionality reduction. In classification, it is used to find an optimal subset of relevant features so that the overall accuracy is increased while the data size is reduced. To increase the performance of classifiers, we have selected the optimal genes from the high dimensional dataset through Relief-F (Sun 2007) approach; it is a revised version of Relief (Shukla et al. 2019b). It can also discriminate conditional dependencies between features that current in the candidate dataset and provide a unified view of the features evaluation in the classification process. In contrast to Relief-F, does not openly reduce the relevancy in selected genes. It elites the high weighted features that distinguish the tissues from neighbors of different categories. The weight of each ith feature (wi) is updated according to Eq. (10).

$$w_i = w_i - \frac{\psi(A_i, R, H)}{n} + \sum \frac{p(c) * \psi(A_i, R, n(c))}{n} \tag{10}$$

where R is a trial instance sample from n instances in different sample category, the function $\psi(A_i, R, H)$ evaluates the distance between instance samples (R) and nearest hit (H) or miss n(c).

Statistical based feature selection

Chi square Chi Square (χ^2), which is standard feature selection method, is evaluated genes individually with respect to the labels. The range of continuous-valued features needs to be discretized into intervals. Chi square is based on comparing the obtained values of the frequency of a class because of the split to the expected frequency of the label. The scientific formulation of χ^2 depicts in Eq. (11) for calculating the value for two adjacent intervals.

$$\chi^2 = \sum_{j=1}^{2} \sum_{k=1}^{l} \frac{\left(\alpha_{j,k} - \beta_{j,k}\right)^2}{\beta_{j,k}} \tag{11}$$

where l represents the class labels, $\alpha_{j,k}$ shows the instances present in the jth interval with label k, Rj shows instances in the jth range, l_k shows the is instances into label k in the interval of two, N shows the instances in the interval of two, and $\beta_{j,k}$ is the expected frequency of $\alpha_{j,k} = R_j * l_k / N$.

T-test In the classification problem, each tissue can be classified either class l1 or class l2. It can helps us to assess whether the values of precise feature for class l1 are significantly different from benefits of the same functionality for class l2. If this holds, then the feature can help us to differentiate our gene data better. Using this, we compute the mean and variance of the observations separately. The scoring functions are portrayed as a function f(s1, s2) with s1 and s2 gene instances. The scoring ability is represented in gathered in two structure such as fold-change as displays in Eq. (12).

$$t(X_i) = \left| \frac{\left(\mu_{i,1} - \mu_{i,2}\right)}{\sqrt{\left(\frac{\sigma_{i,1}^2}{n_1} + \frac{\sigma_{i,2}^2}{n_2}\right)}} \right| \tag{12}$$

where, $\mu_{i,j}$ shows the mean of ith feature Xi for class Cj and σij shows the standard deviation of ith feature Xi for class Cj. The class index is denoted by j i.e. j = 1 or j = 2. After assessment the values of t-test for each gene, it can be sort in descending order based on the score in order to select the important genes.

Classification algorithms

Classification algorithms often have faced difficulty to deal with high dimensional data sets that include a large number of features, which seriously increase the temporal and spatial complexity. Many of the features in an input data set are irrelevant or redundant and must be eliminated. Feature selection is the process of identifying the subset of features that can allow the classifier to perform most effectively. Several earlier studies have combined classification algorithms with feature selection methods based on ranking methods. In this study, we employed four different machine learning algorithms such as Naïve Bayes (NB), Support Vector Machine (SVM), k-Nearest Neighbor (k-NN), and Decision Tree (DT) (X. Wu et al. 2008) (Kohavi 1995) for better cancer prediction.

Naïve Bayes

In machine learning, Naive Bayes (NB) is a native simple probabilistic classifier based on Bayes theorem; it is the extension of Bayes theorem with the hypothesis of independence of all features. Let select the frequent class of instance C from the dataset and assume the random feature vector $X = \left(X_1, X_2, \dots, X_i\right)$ by using the observed features. Let c_j represent jth class label and $x = \left(x_1, x_2, \dots, x_i\right)$ represent a predicted feature vector. To identify the correct class label of a testing instance x, so we can use Bayes' theorem to calculate explicit probabilities as express in Eq. (13).

$$Pr(C = c_j | X = x_{1,\dots,i}) \alpha Pr(C = c_j) \prod_{i=1}^{m} Pr(X_i = x_i | C = c_j) \tag{13}$$

We note that one may associate a notion of a minimalist with the definition of a naïve Bayesian as done (Friedman et al. 1997).

Support vector machine

Separating the feature vector and predict the correct label is the vital task for classification algorithms. Support Vector Machine (SVM) is one of the important discriminative classifiers formally defined by separating hyperplane. In other words, given labelled training data, the algorithm outputs an optimal hyperplane which groups new samples (Edsgärd et al. 2018). It has become nearly usual technique in many domains there are several numbers of reasons. First, SVM objective is maximizing margin, has a theoretical basis tied to achievement of good generalization accuracy (see Fig. 2). Second, there is unique, globally optimal solution to the SVM training problem. Third, there are improvements in representation power through nonlinear kernels, which map to high or even infinite-dimensional feature space and, via the kernel trick do so without vast increase in decision making and classifier training complexities. Fourth, SVM achieves good results on a variety of domains such as intrusion detection and pattern recognition.

Decision tree

In the reported literature, several learning engines such as statistical classifier, support vector classifier, and other classifier. Here, we selected a tree-based classifier because of its simple properties, the explicit meaning, and easily transformation to if-then rules. The definition of decision tree is decision support device that uses dendritic graph and their possible concerns, including accidental event outcomes, resource costs, and utility. Decision trees are commonly used in bioinformatics, specifically in decision analysis in order to help identify a type of disease most likely to reach a specific symptoms. Another use of decision trees is descriptive means for manipulative conditional probabilities. It works on divide and conquers (Quinlan 1993) approaches for constructing a decision

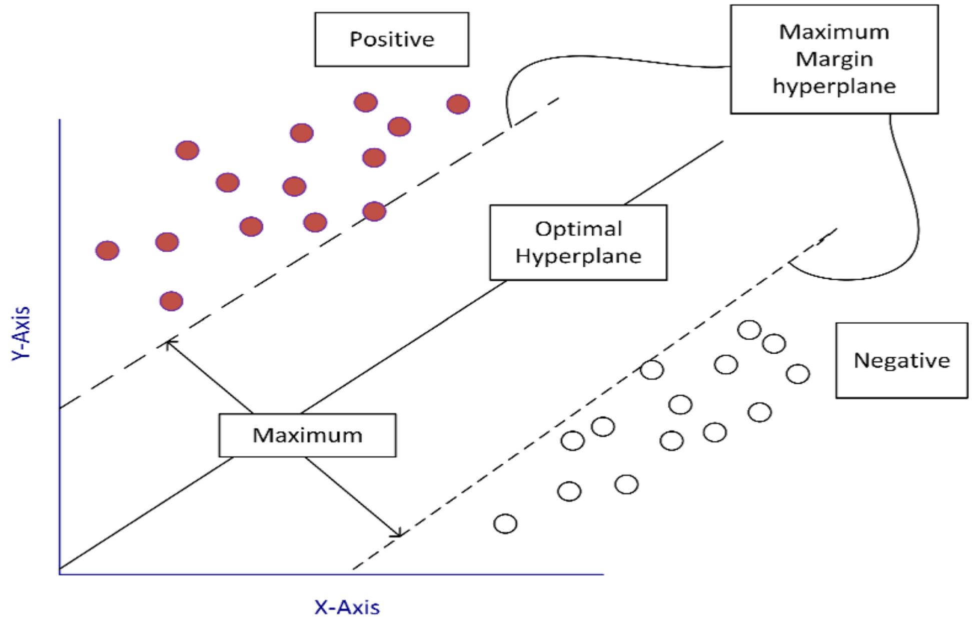


Fig. 2 SVM classification process

tree. As shown in Eqs. (14) and (15), the gain ratio is a parameter of the decision tree to measure the performance. It is defined as:

$$\text{Entropy (D)} = -\sum_{i=1}^{n} p_i \log(p_i) \tag{14}$$

$$\text{Gain Ratio} = \frac{Gain(p)}{SplitInfo(p)} \tag{15}$$

The function Split into is described in Eq. (16)

$$\text{Split Info (p, test)} = -\sum_{i=1}^{n} p\left(\frac{i}{p}\right)\log\left(p\left(\frac{i}{p}\right)\right) \tag{16}$$

where p represents the probability distribution of the data sample, and as Eq. (16) mentioned above uses the log function as base two due to measuring statistics as 'bit'.

K-nearest neighbor

The recent trends of classification problems, K-nearest neighbor is applied for distinguishes the classification of an unknown data point on the basis of its nearest neighbor when the class is previously known. It can solve real-world problems with the availability of the inexpensive platform. Stevens et al. (1967) investigated the purpose of k-nearest neighbor (k-NN), is to find a group of k objects in the training set which is nearest to the test object, and based on the label of the majority of particular dataset (Han et al. 2006) in this neighborhood.

Proposed methodology

Let $£_{\mathrm{n}} = \{£1, £2, \ldots £\mathrm{n}\}$ represents the group of genes and $\mathrm{S}_j = \{\mathrm{s}_1, \mathrm{s}_2, \ldots, \mathrm{s}_{\mathrm{M}}\}$ represents a group of samples. The vector representation of microarray dataset is expressed as $\mathfrak{R}_M^n$. The n genes form feature space ($\mathfrak{R}_M^n$) corresponding to an instance space (τ);.it can be represented as $f : \mathfrak{R}_M^n \to \tau$. More interestingly, there is l class label for all the given instances then there exists an association between gene expression patterns and the ith class which can be defined as:

$$l^i = f(\varrho_j^n)$$

where $l^i \epsilon \tau \, \forall 1 \le i \le c$ and $\varrho_j^n \, \epsilon \, \mathfrak{R}_M^n$. Generally, gene expression dataset has ϱ_j^n matrix which ordered sequence rows is concerning to n genes for the jth instances.

$$\varrho_j^n = (\varrho_1, \varrho_2, \ldots, \varrho_{\mathrm{M}})$$

Suppose that complete microarray dataset associated with M tissues or observations which is described as:

$$\mathfrak{R}_M^n = \{(\varrho_j^n, l^i) : j = 1, \ldots, M, i = 1, \ldots, c\} \subseteq \varrho_j^n * \tau$$

The vector ϱ_j^n (j = 1, 2,…, M) consists of gene expression values for n number of genes and $l^i \in \{1, 2, 3, \ldots, \mathrm{c}\}$ is the class label assigned to the gene expression vector.

The increasing microarray related activities provide a wide range of indicators and tissues for microarray analysis. Facing such a large amount of data, how to detect useful information from it has drawn extensive attention in bioinformatics domain. Traditional filter methods cannot embrace the high-dimensionality data since they only involve limited economic factors for model construction based on past experiences. For example, some scientists analyze economic development from the perspective of bioinformatics structure. Besides, the existing statistical analysis software would generate runtime errors when dealing with the high-dimensionality and huge-volume biological data. While some methods are able to process massive data, their computation costs are expensive. Therefore, we aim to provide an efficient way to bridge the gap between data analysis methods and in biological real-world dataset. Therefore, in this work, we introduce a distributed filter approach that can help to increase classification performance results on microarray data and reduced the computational time.

From the available literature, researchers have been observed that existing FS method showed the composite computational burden to determine the relevant genes from the data due to the indecency of dimension. Keeping this in mind, in this study, we have developed two-stage distributed correlation-based FS approach by using Spearman's correlation and well-known filter FS approach that can select the highly discriminative genes for distinguishing tissues. Here, we focused on three main aspects: Calculate the Spearman

