Applied Soft Computing 111 (2021) 107729

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

An efficient feature selection framework based on information theory for high dimensional data

G. Manikandan [a,*], S. Abirami [b]

[a] Department of Computer Science and Engineering, College of Engineering Guindy, Anna University, Chennai 25, India
[b] Department of Information Science and Technology, College of Engineering Guindy, Anna University, Chennai 25, India

ARTICLE INFO

Article history:
Received 12 January 2021
Received in revised form 9 July 2021
Accepted 13 July 2021
Available online 23 July 2021

Keywords:
Feature selection
Feature fusion
Feature relevancy
Feature redundancy
Microarray
Bioinformatics
High dimensional data
Mutual information

ABSTRACT

Feature selection plays a vital role in many fields, particularly in pattern recognition and bioinformatics, for selecting informative and relevant features from high dimensional datasets. The increase in dimensionality of data along with the existence of redundant and irrelevant features leads to challenging performance issues when processing and analysing the data. In this paper, an effective feature selection technique called mutual information and Monte Carlo based feature selection (MIMCFS) is proposed. It comprises of two stages. The first stage aims to select predominant features from the high dimensional data. The second stage involves elimination of redundant features that were selected in the first stage. For the purpose of implementing the first stage, a new feature selection strategy based on the approximate Markov blanket and the concept of mutual information is proposed to find out irrelevant and redundant features. In second stage, to avoid misjudgement of redundant features as relevant features, a new strategy based on Monte Carlo tree search technique is proposed in order to completely eradicate redundant features and to improve feature interaction. For experimental evaluation, eight benchmark microarray datasets including imbalanced ones pertaining to cancer analysis are used. Further, in order to compare and justify the performance of the proposed feature selection method, seven state-of-art feature selection techniques namely CFS, Relief, DISR, JMI, CMIM and CMI are employed. The outputs from these feature selection techniques are provided to three standard classifiers namely Naive Bayes, SVM and C4.5 in order to assess the significance of the selected features in building classification models. 10-fold cross validation is adopted to evaluate the classifiers. Accuracy, precision, recall, f-measure, standard deviation, statistical significance metrics are measured to quantify the classifier performance. Experimental results demonstrate the outstanding performance of the proposed algorithm when compared to that of the standard existing methods.

1. Introduction

In the recent years, technology advancements in various application domains such as social media, bioinformatics, text analytics, image processing, clinical medicine, natural language processing and so on leads to the emergence of high dimensional datasets [1–5]. The primary concerns associated with the increase in dimensionality of the datasets are the computational cost and degradation in classification performance [6]. Generally, if dimensionality of the data increases, the number of non-informative features in the data grows as well. Further, it makes identification of common patterns from the data difficult and thereby, affects the performance of classifiers. Another major issue is that all the features of the dataset do not characterize equal predictive information. In general, features in a dataset are classified into strongly relevant, irrelevant, weakly relevant and non-redundant, and weakly relevant and redundant features [7]. The relevant features hold predictive information that aid in building learning models and concepts. Therefore, inclusion of these relevant features into the learning model improves the performance of classifiers. On the other hand, the features are considered as irrelevant if they do not convey any predictive information in building learning models and concepts. These irrelevant features tend to decrease the efficiency of classifiers. Further, redundant features are those that contain predictive and distinguishing information needed to build the learning models and concepts but, the information would have already been conveyed through any of the previously included features or subset of features. Thus, retaining these redundant features add to the computational complexity. Thus, inclusion of redundant and irrelevant features will affect the accuracy of the system as well as the learning time of the classification models as these features ground high dimensionality to the dataset leading to complexity in finding non-trivial

* Corresponding author.
E-mail address: manitamilm@gmail.com (G. Manikandan).

https://doi.org/10.1016/j.asoc.2021.107729
1568-4946/

patterns. Generally, these irrelevant and redundant features are identified and removed prior to building of classification model so that they do not affect the learning task and classification performance. Hence, to build an efficient classification model, dataset with less number of features conveying high predictive information is required. In this regard, feature selection plays a highly significant role [8,9].

Feature selection improves the quality of a dataset not only by identifying significant features but also by removing the unwanted features that degrade the quality of the dataset [7]. Feature selection process selects and finds the best meaningful inputs for further processing. In the context of microarray gene selection [10,11], an additional challenge is that microarrays reveal enormous information about the cell through huge number of genes leading to very high dimensionality of data; however, with comparatively small number of samples. As the genes correspond to the features in a microarray dataset, it may include noisy or irrelevant or missing data making prediction more complicated. Further, it potentially incurs high computational overhead in terms of prediction and classification time and cost, involves high computational complexity and leads to large prediction errors as well. Therefore, the primary concern is to identify the genes that predict and determine the type of tissues for classification and to eliminate the remaining ones. Identification of highly important and informative genes/features is considered as the main challenge in the existing methods towards gene data analysis [12,13]. Feature selection techniques are normally used for simplifying classification process, reducing training times and enhancing generalization.

Numerous literature methods have been reported [14] to select highly important, relevant and informative features. It includes t-test, KNN, Shannon's entropy, clustering, logistic regression, and information gain. The limitation of these methods is that it requires different framework to achieve feature selection with respect to feature relevance and redundancy. To alleviate these limitations, feature grouping [15] has been used for dimensionality reduction. This method efficiently reduces the feature dimensionality by finding optimal subset of features. It improves the stability and reduces the complexity of the model as well. This approach provides important features in the form of groups wherein each feature is highly correlated with the other, in contrast to other techniques that provide a single subset of features [16].

Based on the choice of evaluation metric and the combination of selection algorithm and building of learning model, feature selection techniques are categorized into five classes: filter method, wrapper method, embedded method, hybrid method and ensemble method [17–20]. Filter method selects the features based on its internal characteristics before classification task. It filters the significant features on the basis of the incorporated evaluation criteria. As these methods do not employ any learning model, it is computationally less intensive and fast to compute. However, these approaches tend to select redundant variables as they do not consider the relationship between the features. Further, these methods generally yield lower performance when compared to that of the wrapper methods [21,22]. Filter methods measure the mutual information, inter or intra class distance, or significance tests scores of the features or class for selecting the features. Some of the standard filter methods include max-relevancy and min-redundancy (mRMR), fast correlation-based filter method (FCBF), feature selection technique-based interaction capping, and Relief-F relevant feature selection. Further, filter methods [23,24] can be divided into two types: univariate and multivariate. The univariate method assesses the feature importance individually using the evaluation criteria, whereas the multivariate method evaluates the features based on the dependencies among the features [25]. Some popular methods of univariate filter methods [26] include information gain, Fisher score, Gini index, Laplacian score, symmetrical uncertainty, variance [27], SVD-Entropy [28], spectrum decomposition (SPEC) [29], and unsupervised spectral feature selection method for mixed data (USFSM) [30]. A few of the multivariate methods include max-relevance and min-redundancy (mRMR) [21], relevance–redundancy feature selection (RRFS), multi-cluster feature selection [30], unsupervised feature selection on ant colony optimization (ACO) [26] [31] and random subspace method (RSM).

Wrapper method incorporates learning algorithms to score and select the feature subsets. In this regard, each and every subset is given as input to train the learning model. Then, the model is tested based on the hold-out strategy. After that, based on the error rate of built models, the relative scores of each subset are identified. Based on the scores, significant subset of features is selected. As the wrapper methods train each subset, it yields better subset of features at the expense of high computational cost [32] [33]. The two main approaches adopted to select features in this regard include sequential and random search strategies. Sequential search technique selects the features sequentially and is prone to get struck in local optima. Alternatively, random search technique applies random strategies to select the features and thereby, escapes the local optima. Embedded methods [34] employ group of machine learning techniques for model construction and thereby, select the significant features. This method avoids repetitive running of the classifier and analysis of each feature subset. In the context of embedded feature selection methods, decision tree and support vector machine (SVM) algorithms are widely employed for building learning models.

Another well-known category of feature selection technique is the hybrid method. This method forms the central focus of the recent research in the domain of the feature selection. Hybrid methods carry out the feature selection process by combining various methods (like filter and wrapper) to select the features with the view to achieve better performance. Hybrid methods handle the over-fitting issue effectively. Further, it incorporates different search strategies in different evaluation criteria with the intention of achieving higher performance. Some of the existing feature selection works with respect to microarray data can be found in [35–38]. Ensemble feature selection methods [39–41] employ machine learning models to select feature subset groups. Ensemble model involves aggregation of outputs of all the machine learning algorithms considered by the model. These ensemble methods provide good approximation in selecting the best subset or ranking of features through aggregating the features from outputs of several feature selection algorithms [42,43]. Apart from this, feature selection techniques can be further categorized into supervised and unsupervised categories [14]. In supervised feature selection, the feature selection process is carried out with the help of labelled data, whereas in unsupervised feature selection [44], the process is carried out without the assistance of labelled data.

The main aim of feature selection algorithm is to determine the predominant features that can yield the highest classification accuracy. However, the number of features that are chosen to be significant is a trade-off as very less number of features may not hold sufficient information to build an efficient model while large number of features may encompass irrelevant and redundant features that may degrade classification performance. Hence, an optimal number of features that hold sufficient predictive information must be selected. For achieving this goal, a good feature selection model, which finds and assesses all the feature subsets in order to identify the important features

by removing redundant and irrelevant features, is essential. Redundancy among the features is identified based on the correlation among the features, whereas the relevance is identified by finding the relationship between the feature and the class. Correlation is a statistical measure for finding the relationship between the features. If the systematic changes in one feature affects changes in the other feature, then the two features are said to be strongly correlated [45]. In order to measure the correlation among the features or between the feature and the class, computational methods based on information theory such as mutual information (MI) [25] [46], symmetrical uncertainty (SU) [47], joint mutual information (JMI) [48], conditional mutual information (CMI) [49,50], conditional mutual information maximization (CMIM) [34,35], and double input symmetrical relevance (DISR) [51] can be adopted as they can predict both linear and non-linear relationship among the features.

In order to determine the redundant features approximately, approximate Markov model based on the symmetrical uncertainty has been used recently [52]. Among distinct set of selection criteria, Mutual Information (MI) based approaches are widely used when compared to the techniques based on symmetrical uncertainty as it measures the dependencies between the features as well as assesses the informative quantum of features in dataset [19]. However, there is a chance that some of the strongly relevant features be mistreated as redundant ones. To avoid this scenario and to select the strong predominant features from high dimensional data, a new strategy called approximate Markov blanket based on mutual information is proposed in this work.

Thus, this work considers MI based feature selection applied in the approximate Markov blanket to assess the importance of the features. Accordingly, if feature f_1 forms the approximate Markov blanket with feature f_2 and if the mutual information (MI) between f_1 and target class is not less than the MI between f_2 and target class, then f_1 contains more information than f_2. Further, if the MI between f_1 and f_2 is not less than the MI between f_2 and the target class, based on the threshold of MI, it suggests whether the correlation between these features is strong or not. From these two conditions, the strongly relevant features can be identified. Using this technique, highly relevant features are identified and selected by ensuring the strong correlation between the features. Further, it avoids misjudgement of predominant features as redundant ones.

The main contributions of this work are as follows:

(1) A new strategy is introduced to group and select the strongly relevant features based on the Markov blanket. It ensures strong correlation between the features in order to identify the prominent features.

(2) A novel feature selection method is proposed to select the prominent features in high-dimensional dataset. It selects the features based on mutual information between the features. Feature reduction is obtained by applying the approximate Markov blanket that involves pairwise comparison of the mutual information. Through this new strategy, the redundant features and irrelevant features are identified and removed from the dataset.

(3) In order to avoid misjudgement of redundant feature as relevant feature, a new method based on Monte Carlo tree search technique is proposed. It ensures the identification of strongly relevant features and ignores the redundant features by finding the interdependencies of the features.

(4) The proposed method discovers the redundant features with and without consideration of the target class. In this approach, the redundant features are completely removed. It provides relevant features without redundant features. Hence, the proposed method considers both relevance and redundancy among the features while selecting the features.

(5) The proposed method yields better average number of features with good classification accuracy when compared to that of the features yielded through standard state-of-art feature selection techniques and competitive feature selection methods.

In this work, the proposed method is evaluated with eight microarray biomedical benchmark datasets. Optimal features are selected without the irrelevant and redundant features. When these features are presented to a classification model, it results in good classification accuracy. Furthermore, the proposed method is compared with other standard techniques to justify its better performance in terms of selection of strongly relevant features and improved classification accuracy. The reminder of the paper is organized as follows: Section 2 briefs the various existing methods related to feature selection on high dimensional medical datasets. Section 3 provides the background on which the proposed work is rooted. Section 4 describes the concept of mutual information and the proposed feature selection method. Section 5 presents the experimental validations and discusses the outcomes of all the experimental algorithms. Finally, Section 6 concludes the work.

2. Related work

In this section, a brief literature review pertaining to various feature selection methods both that are based on information theory and that do not rely on information theory (non-information theory based) is concisely presented. In general, feature selection techniques aim to select a subset of important and informative features from the original feature set, such that the irrelevant and redundant features are discarded, with the view to achieve good performance. Many feature selection techniques that are grounded on information theory have been proposed [53] to select predominant features from high dimensional data. Further, it is crucial to evaluate and measure the relevance between the features and the class as well as the redundancy among the features. Therefore, it is vital to find and define the evaluation criteria adopted to select the important features and eliminate irrelevant and redundant features. Various evaluation criteria are used for selecting predominant features from high dimensional datasets. Information theory based evaluation techniques have been widely used in filter-based feature selection approaches. However, Correlation based Feature Selection and Relief feature selection (commonly used feature reduction techniques) do not rely on information theory based techniques. Correlation based Feature Selection (CFS) [54] is one of the filter-based feature selection methods rooted on the test theory. It evaluates the adequacy of the feature subset by assessing the correlation among the features and between feature and class. Relief algorithm [55], another filter based approach, measures the relevance of the features and the class by sampling the instances from the training dataset randomly. Further, for each feature, it updates the relevance score based on the difference among the instances selected and the two nearest instances of different and same class. This algorithm scales well for high dimensional data whereas it does not remove redundant features.

Further, the focus is placed on information theory based methods for selecting predominant features from high dimensional data. In this literature review, f_k denotes the candidate feature, f_j indicates the selected feature, S represents the selected subset and Y signifies the class. In information theory based methods, Mutual information maximization [56], also known as information gain (IG), is one of the straight forward feature selection strategies that evaluates the correlation between the features and class by applying the mutual information method. However, this IG technique does not consider the redundant information among the candidate features and the selected features subset.

To deal with this issue, maximal relevancy minimal redundancy (mRMR) that takes redundancy among features into account is proposed [57]. The evaluation function for the mRMR technique is given in Eq. (1).

$$J_{mRMR}(f_k) = I(f_k; Y) - \frac{1}{|S|}\sum_{f_j \epsilon S} I(f_j; f_k) \quad (1)$$

In Eq. (1), $J_{mRMR}(f_k)$ denotes the evaluation function, S represents the subset that has already been selected, $I(f_k; Y)$ provides the relevance measure between the candidate feature and class Y, $I(f_j; f_k)$ measures the redundancy factor between the candidate feature f_k and the selected feature f_j, and $\frac{1}{|S|}$ represents the normalization factor between the feature redundancy and the feature relevancy term. Alternatively, joint mutual information (JMI) that measures the importance of features between the candidate features and the already selected features subset is proposed [58]. It is expressed as in Eq. (2).

$$J_{JMI}(f_k) = \sum_{f_j \epsilon S} I(f_k, f_j; Y) \quad (2)$$

The Equation pertaining to JMI can be re-written as in Eq. (3).

$$J_{JMI}(f_k) = I(f_k; Y) - \frac{1}{|S|}\sum_{f_j \epsilon S} I(f_j; f_k) + \frac{1}{|S|}\sum_{f_j \epsilon S} I(f_j; f_k|Y) \quad (3)$$

In Eq. (3), $I(f_j; f_k|Y)$ quantifies the aggregation of selected features f_j and candidate features f_k, given the classification information Y.

Further, another evaluation metric namely conditional mutual information (CMI) [59] selects the features that hold the maximum mutual information with the class Y. The evaluation criterion for CMI is as provided in Eq. (4).

$$J_{CMI}(f_k) = I(f_k; Y|S) \quad (4)$$

Using the identity $I(A: B|Y) - I(A; B) = I(A; Y|B) - I(A: Y)$, Eq. (4) can be re-expressed as in Eq. (5),

$$J_{CMI}(f_k) = I(f_k; Y|S) = I(f_k; Y) - I(f_k; S) + I(f_k; Y|S) \quad (5)$$

Conditional mutual information maximization (CMIM) [60] is one of the competitive feature selection methods that select the features by maximizing the conditional mutual information among the candidate features f_k and the class Y, given the already selected features. The mathematical representation of CMIM is given in Eq. (6).

$$J_{CMIM}(f_k) = \min\left\{I(f_j; Y|f_k)\right\} \quad (6)$$

In Eq. (6), $I(f_j; Y|f_k)$ quantifies the aggregation of selected features f_j and candidate features f_k, given the classification information Y.