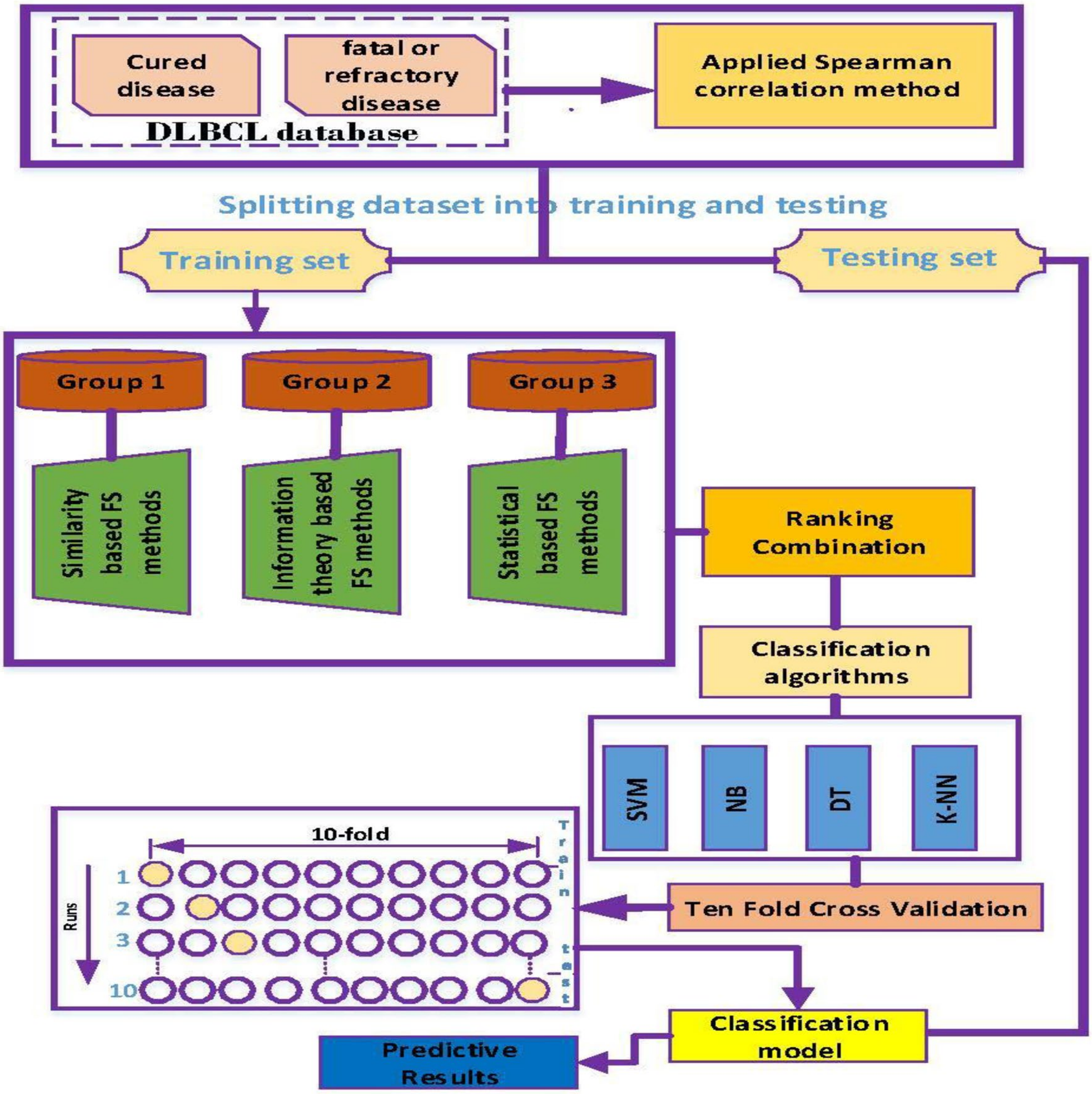

Cured disease
fatal or refractory disease
DLBCL database
Applied Spearman correlation method
Splitting dataset into training and testing
Training set
Testing set
Group 1
Group 2
Group 3
Similarity based FS methods
Information theory based FS methods
Statistical based FS methods
Ranking Combination
Classification algorithms
SVM
NB
DT
K-NN
10-fold
Runs
1
2
3
10
Train
test
Ten Fold Cross Validation
Classification model
Predictive Results

Fig. 3 Proposed evaluation framework

correlation between gene and label, after that relevance/redundancy concept (i.e. Dispersion measure) (Ferreira and Figueiredo 2012) is used to select ideal genes then distributed filter ranking base feature selection methods are applied. In addition to this, in computationally speed and scalability, proposed method is less expensive as compared to other filter techniques.

The proposed method involves several filter methods that are applied to partitions of the data, combined after that into unique features ranking to estimate the performance. Here, we divided dataset D into several small disjoint subsets Di. The filter method is applied to each of them, generating FS ranking on Si. After all the small datasets Di have been used, the combination method builds the final ranking of S as the result of the filter process.

The main objective of this study is enterprise an efficient framework to select a relevant set of features that are capable to enhance the classification performance. By using developed framework, we have selected top-most r features according to the ranking of correlated features corresponding to labels. It can evaluate by using evaluation equation, is highly correlated with the class and uncorrelated to all features. The overall procedure of the proposed method are:

Step 1: Initially, Spearman correlation (Yu and Liu 2004) method is applied to measure the tremendous goodness of genes based on the correlation of features ρ; mathematical expression shown in Eq. (17). According to these two characteristics: firstly, we can decide which genes are meaningful with respect to class or not; and then decides whether such relevant gene is redundant or not when considered it with other meaningful genes. Select the high ranked features and permit to subsequent step for better classification.

$$\rho = \frac{\sum_{i=1}^{i=n}\left(x_i - \bar{x}\right)\left(\left(y_i - \bar{y}\right)\right)}{\sqrt{\sum_{i=1}^{i=n}\left(x_i - \bar{x}\right)^2 * \sum_{i=1}^{i=n}\left(y_i - \bar{y}\right)^2}} \tag{17}$$

where ρ shows the Spearman correlation, x_i and y_i are the randomly selected genes from n number of samples in different categories of samples,$\bar{x}and\bar{y}$ shows the mean of each random gene subsets. Each gene is sorted in ascending order. And calculate the rank of each gene based on the score function value of selected genes.

Step 2: Separating the dataset into some disjoint subsets of nearby the same size that cover the complete dataset. In this study, the distribution is completed as vertically. From the reported literature (Shukla 2020), two, unlike methods, are used for partitioning datasets such as random partition and ranking the original features before generating the subsets. The second method is lively in the current direction of research to improve performance. By having an ordered ranking, features with similar relevance to the class will be in the equivalent subset, which will support the task of the subset filter which can be applied later. This technique for partitioning the data will generate FS approach for the distributed method called distributed filter FS (DFS).

Step 3: Used well-known filter FS method after the SC method and then Dispersion measure is employed for getting the ranking of features so, we can select the relevant subset of genes at a low dimension. Our aim is to choose the significant features and to find the maximum score and important, relevant predictive subset features with the size of n and the most suitable optimal subset of size m.

One probable solution to above-mentioned issue is design which can be seen in Fig. 3. In the proposed method, front-end strategy (by using SC method) adopts a correlation concept, to evaluate the degree of association between features while the following way accelerates the search and identify the essential biomarker to better classification performance.

Algorithm 1: Distributed filter FS method

```
Input: D - $\varrho_j^n \times l^i$
        €- Performance measure through learning algorithms
        $\delta -$ Threshold value
        $\Theta_j$- Number of ranker methods
Output: $\partial$ - Compact datasets                    //reduced dataset with minimum features
Initialization: $F' =$ NULL and $\chi_R = \{\}$
begin
R(X,Y) = SC (D)                                           // Apply Spearman correlation
$\Delta$ = Split the training dataset (X) in n(R)/ $\Theta$ times;   // number of selected genes in each bunch of
genes
 For $\aleph = 1$ to $\Delta$                              //Distributed feature ranking
   For each $\Theta_j \epsilon R_{\Delta,\aleph}$
     Evaluate $\chi_R$ for $R_\Delta$ with the help of ranking vote scheme   // filters method based on ranking
   End for
        $\boldsymbol{select}\ (\boldsymbol{f}_\aleph) \leftarrow \boldsymbol{f}\ (\chi_R,\delta)$
      select the optimal feature sets as F' from $f_{m'}$ ; $F' = F - f_\aleph$
    End for
return $\partial(\varrho_j^{m'} \times l^i)$
For i=1: € do
        Randomly split using ten cross-validation
        Training set X with $F'$
         Testing € classifiers from testing data Y
         Evaluate performance
End for
     Return Accuracy, Precision, Sensitivity, F-measure, and ROC value.
End
```

Performance measures

We evaluate the classification performance of proposed and other methods using four classifiers, i.e., SVM, DT, k-NN and NB. The performance is assessed by four measures, such as accuracy, sensitivity, precision, and f-measure on DLBCL dataset. These performance measures are defined as:

$$\text{Accuracy} = \frac{T_p + T_N}{T_p + T_N + F_N + F_P}$$

$$\text{Sensitivity} = \frac{T_p}{T_p + F_N}$$

$$\text{Precision} = \frac{T_p}{T_p + F_P}$$

$$\text{F-measure} = \frac{2T_p}{2T_p + F_N + F_P}$$

Here T_p, T_N, F_p, and F_N are true positive, true negative, false positive, and false negative in the independent datasets. Based on the confusion matrix, we evaluated the performance of the proposed method and rival gene selection.