If $I(f_j; Y|f_k)$ is low, then it signifies that the candidate feature f_k is redundant with that of the selected feature or it suggests that the candidate feature f_k is irrelevant with respect to class Y. Meanwhile, on comparing $I(f_k; Y)$ and $I(f_j; Y|f_k)$, it does not concern the redundant information among the pairwise features in the subset. Thus, it selects features that are weakly redundant and irrelevant. Furthermore, double input symmetrical relevance (DISR) [61] is a feature section method that normalizes joint mutual information similar to the JMI technique. DISR function is expressed as in Eq. (7).

$$J_{DISR}(f_k) = \sum_{X_j \epsilon s} \frac{I(Y; f_k, f_j)}{H(Y, f_k, f_j)} \quad (7)$$

DISR technique obtains information pertaining to Y, as provided by f_k and f_j. In addition, this information $I(Y; f_k, f_j)$ is divided by the joint entropy $I(Y; f_k, f_j)$.

Recently, a feature selection algorithm has been introduced for selecting the features based on uncertainty change ratio (UCRFS) [62] on the high dimensional datasets. In this technique, the difference between the reduced uncertainty and remained uncertainty of the class from the already selected features and candidate features are considered in a different manner in order to select the features with less redundancy and relevance. The method uses uncertainty change ratio with traditional feature relevance and redundant terms to evaluate candidate features. Furthermore, another feature selection technique called dynamic feature importance (DFIFS) [63], which assesses the feature importance and feature redundancy, has been reported. In this method, Gini importance (GI) method that is employed in random forest is used in order to compute feature importance. In order to assess the feature redundancy, maximum information coefficient (MIC) is used. Moreover, this technique also incorporates a few existing filters with the view to achieve high accuracy with less number of features in originally high dimensional datasets.

In the context of microarray gene selection, in order to improve the gene selection and the performance effectively with low complexity, a multivariate feature ranking method has been put forward [64]. In this technique, Markov blanket (MB), which simultaneously considers both redundancy among the features and relevance with the class labels, is used for selecting significant features from high dimensional datasets. This method mainly focuses on the prediction accuracy rather than the interpretability of important features.

In this work, information theory based methods such as CMI, CMIM, JMI, DISR, mRMR and methods that are not based on information theory such as CFS and Relief are considered for comparing and justifying the performance of the proposed feature selection method. The following section provides an account on the background needed for the proposed work.

3. Background

This section provides information pertaining to the datasets used in the experimental analysis and accounts on the background concepts involved in the proposed work. For the purpose of experimental analysis, benchmark datasets with respect to gene data have been selected. The detailed analysis of the datasets is reported in Section 5. Owing to the complex nature of the gene microarray data sources, proper analysis and enhancement is essential for understanding the patterns in the datasets. This work investigates eight microarray datasets namely Leukemia, Leukemia_3c, Lung, SRBCT, Lymphoma, Leukemia _4c, Ovarian and Breast with respect to identification of cancer. The proposed work aims at determining the predominant genes that effectively interpret the problem in hand (identification of cancer) for further analysis. The proposed feature selection technique is rooted on the concepts of relevance, redundancy and interaction analysis among the features. In this regard, preliminary concepts related to information theory, with emphasis on mutual information and entropy, are presented here. Further, the rationale for incorporating these concepts in the proposed work is explained.

In information theory [65–68], mutual information (MI) is used to analyse the relationship between any two variables. This can be applied to quantitatively analyse interdependency between two features or between a feature and the target variable. Let $X = \{x_1, x_2, x_3, \ldots..., x_m\}$ be a random variable, then the entropy $H(X)$ can be defined as in Eq. (8).

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i) \quad (8)$$

In Eq. (8), $p(x_i)$ is the probability function that denotes the distribution of x_i. When X is discrete, the value of $p(x_i)$ is denoted as $p(x_i) = \frac{number\ of\ instances\ with\ value x_i}{total\ number\ of\ instances(n)}$.
Suppose X = $\{x_1, x_2, x_3, \ldots.., x_m\}$ and Y = $\{y_1, y_2, y_3, \ldots.., y_n\}$ be the two discrete random variables, then the joint probability of the variables X and Y is $p(x_i, y_j)$, where $i = 1, 2, ..m; j = 1, 2, ..n$. Then, the joint entropy $H(X, Y)$ of the variables X and Y can be defined as in Eq. #(9).

$$H(X, Y) = -\sum_{i=1}^{m}\sum_{j=1}^{n} p(x_i, y_j) \log p(x_i, y_j) \tag{9}$$

In case, if two variables X and Y are statistically independent, then the joint entropy $H(X, Y)$ is represented by Eq. (10).

$$H(X, Y) = H(X) + H(Y) \tag{10}$$

In Eq. (10), the joint entropy can be expressed in terms of conditional entropy $H(X|Y)$ and $H(Y|X)$ as in Eqs. (11) and (12).

$$H(X, Y) = H(X|Y) + H(Y) \tag{11}$$

$$H(X, Y) = H(Y|X) + H(X) \tag{12}$$

Based on the Eqs. (11) and (12), the mutual information can be represented in the form of entropy as provided in Eq. (13).

$$MI(X; Y) = H(Y) - H(X|Y)\,; MI(X; Y) = H(X) - H(Y|X) \tag{13}$$

Assuming that the entropy of X is known and entropy of Y is observed after observing the entropy of X, the conditional entropy $H(Y|X)$ is defined as in Eq. (14).

$$H(Y|X) = H(X, Y) - H(X) = -\sum_{i=1}^{m}\sum_{j=1}^{n} p(x_i, y_j) \log p(y_j|x_i) \tag{14}$$

Proof.

$$\begin{aligned} H(Y|X) &= -\sum_{i=1}^{m}\sum_{j=1}^{n} p(x_i, y_j) \log p(x_i, y_j) \\ &= -\sum_{i=1}^{m}\sum_{j=1}^{n} p(x_i, y_j) \log[p(x_i)\, p(y_j|x_i)] \end{aligned}$$

$$\begin{aligned} MI(X, Y) &= H(Y|X) = H(X, Y) - H(X) \text{ from Eq. (13)} \\ &= H(X) - H(X|Y) \\ &= -\sum_{i=1}^{m}\sum_{j=1}^{n} p(x_i)\, p(y_j|x_i) \log p(x_i) \\ &\quad + \sum_{i=1}^{m}\sum_{j=1}^{n} p(x_i, y_j) \log p(y_j|x_i) \\ &= -\sum_{i=1}^{n} p(x_i) \log p(x_i) \left[\sum_{i=1}^{n} p(y_i|x_i)\right] + H(Y|X) \\ &= H(Y|X) + H(X) \end{aligned}$$

In this way, the mutual information measures the interactive information and the statistical dependency between X and Y that share information with each other. Similarly, mutual information of the two joint random variables X *and* Y with conditional entropy can be calculated through Eqs. (15) and (16).

$$MI(X; Y) = -\sum_{i=1}^{m}\sum_{j=1}^{n} p(x_i, y_j) \log \frac{p(x_i|y_j)}{p(x_i)} \tag{15}$$

$$MI(X; Y) = H(X) - H(X|Y)\,;\ MI(Y; X) = H(Y) - H(Y|X)\,;$$

$$MI(X; Y) = MI(Y; X) = H(X) + H(Y) - H(X|Y) \tag{16}$$

Proof.

$$MI(X; Y) = -\sum_{i=1}^{m}\sum_{j=1}^{n} p(x_i, y_j) \log \frac{p(x_i|y_j)}{p(x_i)}$$

$$MI(X; Y) = \sum_{i=1}^{m}\sum_{j=1}^{n} p(x_i, y_j) \log \frac{1}{p(x_i)} - \sum_{i=1}^{m}\sum_{j=1}^{n} p(x_i, y_j) \log \frac{1}{p(x_i, y_j)}$$

$$MI(Y; X) = \sum_{i=1}^{m}\sum_{j=1}^{n} p(x_i, y_j) \log \frac{1}{p(y_i)} - \sum_{i=1}^{m}\sum_{j=1}^{n} p(x_i, y_j) \log \frac{1}{p(x_i, y_j)}$$

$$MI(X; Y) = MI(Y; X) = H(X) + H(Y) - H(X|Y)$$

$$MI(X; Y) = \sum_{i=1}^{m}\sum_{j=1}^{n} p(x_i, y_j) \log \frac{1}{p(x_i)} + \sum_{i=1}^{m}\sum_{j=1}^{n} p(x_i, y_j) \log \frac{1}{p(y_i)} - \sum_{i=1}^{m}\sum_{j=1}^{n} p(x_i, y_j) \log \frac{1}{p(x_i, y_j)}$$

Based on Eqs. (15) & (16), interaction information and MI can be mathematically defined as in Eq. (17).

$$MI(f_j; C|S) = \sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{s} p(f_j, S_k, C_i) \log \frac{p(C_i|f_j, S_k)}{p(C_i|S_k)} \tag{17}$$

$p(C_i|f_j, S_k) \neq p(C_i|S_k)$ *i.e.*, $\frac{p(C_i|f_j,S_k)}{p(C_i|S_k)} \neq 1$, hence $MI(f_j; C_i|S_k) \neq 0$

According to Eq. (17), feature f_j is said to be strongly relevant to the target class C if and only if $p_f(C_i|f_j, S_k) \neq p_f(C_i|S_k)$. This clearly indicates the importance of features with respect to the feature and class. Hence, it is evident that the feature selection technique based on the mutual information is supremely effective for selecting highly informative features.

3.1. Relative definitions

Let the feature selection (FS) problem involve n features and m instances. Let definition of variables be as follows: $F = f_1, f_2, f_3, \ldots\ldots f_n$ represents the feature set, C denotes the entire set of class labels, p_f defines the probability function, f_j, where, $j \in \{1, \ldots\ldots, n\}$ indicates the j^{th} feature, $CS_j = F - \{f_j\}$ denotes the complementary set of feature f_j and the class label C [52] [67].

Definition 1 (*Strong Relevance [69]*)**.** Given the target class C, a feature f_j is strongly relevant to the target class C if and only if $p_f(C|f_j, CS_j) \neq p_f(C|CS_j)$.

In other terms, a feature is strongly relevant if the feature holds high discriminative power and prediction accuracy, i.e., the feature contains strong information about the class C. Strong relevance of the feature is consistently required for predicting the optimal subset.

Definition 2 (*Weak Relevance [69]*)**.** If the target class C is given, a feature f_j is weakly relevant to the target class C if and only if $p_f(C|f_j, CS_j) = p_f(C|CS_j)$ and $\exists CS'_j \subset CS_j$, such that $p_f(C|f_j, CS'_j) \neq p_f(C|CS'_j)$.

In other words, weakly relevant features also contribute in improving the prediction accuracy under certain conditions — the feature should be non-redundant and well suited with the evaluation criteria. On the other hand, the weakly relevant features with redundant information can be ignored since they do not provide any significance in improving the prediction accuracy.

Definition 3 (*Irrelevance [69]*)**.** Given the target class C, a feature f_i is irrelevant to the target class C if and only if $p_f\left(C|f_j, CS'_j\right) = p_f\left(C|CS'_j\right)$ and $\forall CS'_j \subset CS_j$.

Definition 4 (*Markov Blanket [69]*)**.** Redundancy of the features can be determined by using Markov blankets theory. In general, it is defined as follows.

Given target class C and feature f_j, let $\mathbb{M}_j \subset F\left(F_j \notin \mathbb{M}_j\right)$. $\mathbb{M}_j$ is considered to be the Markov blanket for F_j iff $p_f\left(F - \mathbb{M}_j - \left\{f_j\right\}, C|F_j, \mathbb{M}_i\right) = p_f(F - \mathbb{M}_j - \left\{f_j\right\}, C|\mathbb{M}_j)$.

The Markov blanket condition states that $\mathbb{M}_j$, not only absorbs the information that the feature f_j has about the target class C, but also absorbs all the other features in the subset. The optimal subset from the entire set is acquired by backward elimination process, also known as Markov blanket filtering. Let $\mathbb{S}$ be the set of features present in the set, where $\mathbb{S} = F$ in the early stage at any phase. If there exists a Markov blanket $\exists\mathbb{M}_j$ for F_j within the present set $\mathbb{S}$, then F_j is removed from the set $\mathbb{S}$. From the above expression, it is evident that the strong predominant features do not form a part of any of the Markov blankets. This condition requires that the Markov blanket $\mathbb{M}_j$ should not only contain information about f_j with respect to target class C but also encompass the feature set $F - \mathbb{M}_j - \{f_j\}$. In this way, the redundant features are identified by using Markov blanket and removed from the entire feature set *FS* without losing any of the strongly relevant features.

Definition 5 (*Approximate Markov Blanket [52]*)**.** Given two relevant features f_j and f_k (j ≠ k), f_k forms an approximate Markov blanket for f_j if and only if $SU_{k,C} \geq SU_{j,C}$ *and* $SU_{j,k} \geq SU_{j,C}$. As mentioned earlier, Markov blanket-based feature filtering involves backward elimination. The Markov blanket set guarantees that the redundant features are removed without losing the relevant features in an earlier phase. It is also possible to remove the redundant features in any later phase. For example, if the feature f_j is the only feature that forms an approximate Markov blanket on the feature f_i, and f_k forms an approximate Markov blanket for the feature f_j, then if f_j is eliminated based on f_k, there will be no approximate Markov blanket for f_i in the present set.

Definition 6 (*C-Correlation [69]*)**.** Given feature f_j with the target class C, the correlation between the feature f_j and the target class C is denoted as $SU_{j,C}$. Symmetrical incertainty (SU) is the correlation measure used to develop an approximation method for defining and analysing the relevance and redundancy between the features.

Definition 7 (*F-Correlation [69]*)**.** Given features f_j and f_k, the correlation between the features f_j and f_k, where $i \neq j$, is denoted as $SU_{j,k}$..

Having provided the basic definitions involved in the proposed work, the following section details the proposed work.

4. Proposed method

Approximate Markov blanket is an effective approach for finding the redundant features from the original feature set. In general, Symmetrical uncertainty (SU) (discussed in Section 3: Definition 5) is a non-linear correlation measure used to find the correlation between features [42]. In the proposed work, approximate Markov blanket procedure is employed in order to remove the redundant features without removing the relevant features. However, there is a possibility that strongly relevant features may be identified as redundant features. In order to identify relevant features, mutual information based approximate Markov blanket is introduced. The proposed mutual information based approximate Markov blanket aims at determining the prominent features without misjudgement of relevant features as redundant features (presented in Definition 8). In this definition, the approximate Markov blanket measures the feature information based on mutual information. However, it is different from the symmetrical uncertainty measure in that SU measure is non-parametric and considers only non-linear dependencies between the features. Nevertheless, it is important to consider both linear and non-linear dependencies between the features. In this context, MI incorporated correlation-based feature selection technique is proposed. Using mutual information as correlation measure integrated with Monte Carlo Tree Search (MCTS) model, an approximation method for analysing the feature relevance and redundancy is proposed. The proposed framework is shown in Fig. 1.

Definition 8 (*Modified Approximate Blanket*)**.** Given two relevant features f_j *and* f_k ($j \neq k$), f_k forms an approximate blanket for f_j iff $MI_{k,C} \geq MI_{j,C}$ *and* $MI_{j,k} \geq MI_{j,C}$.

$$MI\left(f_j, f_k\right) = \sum_{f_j \epsilon \mathbb{S}} \sum_{\substack{f_k \epsilon \mathbb{S} \\ f_j \neq f_k}} \frac{MI\left(f_j, f_k\right)}{m\,(m-1)} \tag{18}$$

$$MI(f_j, C) \geq MI(f_k, C)$$

$$MI(f_j, f_k) \geq MI(f_k, C)$$

Definition 9 (*C-Correlation Based on MI [69]*)**.** Given feature f_j with the class C, the correlation between the feature f_j and the target class C is denoted as $MI_{j,C}$. The C-correlation for each feature f_j and class C is computed and f_j is selected if it is strongly correlated with the class C.

Definition 10 (*F-Correlation Based on MI [69]*)**.** Given feature f_j *and* f_k, the correlation between the feature f_j and f_k where $j \neq k$, is denoted as $MI_{j,k}$.

Definition 11 (*Predominant Feature [52]*)**.** Given a feature set, a subset $S \subset F$ from the entire set F is to be identified such that S encompasses predominant features. Feature f_j is considered as a predominant feature and added to subset S if it does not form any approximate Markov blanket in the current set. These predominant features are highly important in prediction and at any stage they will not be ignored from the set. If a particular feature f_j is eliminated based on its strong relationship with some other predominant feature f_k in earlier stage, it is possible that the removed feature f_j will find an approximate Markov blanket in any of the later phases for the same feature f_k.