Experimental results and analysis

In this section, the experimental results of the proposed method have been applied to the DLBCL database to analyze the system efficiency. The software for simulation is Matlab R2016 and the hardware configuration is an Intel Core i7 processor with 8 GB of RAM and 2.40 GHz CPU on the platform Windows 8. DNA microarrays indeed are a relatively variety-new, complex technology found in molecular biology and medicine recognized in Fig. 4. In this experiment, we run our algorithm on diffuse large B-cell lymphoma benchmark dataset, and compare the results with three other measurements: Statistical based, Similarity-based, and Information theory based on commonly used biomedical gene expression data. We calculate the overall performance of popular gene selection methods using microarray cancer dataset as DLBCL, which was downloaded from http://www.gems-system.org. The dataset reports for a total of 47 patients, 24 of them are from germinal center

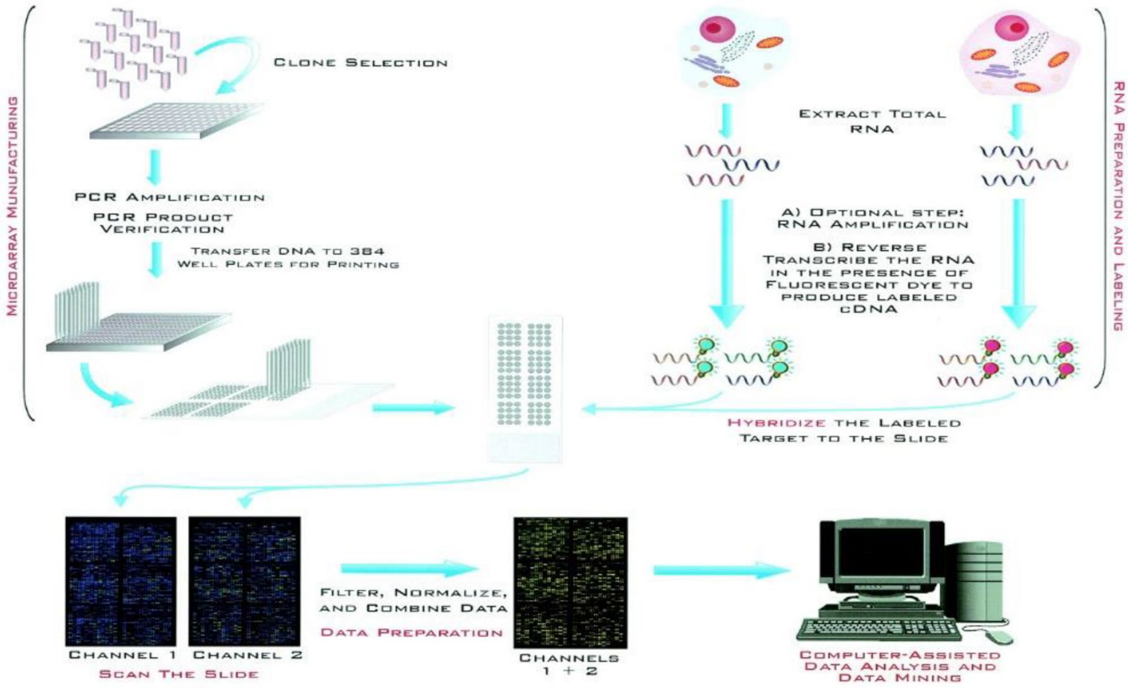

Microarray Munufacturing
Clone Selection
PCR Amplification
PCR Product Verification
Transfer DNA to 384 Well Plates for Printing
Extract Total RNA
A) Optional step: RNA Amplification
B) Reverse Transcribe the RNA in the presence of Fluorescent Dye to produce labeled cDNA
RNA Preparation and Labeling
Hybridize the Labeled Target to the Slide
Channel 1
Channel 2
Scan The Slide
Filter, Normalize, and Combine Data
Data Preparation
Channels 1 + 2
Computer-Assisted Data Analysis and Data Mining

Fig. 4 DNA microarray process (Macgregor and Squire 2002)

B-like group while 23 are activated, B-like groups. Samples contain the expression measures of 4026 genes.

Comparison of existing filter methods and proposed method

In the experimental study, we show the results of the application of algorithms for the selection of filter-based characteristics in the DLBCL data. To evaluate the performance of the proposed algorithm, firstly we chose the hundred best genes through Spearman's correlation (SC) and then applied the distributed feature ranking method to minimize the related features selected through three measures of FS that are based on similarities and statistics and evaluated performance with the four classifiers such as SVM, k-NN, where k is 3, DT and NB tested with ten times the cross-validation. Six filter methods and proposed methods employed to measure the quality of the models regarding the accuracy, precision, sensitivity, f-measure and execution time seen in Table 1.

From Table 1, we observed that the classification results obtained from the DLBCL data by the proposed method and six filters gene selection methods. We can see that the proposed method is outperformed in all most each classifier. Although the upgrading is not substantial, such a movement of upgrading is obvious.

From Table 1, we perceived that the tenfold CV method is employed on the top 100 gene subsets to achieve better classification performance using various classifiers regarding classification accuracy, sensitivity, precision, and f-measure. According to Table 1, SVM classifier has higher performance in DLBCL dataset. Besides the high performance, robustness is an important factor in evaluating a classifier. This shows that Support vector machine is a robust classifier. In Table 1, we can find that the F-measure obtained with the SVM classifier using the proposed as 93.52%. These results are reasonably better than the other learning classifiers acquired: the comprehensive maximum accuracies by the proposed method using each classifier set in the interval of 83.63% to 97.62%, where 97.62%, 95.12%, 90.87%, and 83.63% for the SVM, NB, k-NN, and DT classifiers, respectively. This article reveals that it supports the investigation of the other attributes of cancer or tumor, in other words, SVM supremacy on conventional methods when treated with gene data.

In this experiment, we carried out a comparison between the six of the most common filters reported in the literature to rank genes from microarray data can see in Fig. 5. For the DLBCL dataset, there is an extensive enhancement in the accuracy using the FS method of the amalgamation of SC and distributed FS methods for each subset of features starting between 5 and 100. To better illustrate the performance of the classification, the changes in the average classification performance of the four classifiers in the dataset. The number of selected features increases from 5 to 100 in the range of 10. The best results are 90.15%, 98.50%, 96.75%, and 99.05% compared to the DT, SVM, NB, and k-NN classifiers.