The process involved in the proposed is presented in Algorithm 1. In Algorithm 1, the feature groups are generated in order to find out the redundant features. The steps involved in generating feature groups are as follows: Initially, all the features f_j are listed in the descending order based on the $MI(f_j, C)$ value and the search sequence is determined. Subsequently, the first feature is selected in the list and MI between the features f_j *and* f_k and MI between f_j and the class are computed. Further, this step is repeated sequentially for all the other features in the list. Then, all the features with MI measure less than or equal to the threshold τ are removed as the correlation value will be zero or less than

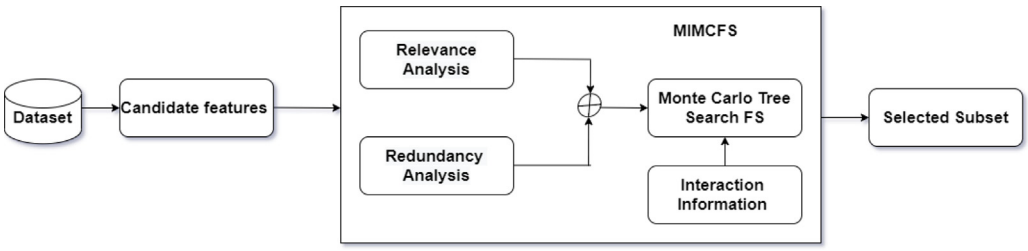


Fig. 1. Proposed feature selection framework.

zero if the features are not correlated with each other. Through this step, not only the most prominent feature is identified but also the prominent feature groups are detected. After that, the first feature in the list is considered and it is checked if it forms a Markov blanket for any other feature in the list. If it forms a Markov blanket, then that feature is considered as redundant and is added to the prominent group. Then, the process is repeated for all the features in the list. Once all the features are investigated, a subsequent prominent feature will be selected based on the next relevant feature with the highest MI value in the list. Also, the feature should not belong to any of the prominent groups. This iterative procedure repeats until no prominent feature is found in the list.

Algorithm 1:

$begin$

1: $\mathcal{P}_{set} \leftarrow \phi, Q_G \leftarrow 0, \lambda \leftarrow 0, m \leftarrow 0$

2: $\tau = \sum_{f_j \epsilon\, \mathbb{S}} \sum_{\substack{f_k=1 \\ f_j \neq f_k}} 2MI(f_j, f_k) \,/\, m\,(m-1)$

3: $m_j \leftarrow MI(f_j, C), m_k \leftarrow MI(f_k, C)$

4: $foreach f_j \epsilon\, S$

5: $if (MI(f_j, C) > \tau$

6: $\mathcal{P}_{set} \leftarrow f_j$

7: $endif$

8: $endfor$

9: $\mathcal{P}_{MI} \leftarrow sort(\mathcal{P}_{set})\ where\ m_j > m_k, \forall j < k$

10: $foreach\ f_j \epsilon\, \mathcal{P}_{MI}: f_j \notin \mathrm{Q_G}$

11: $\mathcal{P}_{set} \leftarrow f_j$

12: $foreach\ f_k \epsilon\, \mathcal{P}_{MI}: f_k \notin \mathrm{Q_G}$

13: $if (MI(f_j, f_k) \geq MI(f_k, C))$

14: $\mathcal{P}_{set} \leftarrow f_j$

15: $endif$

16: $endfor$

17: $Q_G \leftarrow \mathcal{P}_{set} \cup Q_G$

18: $\lambda \leftarrow \lambda + 1$

19: $endfor$

20:end

After identifying the redundant features, the proposed work involves determining the irrelevant features. Algorithm 2 presents the procedure for removing the irrelevant features from the subset. A strategy known as variable neighbourhood search is incorporated to evaluate the quality of the features. It is widely used for solving optimization problems which adopted from [70]. This approach is rooted on the concepts of local minima and global minima. It starts generating a solution with the initial value $\mathbb{S}$, then it generates a new solution $\mathbb{S}'$ with the first neighbourhood value of S. This process is continued for acquiring the improved solution $\mathbb{S}''$. If $\mathbb{S}''$ does not improve the value S, the search continues until the stopping criterion, i.e. the maximum value of k, is reached. This function evaluates the features subset based on the correlation measure between features and class. Mutual information technique is implemented for finding the quality of features. Based on the correlation measure, strongly relevant features will be examined and irrelevant features will be ignored.

The optimized feature subset will be strongly correlated with the class whereas uncorrelated with each other. Based on this strategy adopted in correlation-based feature selection (CFS) [52] [54] [70], the proposed evaluation function is formulated based on mutual information as in (19) with the view to identify good subset of features in $\mathbb{S}$.

$$J(\mathbb{S}) = \frac{m.\overline{MI}(\mathbb{S}, C)}{\sqrt{m + m(m-1).\overline{MI}(\mathbb{S}, \mathbb{S})}} \tag{19}$$

where, $\overline{MI}(\mathbb{S}, C) = \sum_{f_j \epsilon \mathbb{S}} MI(f_j, C)/m$ and $\overline{MI}(\mathbb{S}, \mathbb{S}) = \sum_{f_j \epsilon \mathbb{S}} \sum_{\substack{f_k \epsilon \mathbb{S} \\ f_j \neq f_k}} 2MI(f_j, f_k)/m(m-1)$

In Eq. (19), $J(\mathbb{S})$ is the evaluation function for feature subset $\mathbb{S}$ consisting of m features, $\overline{MI}(\mathbb{S}, C)$ is average correlation measure between features and the target class and $\overline{MI}(\mathbb{S}, \mathbb{S})$ is the average correlation between features. There are different attribute quality evaluation methods. In the proposed work, mutual information method is employed to measure the quality of features in CFS.

Algorithm 2:

1: $begin$

2: $k \leftarrow 1$

3: $repeat$

4: $if\ J(\mathbb{S}'') \geq J(\mathbb{S})$

5: $\mathbb{S} \leftarrow \mathbb{S}''$

6: $k \leftarrow 1$

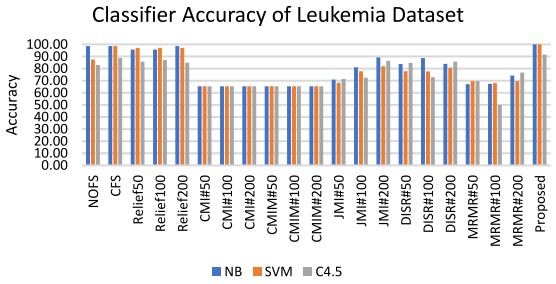
7: $else$

8: $k \leftarrow k + 1$

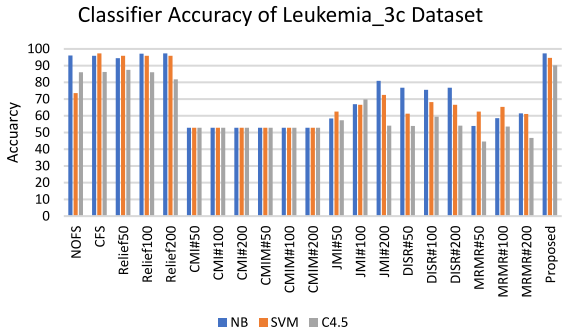
9: $untill\ (maximum\ of\ k)$

10: $return\ \mathbb{S}$

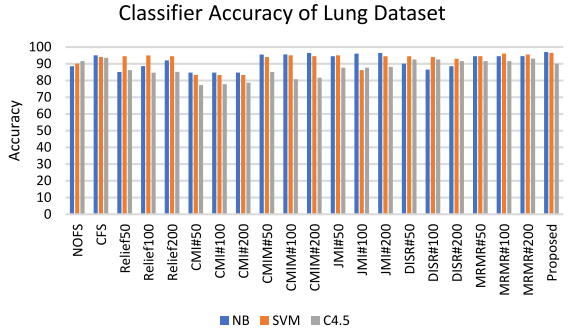
10: end



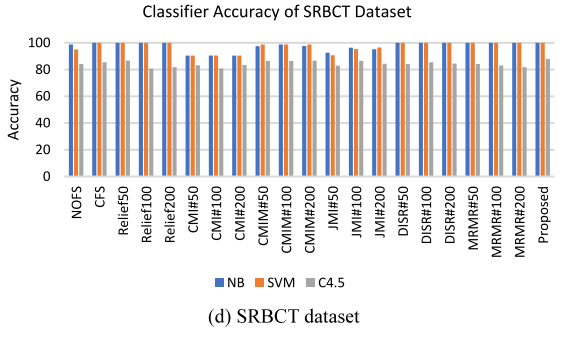
(a) Leukemia dataset



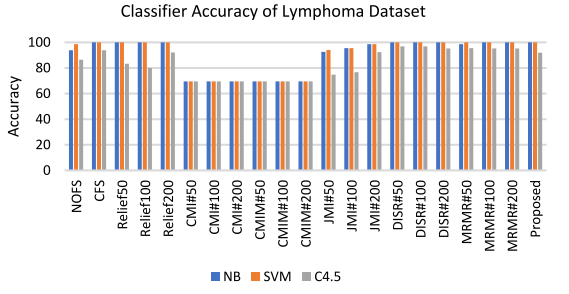
(b) Leukemia 3_c dataset



(c) Lung dataset

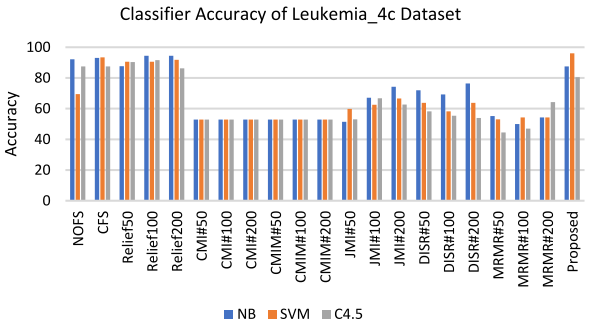


(d) SRBCT dataset

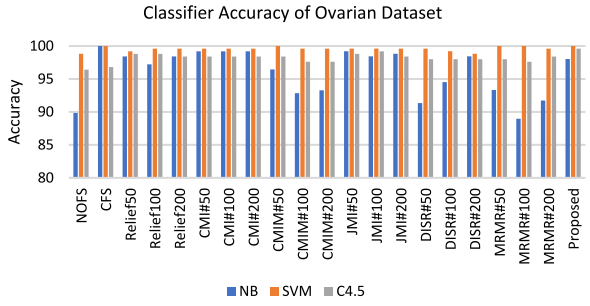


(e) Lymphoma dataset

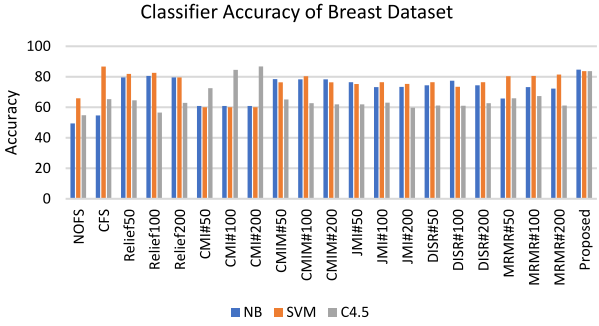
Fig. 2. Performance comparison (in terms of accuracy %) of NB, SVM and C4.5 classifiers when provided with features selected through the proposed and other existing feature selection algorithms for each benchmark dataset.



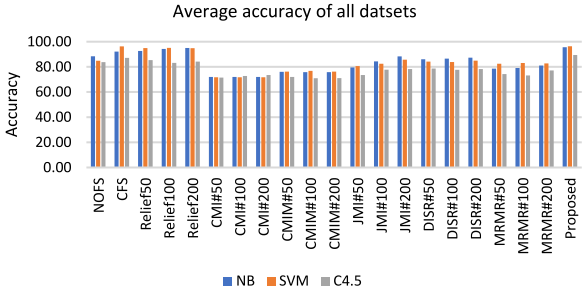
(f) Leukemia_4c dataset



(g) Ovarian dataset



(h) Breast dataset



(i) Average across all datasets

Fig. 2. (*continued*).

Algorithm 1 generates the predominant groups and Algorithm 2 is implemented to identify the optimal feature subset. Usually, the number of predominant groups will be same as the number of predominant features as a predominant feature can belong to only one predominant group. However, a redundant feature may belong to more than one predominant group as the Markov blanket for the redundant feature consists of many predominant features. In Algorithm 1, initially the features f_i are sorted in the decreasing order based on the mutual information value $MI(f_i, C)$ between the feature f_i and the class C. Then, based on the $MI\ (f_i, C)$ values all the features less than the threshold are

removed from the feature set. In this way, all the uncorrelated features that exist in the feature set are eliminated. Meanwhile, this is followed by the identification of the predominant features and formation of the predominant groups P_{set}. As the first feature in the list will not have an approximate Markov blanket, it is considered as a predominant feature. Then, it is checked if it forms the Markov blanket for any other feature present in the list. If it is present, then the feature for which the Markov blanket is formed is added to its predominant group. These steps are repeated for other predominant features, that is, the feature with the next highest MI value that are not part of any predominant group. These steps are repeated till all the predominant features are checked. The input to Algorithm 2 is predominant feature group from Algorithm 1.

The objective of Algorithm 2 is to find the optimal feature subset based on the evaluation function proposed in Eq. (19). Since exhaustive search of all 2^X possibilities is computationally expensive and not feasible, a heuristic method called neighbourhood search algorithm is used in the proposed work. The neighbourhood of the solution S is the solution which is obtained by slightly modifying S. Also, the algorithm traverses the solution space by moving a solution to its neighbourhood solutions. It starts by generating the solution with the initial value of S, then it generates a new solution S' within the neighbourhood solutions of S. If the solution S' is better than the current best solution S, then the search is continued in its neighbourhood for acquiring an improved solution S'. If the solution of S' is not better than the solution S, then the next neighbourhood is chosen and the search continues until the stopping criterion is met, i.e. the maximum value of k is reached. Finally, the optimal feature subset is obtained as the output of Algorithm 2.

4.1. Time complexity analysis

As mutual information value is to be calculated for all the features in order to filter the features with the mutual information value greater than the threshold, the computational complexity of step 4 to 8 of Algorithm 1 is O(S). For step 9, the computational complexity is $O(|P_{set}| \log |P_{set}|)$, as the fastest sorting algorithm takes $O(n \log n)$ time, where n indicates the number of elements to be sorted. Finally, steps 11 to 16, related to selection of the predominant features, incur computational complexity of $O(|P_{MI}|^2)$, as it requires comparing the mutual information for all possible pairs of features. Hence, the overall time complexity of Algorithm 1 is $O\left(S + |P_{set}| \log |P_{set}| + |P_{MI}|^2\right)$, which is $O(|P_{MI}|^2)$. The overall time complexity of Algorithm 2 is $O(k_{max})$, since the steps 3 to 8 are repeated k_{max} times ($k = 1, 2, \ldots..k_{max}$).

4.2. Monte Carlo Tree search based feature selection

In general, most of the feature selection techniques are based on the information theory methods. It adopts mutual information measure to estimate the relevance between the features. Mutual information-based methods favour the features that characterize high information values and ignores the interaction among the features. To solve this problem, a new idea based on the Monte Carlo Tree Search (MCTS) method [71] is incorporated and integrated with the proposed technique to find out the relative importance of the features (feature interaction) with the view to improve the performance. This technique is entirely different from that of the usual techniques. It aims to select the features by discovering the interdependencies among the features by considering the multidimensional dependency among the classes and the sequences of the features. Let there be n features and s subset of d features be selected. Then, t trees are constructed randomly and performance of each tree is assessed. To obtain the relative importance of the features, numerous trees are constructed by the randomly selected feature subsets. Each of the selected t trees in a particular loop is trained and performance of the trees is assessed by the randomly selected training and test set. To determine the relative importance of the feature, weighted accuracy of the tree is considered to overcome the influence of majority and minority class. Let m_{ij} be the number of instances from the class i that are classified as the class j, where $i, j = 1, 2, 3 \ldots\ldots..t_p$,. i.e mean of t_p true positive rates. Then, the weighted accuracy is defined as in Eq. (20).

$$wA = \frac{1}{t_p} \sum_{j=1}^{t_p} \frac{m_{ii}}{m_{i1} + m_{i2} + \cdots + m_{it_p}} \tag{20}$$

From Eq. #(20), the weighted accuracy is predicted. Among the features that form the tree, highly informative feature f_j is the one whose split yields the highest wA in the entire tree. After identifying the informative feature f_j, its relative importance (RI_{f_j}) is computed through gain ratio evaluation metric. The computation is mathematically expressed in Eq. (21).

$$RI_{f_j} = \sum_{t=1}^{st} (wA)^u \sum_{n_{f_j}} IG\left(n_{f_j}(t)\right) \left(\frac{no.in\ n_{f_j}(t)}{no.in\ t}\right)^v \tag{21}$$

In Eq. (21), $IG(n_{f_j}(t))$ denotes the information gain measure for the node $(n_{f_j}(t))$, $no.in\left(n_{f_j}(t)\right)$ indicates the number of samples in the node, $no.int$ represents the number of instances in the root node of the t^{th} tree and u *and* v signify the fixed number of positive reals. Further, it is to be noted that trees with lower wA are penalized and are not taken into account for further computation [72]. Similarly, greater the value of v, lesser is the influence of the node $n_{f_j}(t)$, with the given ratio $\left(\frac{no.in\ n_{f_j}(t)}{no.in\ t}\right)$ on RI_{f_j}, unless $n_{f_j}(t)$ is the root of the tree. Furthermore, it can be derived that, for any fixed positive reals on v, the node value on RI_{f_j} decreases with large number of samples.

For the purpose of performance comparison (for computing weighted accuracy), K-NN and RIPPER (Repeated Incremental Pruning to Produce Error Reduction) classifiers are employed. K-NN is a simple, non-parametric and effective learning algorithm. Further, it yields better results in a similar experimental setup [73]. RIPPER algorithm is generally a rule-based learner which builds a set of rules in order to minimize the amount of errors. RIPPER uses a three step process (grow, prune and optimize) to achieve the best results. The R source code for MCTS is available in [74]. After computing the weighted accuracy, based on the relative importance, the low-level nodes in the trees are ignored. Thus, the informative features are obtained for the classification task.

The following section presents the experimental setup, outcomes and the related discussions.

5. Results and discussion

This section presents the performance comparison in terms of empirical evaluation of the proposed feature selection work with various existing feature selection methods. Firstly, various experiments are carried out to highlight the effectiveness of the proposed method. Secondly, in order to justify the performance of the proposed feature selection method, seven standard feature selection algorithms namely CFS, Relief, mRMR, DISR, JMI, CMIM and CMI are compared.

Table 1
Specification of the considered benchmark datasets.

S.no	Datasets	ID	Number of samples	Number of features	Number of classes	IR
1	Leukemia	leuk	72	7130	2	1.88
2	Leukemia_3c	le3c	72	7130	3	4.22
3	Lung	lung	203	12601	5	23.6
4	SRBCT	srbt	83	2309	4	2.63
5	Lymphoma	lymp	66	4027	2	5.11
6	Leukemia_4c	le4c	72	7130	4	9.50
7	Ovarian	ovrn	253	15155	2	1.78
8	Breast	brst	97	24482	2	0.90

5.1. Dataset description

To verify and validate the effectiveness of the proposed system, eight benchmark datasets, publicly available in UCI repository, are used. Table 1 presents the detailed information and characteristics of these datasets.

For these datasets, the imbalance ratio (IR) is computed by dividing the total samples of the majority class to the total samples of the minority class [75]. The larger the value of the ratio, the more imbalanced dataset [76]. An imbalance ratio value of 1 suggests the equality nature of the samples in different classes. On the other hand, if the value of imbalance ratio is greater than 2, then the dataset is considered as an imbalanced one.

5.2. Evaluation criteria

With the view to demonstrate the outstanding performance of the proposed feature selection algorithm, seven existing standard feature selection algorithms namely CFS, Relief, mRMR, DISR, JMI, CMIM and CMI are employed and their performances are compared with that of the proposed algorithm. The outcome of the feature selection algorithms is provided as input to classification algorithms in order to assess the effectiveness of the selected features in building classification models. For this purpose, i.e., to validate the effectiveness of the proposed algorithm, three standard classifiers namely Naïve Bayes (NB), Support Vector Machine (SVM) and C4.5 are employed. Classification accuracy is the common metric used to evaluate the performance of classifiers. Each feature subset produced from the proposed algorithm is evaluated with the classifiers using k-fold cross validation. During k-fold validation, the dataset is randomly divided into k sets. Initially, the first $k-1$ sets are used for training and the k^{th} set is used for testing. Then, the process continues for k times with subsequent $k-1$ sets as training set and the remaining one as the test set. In this work, 10-fold cross validation is adopted. Therefore, the dataset is split into 10 sets, the classification algorithm is run for 10 times, during each time, nine sets are used for training and the remaining one set is used for testing. The primary objective of the proposed work is to achieve the highest accuracy with less number of features. In addition to accuracy, other metrics namely precision, recall and f-measure are also used to demonstrate the effectiveness of the classifiers. These performance metrics are calculated based on the True Positives (TP), False Positives (FP), True Negatives (TN) and False Negatives (FN). The formulae for computing classification accuracy, precision, recall and f-measure are given in Eqs. (22), (23), (24) and (25) respectively.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{22}$$

$$Precision = \frac{TP}{TP + FP} \tag{23}$$

$$Recall = \frac{TP}{TP + FN} \tag{24}$$

$$F - Measure = \frac{2\,(Precision * Recall)}{Precision + Recall} \tag{25}$$

5.3. Classifiers for validation

In order to investigate the significance of the proposed algorithm, three standard classifiers namely Support Vector Machine (SVM), C4.5 Decision Tree and Naïve Bayes (NB) classifiers are used. The quality of the feature subset obtained from the proposed technique is validated by using these three widely used classifiers. SVM classifier recognizes and analyses the patterns in the labelled dataset based on the input and output mapping functions and the hyper-plane that separates the class. It produces fine separation of the class based on the boundaries and the nearest data points of any of the classes. Naïve Bayes is the simplest probabilistic classifier that works on the basis of Bayes theorem. It predicts the class based on the probability of the class with strong independence assumptions between the features. C4.5 is a decision tree based classifier. It builds a decision tree from the training data by computing information entropy. Normalized information gain is used as the splitting criterion. This measure is computed for each feature and the one that characterizes the highest value is chosen at the split to make further decision. Here, the dataset is split into two smaller subsets based on the selected feature and the split value. Once, the entire tree is constructed, overfitting is reduced by pruning.

The output of each of the feature selection algorithm considered for comparison is provided as input to these three classifiers (NB, SVM and C4.5) in order to assess discriminative power of the selected features and performance. Having explained the benchmark datasets and the experimental setup, the following sub-section reports the experimental results.

5.4. Experimental results

In this section, the performance of the proposed technique is compared with that of the existing feature selection methods. Firstly, the number of features selected by various algorithms is highlighted. Out of the seven standard feature selection algorithms and the proposed feature selection method, the number of features is dynamically decided only in CFS and the proposed algorithm. In other methods, the number of features should be mentioned beforehand. Table 2 records the number of features in the original dataset (No-FS), number of features selected by the proposed algorithm and the number of features chosen by the CFS technique. In the case of other state-of-art techniques, three experiments are conducted with number of selected features be set as 50, 100 and 200. A maximum of only 200 is set as the proposed algorithm selects only less than 200 features in majority of the datasets. In the following tables, the number of features selected is indicated by the number preceded by a '#'. Table 3, Tables 4 and 5 tabulates the classification accuracy and its standard deviation with statistical significance p-value yielded by Naïve Bayes, SVM and C4.5 classifiers when provided with the features obtained from the considered feature selection algorithms. The value in the parenthesis denotes p-value of each algorithm on the appropriate dataset. Furthermore, two-tailed t-test [77–80] with 5% significance level is carried out in order

Table 2
Number of features selected by various feature selection algorithms on benchmark datasets.

	leuk	le3c	lung	srbt	lymp	le4c	ovrn	brst
No FS	7130	7130	12601	2309	4027	7130	15155	24482
CFS	79	104	548	111	185	119	35	138
Proposed	31	24	103	97	87	8	242	12

to assess the statistical significance of the proposed technique. The last row $W/T/L$ provides the Wins as (+), Ties as (=) and Loses as (−), when the proposed technique is compared over the state-of-art methods. The benchmark datasets used in this work covers binary class as well as multi-class. Moreover, the number of features varies between 2000 to nearly 25 000. Table 3 represents the classification accuracy and its standard deviation with statistical significance p-value by the Naïve Bayes classifier with the inputs provided by the considered feature selection algorithms on the datasets namely Leukemia, Leukemia_3c, Lung, SRBCT, Lymphoma, Leukemia_4c, Ovarian and Breast. The last column of Table 3 reveals the average accuracy achieved through a particular feature selection method across all the datasets. On analysing the outcomes reported in Table 3, it is evident that the proposed algorithm achieves better accuracy with less number of features. Average accuracy achieved by Naïve Bayes classifier with the feature inputs from the proposed algorithm is 95.57%. In the case of individual datasets, an accuracy of 100%, 97.32%, 97.02%,100%, 100%, 87.5%, 98.03%, 84.67% is achieved for Leukemia, Leukemia_3c, Lung, SRBCT, Lymphoma, Leukemia_4c, Ovarian and Breast datasets respectively. In general, if p-value is less than 5%, the test suggests that the improvement has not occurred by chance. The sign '+' indicates that the proposed method performs better than the compared method; '=' represents that the compared method performs equally with respect to the proposed method and '−' indicates that the proposed method performs worse than the compared method. From the last row of Table 3 that displays the win/tie/loss (W/T/L) statistics of the proposed method relative to that of the other methods, it can be inferred that the proposed method demonstrates a 'Win' with majority of the techniques or exhibits a 'Tie' with few of the existing methods and no 'Loss' is reported. Thus, the proposed algorithm in combination with Naïve Bayes classifier exhibits considerable performance with respect to all the considered datasets and demonstrates high statistical significance.