Table 1 Average classification performance on top-ranked genes with the help of four classifiers

Classifiers	Methods	Performance				
		Accuracy	Precision	Sensitivity	F-measure	Etime
SVM	mRMR	85.62	82.53	84.12	91.88	0.56
	JMI	74.53	75.63	74.85	75.26	0.85
	Relief-F	86.90	85.22	86.32	84.52	0.24
	t test	68.65	64.32	67.55	68.32	0.86
	Chi Square	69.63	67.12	66.52	67.85	0.25
	IG	79.23	78.56	76.26	72.23	0.42
	Proposed	**97.62**	93.22	93.55	93.52	0.16
NB	mRMR	83.06	83.85	84.01	84.98	0.48
	JMI	69.88	69.05	68.08	69.09	0.67
	Relief-F	80.94	79.35	78	79.02	0.25
	t test	74.04	75.85	74.24	74.32	0.8
	Chi Square	66.91	68.05	71.54	71.89	0.13
	IG	74.89	75.14	75.48	76.75	0.37
	Proposed	95.12	92.62	90.63	91.12	0.27
k-NN	mRMR	84.95	86.25	85.45	86.02	0.52
	JMI	76.26	77.85	76.95	77.53	0.71
	Relief-F	79.54	79.33	78.36	80.99	0.24
	t test	65.88	67.52	63.41	68.56	0.45
	Chi Square	66.47	66.85	65.02	65.98	0.35
	IG	64.25	64.89	65.91	67.43	0.32
	Proposed	90.87	87.12	88.32	87.65	0.18
DT	mRMR	83.39	82.89	84.83	83.19	0.55
	JMI	73.15	70.08	70.42	73.81	0.73
	Relief-F	83.07	83.57	84.93	85.36	0.29
	t test	67.12	68.51	69.45	76.15	0.37
	Chi Square	72.39	73.88	72.86	70.31	0.38
	IG	75.54	73.89	75.58	76.88	0.37
	Proposed	83.63	81.12	82.52	81.02	0.09

Bold value indicate the best value

Even by selecting merely the top 70 genes, proposed method achieves classification accuracy as 90.05% using DT for DLBCL dataset in Fig. 5a. Generally speaking, the comparative study follows from Fig. 5a, that mRMR is superior to the other FS criteria of methods except for proposed method. As we can see in Fig. 5a, mRMR method shows the maximum accuracy as 87.13% on 65 relevant genes and a minimum efficiency as 75.01% on 25 genes with Relief-F evaluated by DT. Further to demonstrate the classification performance, the variations of average classification accuracy of the SVM classifier for benchmark data are displayed in Fig. 5b. The number of selected features ranges from 5 to 100 within the interval of 10. With the help of the SVM algorithm, the proposed strategy shows the maximum accuracy as 99.57% in 100 genes and minimum accuracy as 93.32% in 23 genes.

It is also seen in Fig. 5c that the classification performance utilizing the feature selection based on the proposed is significantly better than the popular FS algorithm such as mRMR, T-test, IG, and JMI. Also, the accuracy of the classification, using recommended is regularly amended for several subgroups of selected genes. The proposed algorithm works very well compared to other feature selection algorithms for DLBCL dataset using NB classifier. The number of selected functions increases from 5 to 100 in the range of 10. NB classifier with the proposed method provides an average accuracy of 96.38% in 100 genes and minimum precision of 93.03% for 25 genes.

Furthermore, Fig. 5d shows that the proposed method, mRMR and Relief-F achieve significant improvements over IG and Chi Square and that Chi square performs slightly better than IG. Proposed and Relief-F converge extremely fast toward the optimal accuracy and generate more promising performance compared with t test and JMI when top 50 ranked genes are used. Figure 6 shows that the classification performance in terms of accuracy, precision, sensitivity, and F-measure.

The comprehensive experimental results of the tenfold CV using the six methods regarding accuracy, precision, sensitivity, and F-measure with the help of four classifiers

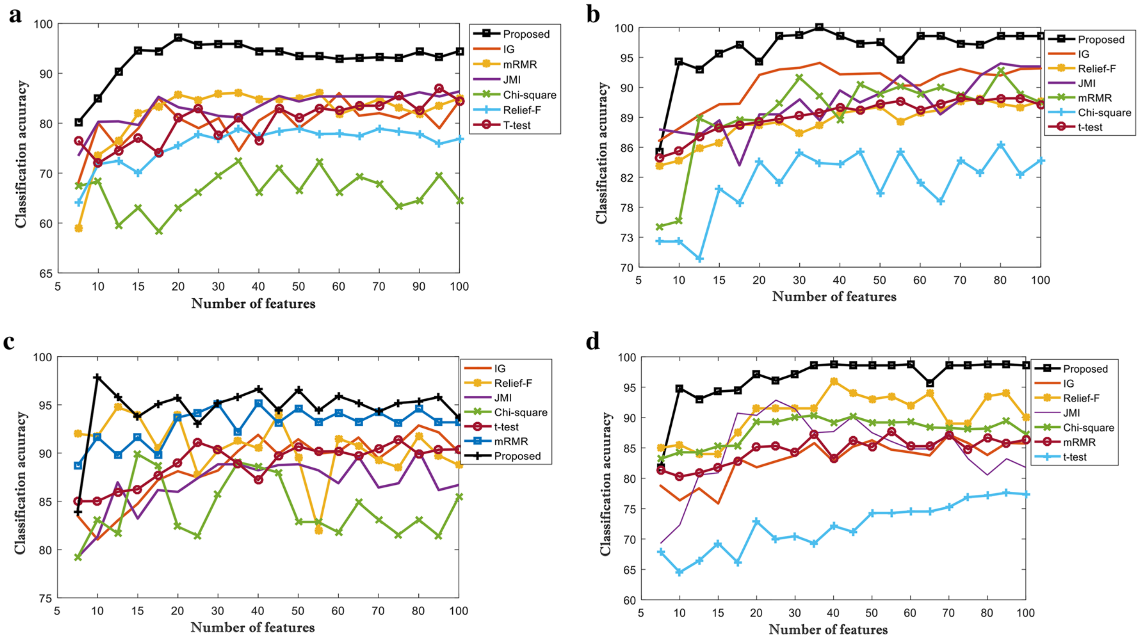


Fig. 5 Classification accuracy on DLBCL based using **a** DT, **b** SVM, **c** NB and **d** k-NN method

are displayed in Fig. 6. In general, the heat map is a graphical representation of data where the individual values contained in a matrix are represented as colors. It is a newer term, but shading matrices have existed for over a century (Mukhopadhyay and Maulik 2013) that helps become an immediate feel for an area by grouping places into categories and displaying their density visually. The consistency of the selected features is analyzed using heat maps showing the frequency of the selected features over the cross-validation in Fig. 7 for the genes selected in the SVM method to identify genes that are differentially expressed between the two disease classes.

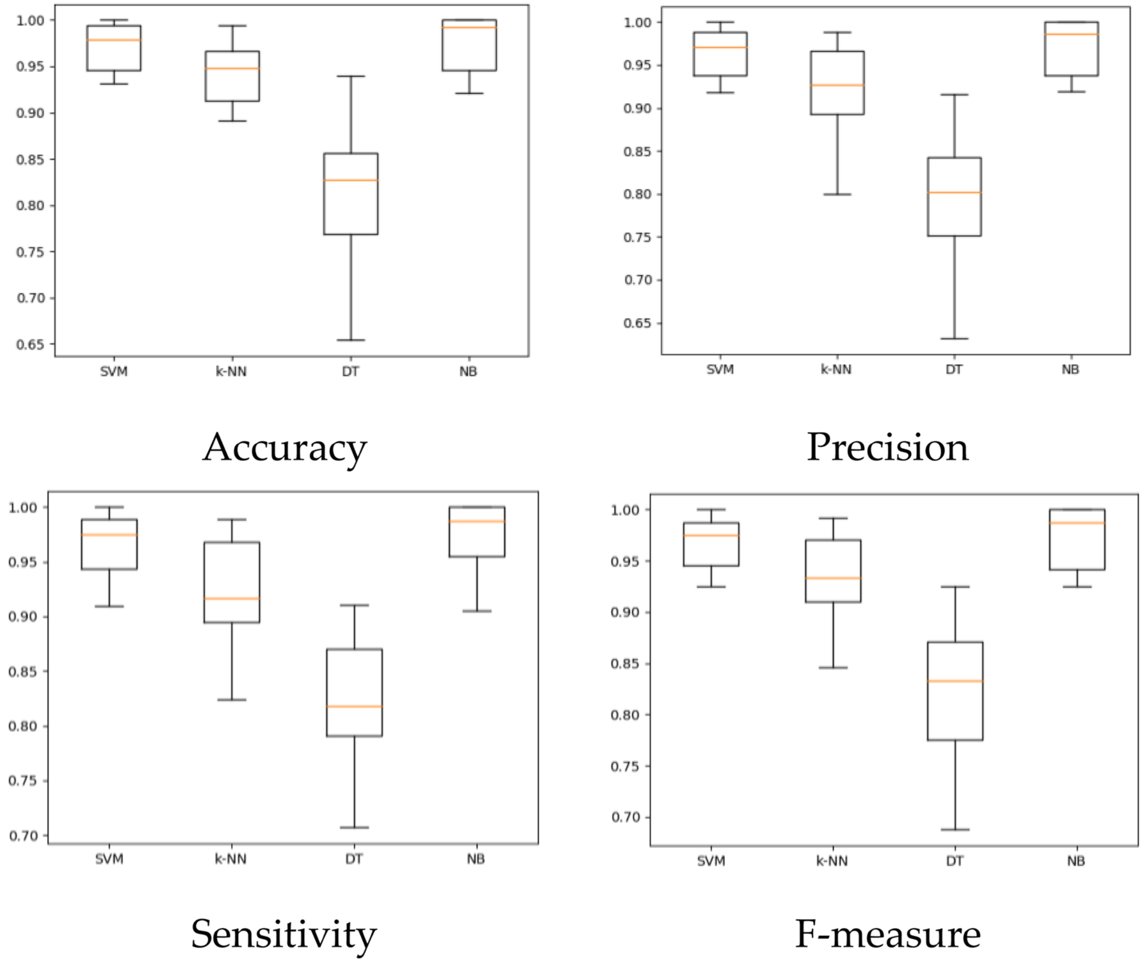
By evaluating the proposed diagnosis model, the average accuracy results of test sets in ten folds for the five methods using every single classifier are shown in Fig. 8. It can be observed from the figure that proposed method performance is superior to those of the other six methods over the whole ten runs. The highest ROC curve value belongs to SVM as 0.997 nearly 1. It means that the support vector machine can classify the data with high precision.

To estimate the performance of each algorithm, the value of k set as 10 using tenfold cross-validation. From Table 2, we show the appreciated statistical measurements of ROC (receiver operating characteristic) values. The area under curves was equal to 0.997, 0.967, 0.896, 0.884, 0.879, 0.864 and 0.804 for proposed method, mRMR, JMI, Chi square, t-test, Relief-F, and IG with the help of SVM classifier respectively. Similarly, from Table 2, we can observe that the highest ROC value of the proposed method in remaining three classifiers as 0.981, 0.954, and 0.946.

Biological interpretation

From the natural point of view, an only small subset of relevant genes obtained from microarray technology is used for diagnostic and prognostic intention of cancer. The objective of the proposed structure is to identify significant gene subsets with the maximum classification accuracy. It is essential to analyze the genes to find biological significance for microarray data. In this section, we investigate the subset of selected genes. Table 3 presents the gene access number and the gene description of the 15 genes selected by the proposed algorithm with SVM for the DLBCL dataset.

Table 4 illustrates the detailed classification results of the various methods in terms of accuracy on the DLBCL dataset. It can be seen in Table 4 that, from among the nine methods, the proposed method performs the best results of accuracy as 97.62%. The detailed comparison results of 10 runs of the tenfold CV for the six filter methods regarding accuracy are displayed in Fig. 8. It can be observed that the proposed method performs better than mRMR and others filter in all ten runs of the tenfold CV, and it achieves comparable results to those of mRMR.



Accuracy

Precision

Sensitivity

F-measure

Fig. 6 Tenfold classification performance of proposed in DLBCL dataset

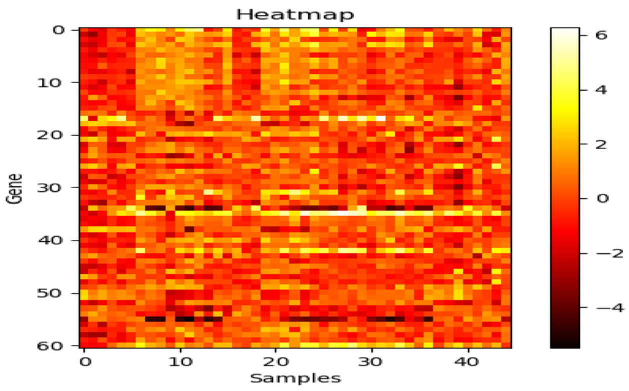


Fig. 7 Heat maps of the genes selected in the SVM predictive model

Conclusion

Two main encounters in the classification process are high-dimensionality and over-fitting. In recent times, there is increasing number of biological datasets exhibiting the characteristics of the combination of over-fitting and high-dimensionality. Filter methods have been successfully applied to solve high-dimensional classification tasks. However, most existing filter methods may suffer from bias performance if the distribution of features is unstructured and has revealed the difficulty in extracting the substantial genes for further data analysis, and until now there is no rule to direct which one could be used for a specified high volume data. To address the problem,

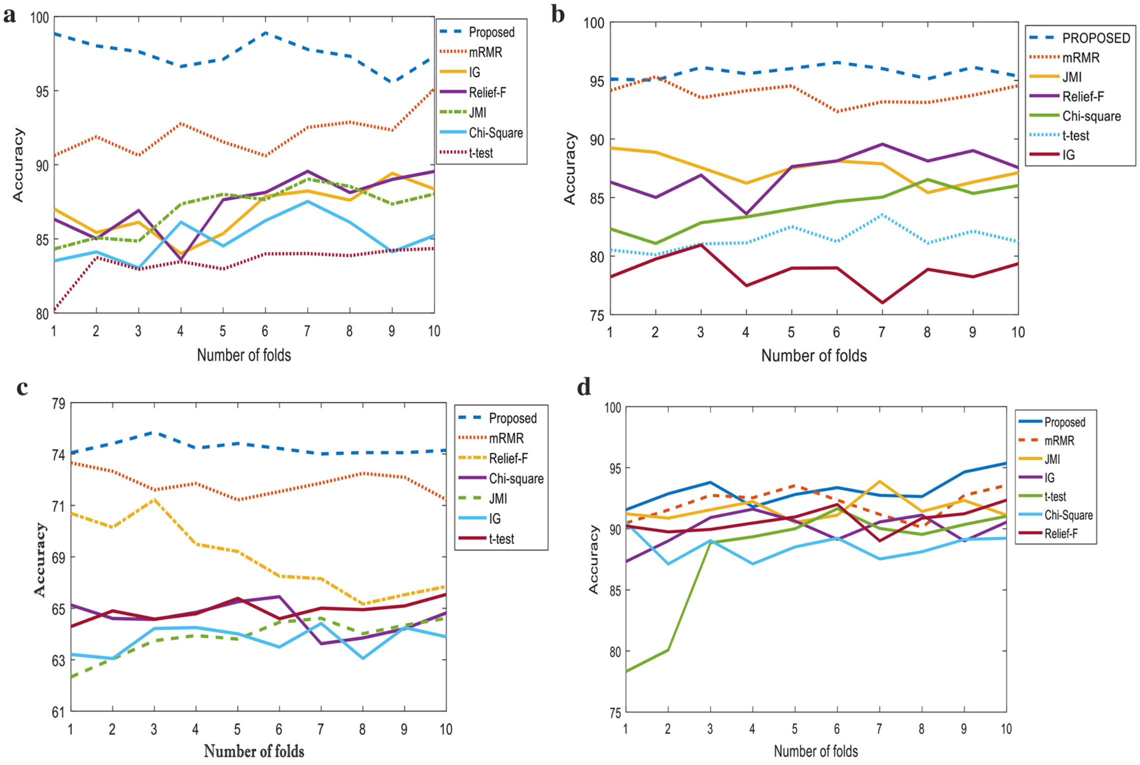


Fig. 8 Tenfold classification performance by using four classifiers such as **a** SVM, **b** NB, **c** k-NN, **d** DT

Table 2 Comparative results of ROC area using four classifiers

Classifiers	ROC-area						
	Proposed	mRMR	JMI	Chi square	t test	Relief-F	IG
SVM	0.997	0.967	0.896	0.884	0.879	0.864	0.804
k-NN	0.981	0.978	0.877	0.844	0.81	0.808	0.818
NB	0.954	0.949	0.861	0.792	0.808	0.795	0.755
DT	0.946	0.932	0.863	0.798	0.791	0.781	0.731

researchers have been employed several powerful feature selection approaches, but they found some limitations. The main goal is to design a method that would be able to successfully distribute the feature selection process. This motivates us to design an operational two-stage framework utilizing Spearman's Correlation (SC) and distributed FS methods which can help to identify the discriminative features and classify disease correctly. The experiment on DLBCL microarray dataset demonstrated that proposed method is able to shrink the execution time significantly in comparison to the centralized filter algorithms along with total number of input features. In terms of execution time, the performance is tremendous, this fact being the most imperative advantage of the proposed method. Furthermore, with regard to classification accuracy, proposed approach is able to match, and even in some cases improve, the standard algorithms applied to the non-partitioned dataset regarding accuracy, sensitivity, precision, and F-measure.

Table 3 Present gene accession number and gene description of the top 15 selected genes by the proposed method

Gene number	Name of genes
AA204640	Unknown UG Hs.193830 ESTs, Highly similar to KIAA0554 protein
AA210719	Similar to novel transcript; similar to transcription factors activation domains; linked at 5′ end to AT-hook motif of HMGI-C
AA211835	CIITA-8 = MHC class II trans activator
AA215688	JkR1 mRNA down regulated upon T-cell activation
AA215651	Trio = LAR trans membrane tyrosine phosphatase binding protein with a kinase domain and separate transpacific and Rho-specific guanine nucleotide exchange factor domains
AA236886	Lst-1 = IC7 = interferon-gamma-inducible gene present in lymphoid tissues, T cells, macrophages, and histiocyte cell lines encoding a trans membrane protein
AA250815	Unknown UG Hs.226360 ESTs, Highly similar to (define not available 4263743)
AA243626	SIP-110 = signalling inositol polyphosphate 5 phosphatases
AA805131	GTP cyclohydrolase 1
AA768003	yotiao = protein of neuronal and neuromuscular synapses that interacts with specific splice variants of NMDA receptor subunit NR1
AA828538	Phosphoribosyl glycinamide formyl transferase, phosphoribosyl glycinamide synthetase, phosphoribosyl aminoimidazole synthetase
AA830622	Unknown UG Hs.173108 Homo sapiens clone 24523 mRNA sequence
AA835845	NK4 = Natural killer cells protein-4 = increased after activation of T cells by mitogens or activation of NK cells by IL-2
AA832473	Similar to PF20 = contains WD repeats and localizes to the inter-microtubule bridges in Chlamydomonas flagella
H98765	Cytochrome P450, subfamily XXVII

Table 4 Comparison of the proposed method with other methods

Methods	DLBCL
MFDPSO (Agarwalla and Mukhopadhyay 2018)	90.01
DRF0-CGS (Bolón-Canedo et al. 2015)	94.67
IWSSr (Wang et al. 2015)	93.60
PSO dICA (Mollaee and Moattar 2016)	94.73
BDF (Medjahed et al. 2017)	89.44
SFS-MB (Wang et al. 2017)	80.90
TSVM (Maulik et al. 2013)	91.83
RMIFS (Tang and Zhou 2016)	95.57
BDE-X Rankf (Apolloni et al. 2016)	92.90
Proposed method	**97.62**

Bold value indicate the best value

Compliance with ethical standards

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

Ethical approval This study was performed using available datasets, as per my compliance with ethical standards there were no human or animal participants, and therefore, the study did not require ethics approval.

Research involving human and animal participants This article does not contain any studies with human participants or animals performed by any of the authors.

References

Agarwalla P, Mukhopadhyay S (2018) Bi-stage hierarchical selection of pathway genes for cancer progression using a swarm based computational approach. Appl Soft Comput 62:230–250

Alirezaei M, Taghi S, Niaki A, Armin S, Niaki A (2019) A bi-objective hybrid optimization algorithm to reduce noise and data dimension in diabetes diagnosis using support vector machines. Expert Syst Appl 127:47–57

Ang JC, Mirzal A, Haron H, Nuzly H, Hamed A (2016) Supervised, unsupervised, and semi-supervised feature selection: a review on gene selection. IEEE/ACM Trans Comput Biol Bioinf 13(5):971–989

Apolloni J, Leguizamón G, Alba E (2016) Two hybrid wrapper-filter feature selection algorithms applied to high-dimensional microarray experiments. Appl Soft Comput 38:922–932

Bolón-Canedo V, Sánchez-Maroño N, Alonso-Betanzos A (2015) Distributed feature selection: an application to microarray data classification. Appl Soft Comput 30:136–150

Daniel RP, Luis R (2019) Distributed ReliefF based feature selection in spark. Knowl Inf Syst 57(1):1–20

Dara RA, Makrehchi M, Kamel MS (2010) Filter-based data partitioning for training multiple classifier systems. IEEE Trans Knowl Data Eng 22(4):508–522

Edsgärd D, Johnsson P, Sandberg R (2018) Identification of spatial expression trends in single-cell gene expression data. Nat Methods 15(5):339–342

Fabris F, Freitas AA, Tullet JMA (2016) An extensive empirical comparison of probabilistic hierarchical classifiers in datasets of ageing-related genes. IEEE ACM Trans Comput Biol Bioinf 13(6):1045–1058

Ferreira AJ, Figueiredo MAT (2012) Efficient feature selection filters for high-dimensional data. Pattern Recognit Lett 33(13):1794–1804

Friedman N, Geiger D, Goldszmidt M (1997) Bayesian network classifiers. Mach Learn 29:131–163

Gonzalez-lopez J, Ventura S, Cano A (2019) Distributed multi-label feature selection using individual mutual information measures. Knowl based Syst 188:105052

Gutkin M, Shamir R, Dror G (2009) SlimPLS: a method for feature selection in gene expression-based DISEASE classification. PLoS One 4(7):6416

Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. J Mach Learn Res 3(3):1157–1182

Han J, Pei J, Kamber M (2006) Data mining: concepts and techniques. Morgan Kaufmann Elsevier, San Francisco

Hu L, Gao W, Zhao K, Zhang P, Wang F (2018) Feature selection considering two types of feature relevancy and feature interdependency. Expert Syst Appl 93:423–434

Kohavi R (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. Int Jt Conf Artif Intell 14(2):1137–1145

Liu J, Lin Y, Lin M (2017) Feature selection based on quality of information. Neurocomputing 255(10):11–22

Macgregor PF, Squire JA (2002) Application of microarrays to the analysis of gene expression in cancer. Clin Chem 48(8):1170–1177

Maulik U, Mukhopadhyay A, Chakraborty D (2013) Gene-expression-based cancer subtypes prediction through feature selection and transductive SVM. IEEE Trans Biomed Eng 60(4):1111–1117

Medjahed SA, Saadi TA, Benyettou A, Ouali M (2017) Kernel-based learning and feature selection analysis for cancer diagnosis. Appl Soft Comput 51(04):39–48

Mollaee M, Moattar MH (2016) A novel feature extraction approach based on ensemble feature selection and modified discriminant independent component analysis for microarray data classification. Biocybern Biomed Eng 36(3):1–9

Mukhopadhyay A, Maulik U (2013) An SVM-wrapped multiobjective evolutionary feature selection approach for identifying cancer-MicroRNA markers. IEEE Trans Nanobiosci 12(4):275–281

Nguyen BH, Xue B, Andreae P (2019) A new binary particle swarm optimization approach : momentum and dynamic balance between exploration and exploitation. IEEE Trans Cybern 1–15

Palma-Mendoza R-J, de-Marcos L, Rodriguez D (2018) Distributed correlation-based feature selection in spark. Inf Sci (NY) 496:287–299

Pang H, Goerge SL, Hui K, Tong T, George SL, Hui K, Tong T (2012) Gene selection using iterative feature elimination random forests for survival outcomes. IEEE ACM Trans Comput Biol Bioinf 9(5):997–1003

Peng H, Long F, Ding C (2005) Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell 27(8):1226–1238

Qu Y, Li R, Deng A, Shang C, Shen Q (2019). Non-unique decision differential entropy-based feature selection. Neurocomputing

Quinlan JR (1993) C4.5: programs for machine learning. Elsevier, New York

Ruiz R, Riquelme JC, Aguilar-ruiz JS (2006) Incremental wrapper-based gene selection from microarray data for cancer classification. Pattern Recognit Lett 39:2383–2392

Shukla AK (2020) Multi-population adaptive genetic algorithm for selection of microarray biomarkers. Neural Comput Appl 1–30

Shukla AK, Singh P, Vardhan M (2019a) A hybrid framework for optimal feature subset selection. J Intell Fuzzy Syst 36(3):2247–2259

Shukla AK, Singh P, Vardhan M (2019b) A new hybrid wrapper TLBO and SA with SVM approach for gene expression data. Inf Sci (NY) 503:238–254

Shukla AK, Singh P, Vardhan M (2019c) A new hybrid feature subset selection framework based on binary genetic algorithm and information theory. Int J Comput Intell Appl 18(03):1950020

Shukla AK, Singh P, Vardhan M (2020) An adaptive inertia weight teaching-learning-based optimization algorithm and its applications. Appl Math Model 77:309–326

Stevens KN, Cover TM, Hart PE (1967) Nearest neighbor pattern classification. IEEE Trans Inf Theory 13(1):21–27

Sun Y (2007) Iterative RELIEF for feature weighting: algorithms, theories, and applications. IEEE Trans Pattern Anal Mach Intell 29(6):1035–1051

Tang J, Zhou S (2016) A new approach for feature selection from microarray data based on mutual information. IEEE ACM Trans Comput Biol Bioinf 13(6):1004–1015

Venkataramana L, Gracia S, Rajavel J, Dodda R (2019) Improving classification accuracy of cancer types using parallel hybrid feature selection on microarray gene expression data. Genes Genom 41(11):1301–1313

Wang A, An N, Chen G, Li L, Alterovitz G (2015) Accelerating wrapper-based feature selection with K-nearest-neighbor. Knowl Based Syst 83:81–91

Wang A, An N, Yang J, Chen G, Li L, Alterovitz G (2017) Wrapper-based gene selection with Markov blanket. Comput Biol Med 81:11–23

Wang H, Tan L, Niu B (2019) Feature selection for classification of microarray gene expression cancers using bacterial colony optimization with multi-dimensional population. Swarm Evol Comput 48:172–181

Wu X, Kumar V, Ross QJ, Ghosh J, Yang Q, Motoda H, Steinberg D (2008) Top 10 algorithms in data mining. Knowl Inf Syst 14(1):1–37

Wu HC, Wei XG, Chan SC (2017) Novel consensus gene selection criteria for distributed gpu partial least squares-based gene microarray analysis in diffused large B cell lymphoma (DLBCL) and related findings. IEEE ACM Trans Comput Biol Bioinf 59:1–14

Yu L, Liu H (2004) Efficient feature selection via analysis of relevance and redundancy. J Mach Learn Res 5:1205–1224

Zhao L, Chen Z, Hu Y, Min G, Jiang Z (2016) Distributed feature selection for efficient economic big data analysis. IEEE Trans Big Data 13(9):1–10

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