Subsequently, classification accuracy and its standard deviation obtained through SVM classifier with the inputs given by the considered seven feature selection algorithms is reported in Table 4. The outcome of pairwise t-test as W/T/L on comparing the proposed technique with that of the existing techniques is presented in the last row. Further, average classification accuracy yielded by SVM across all the selected datasets along with its standard deviation and statistical significance p-value for the various state-of-art feature selection techniques and the proposed method is recorded in the last column of Table 4. The values reported in Table 4 demonstrate the outstanding performance achieved by the proposed feature selection technique even with minimum number of features. An average accuracy of 96.36% is accomplished through SVM classifier with the inputs obtained from the proposed algorithm across all the datasets. Further, in the context of individual datasets, an accuracy of 100% for Leukemia, SRBCT, Lymphoma and Ovarian datasets, 94.64%, 96.52%, 96.02%, 83.67%, for Leukemia_3c, Lung, Leukemia_4c and Breast datasets respectively is attained. Further, on investigating the win/tie/loss (W/T/L) statistics of the proposed method relative to that of the other methods with respect to SVM classifier, it is derived that the proposed method wins over majority of the techniques, performs equally with few existing methods and does not perform worse than any of the existing methods. Thus, the proposed algorithm in combination with SVM exhibits appreciable performance with respect to all the considered datasets and demonstrates high statistical significance.

Following that, Table 5 reveals the accuracies and its standard deviation attained by C4.5 classifier based on inputs from various feature selection methods on considered benchmark datasets. The same observations, as discussed in the previous experimental tables, can be seen here. The proposed approach yields better average accuracy among other compared techniques. Across all the datasets, an average accuracy of 89.43% is attained by C4.5 decision tree classifier when the features selected through the proposed feature selection algorithm is given as input. In the context of individual datasets, an accuracy of 91.61%, 90.18%, 90.02%, 87.92%, 91.91%, 80.54%, 99.6% and 83.67% is yielded for Leukemia, Leukemia_3c, Lung, SRBCT, Lymphoma, Leukemia_4c, Ovarian and Breast datasets respectively. Further, on examining the win/tie/loss (W/T/L) statistics of the proposed method relative to that of the other methods with respect to C4.5 classifier, it is deduced that the proposed method wins over majority of the existing techniques and performs equally with few existing methods; however one loss is observed when the proposed technique is compared with Relief #100 method on Leukemia_4c dataset.

Furthermore, other performance metrics namely precision, recall and f-measure of Naïve Bayes, SVM and C4.5 classifiers with input from the considered feature selection algorithms on the benchmark datasets are investigated. Table 6, Tables 7 and 8 represent the precision, recall and f-measure for Naïve Bayes, SVM and C4.5 classifiers respectively. Analysis of the values presented in these tables (Tables 6–8) reveals that the proposed algorithm exhibits better performance in terms of average precision, recall and f-measure when given to all the three classification algorithms. From Table 6, it is observed that the highest average precision, recall and f-measure of 0.956, 0.940 and 0.952 are achieved by Naïve Bayes classifier when the features selected by the proposed algorithm is given as input. Further, with respect to non-mutual information based methods, Relief and CFS yield subsequent higher average performance measures while in the context of mutual information based methods, features selected by DISR and JMI provide the subsequent higher average precision, recall and f-measure metrics. On the other hand, features chosen through CMI feature selection method on Naïve Bayes give the least performance among all the approaches. Further, these metrics are examined with respect to individual datasets and the following observations are made. Highest precision, recall and f-measure value of 1 is achieved for Leukemia, SRBCT and Lymphoma datasets through the features selected by the proposed feature selection algorithm. The next two highly performing feature selection approaches with respect to Leukemia dataset are CFS and Relief methods. In the context of SRBCT dataset, feature selection methods namely CFS, Relief, DISR and mRMR perform equally with the proposed algorithm. With regard to Lymphoma dataset, CFS, Relief and JMI yield more or less similar performance when compared to that of the proposed algorithm. In case of Leukemia_3c dataset, the proposed algorithm performs the best while Relief method contributes to the second highest performance. With the view to Lung dataset, the proposed algorithm achieves the best results; CMIM, DISR and mRMR methods also perform equally with the proposed algorithm. In case of Leukemia_4c dataset, the proposed algorithm yields the best performance while Relief yields the second highest. In the context

Table 3
Classification accuracy and its standard deviation with statistical significance *p*-value by Naïve Bayes classifier using different feature selection techniques on benchmark datasets.

	leuk	le3c	lung	srbt	lymp	le4c	ovrn	brst	AVG
NoFS	98.57 ± 4.29 (0.331) (=)	96.07 ± 6.02 (0.648) (=)	88.57 ± 6.69 (0.003) (+)	98.75 ± 3.75 (0.003) (+)	93.81 ± 7.62 (0.025) (+)	92.14 ± 10.2 (0.412) (=)	89.85 ± 7.88 (0.008) (+)	49.44 ± 10.44 (0) (+)	88.40
CFS	98.57 ± 4.29 (0.331) (=)	95.89 ± 6.29 (0.611) (=)	95.07 ± 4.47 (0.306) (=)	100 ± 0 (1) (=)	100 ± 0 (1) (=)	93.04 ± 6.98 (0.278) (=)	100 ± 0 (0.037) (+)	54.67 ± 6.12 (0) (+)	92.16
Relief #50	95.71 ± 6.55 (0.065) (=)	94.46 ± 6.8 (0.336) (=)	85.12 ± 7.17 (0) (+)	100 ± 0 (1) (=)	100 ± 0 (1) (=)	87.68 ± 9.46 (0.974) (=)	98.42 ± 2.64 (0.007) (+)	79.56 ± 11.67 (0.238) (=)	92.62
Relief #100	95.71 ± 6.55 (0.065) (=)	97.14 ± 5.37 (0.946) (=)	88.6 ± 7.42 (0.006) (+)	100 ± 0 (1) (=)	100 ± 0 (1) (=)	94.46 ± 6.8 (0.174) (=)	97.23 ± 2.55 (0.001) (+)	80.56 ± 11.22 (0.324) (=)	94.21
Relief #200	98.57 ± 4.29 (0.331) (=)	97.32 ± 5.71 (1) (=)	92.02 ± 6.41 (0.002) (+)	100 ± 0 (1) (=)	100 ± 0 (1) (=)	94.46 ± 6.8 (0.174) (=)	98.42 ± 1.94 (0.002) (+)	79.56 ± 11.67 (0.238) (=)	95.04
CMI#50	65.36 ± 6.35 (0) (+)	52.86 ± 5.71 (0) (+)	84.71 ± 3.55 (0) (+)	90.42 ± 9.87 (0.009) (+)	69.52 ± 2.33 (0) (+)	52.86 ± 5.71 (0) (+)	99.2 ± 1.6 (0.268) (=)	60.89 ± 17.87 (0.001) (+)	71.98
CMI#100	65.36 ± 6.35 (0) (+)	52.86 ± 5.71 (0) (+)	84.71 ± 3.55 (0) (+)	90.42 ± 9.87 (0.009) (+)	69.52 ± 2.33 (0) (+)	52.86 ± 5.71 (0) (+)	99.2 ± 1.6 (0.268) (=)	60.89 ± 17.87 (0.001) (+)	71.98
CMI#200	65.36 ± 6.35 (0) (+)	52.86 ± 5.71 (0) (+)	84.71 ± 3.55 (0) (+)	90.42 ± 9.87 (0.009) (+)	69.52 ± 2.33 (0) (+)	52.86 ± 5.71 (0) (+)	99.2 ± 1.6 (0.268) (=)	60.89 ± 17.87 (0.001) (+)	71.98
CMIM#50	65.36 ± 6.35 (0) (+)	52.86 ± 5.71 (0) (+)	95.52 ± 3.5 (0.362) (=)	97.5 ± 5 (0.005) (+)	69.52 ± 2.33 (0) (+)	52.86 ± 5.71 (0) (+)	96.45 ± 4.85 (0.002) (+)	78.44 ± 11.45 (0.149) (=)	76.06
CMIM#100	65.36 ± 6.35 (0) (+)	52.86 ± 5.71 (0) (+)	95.57 ± 4.09 (0.418) (=)	98.75 ± 3.75 (0.003) (+)	69.52 ± 2.33 (0) (+)	52.86 ± 5.71 (0) (+)	92.85 ± 6.15 (0.032) (+)	78.33 ± 10.65 (0.002) (+)	75.76
CMIM#200	65.36 ± 6.35 (0) (+)	52.86 ± 5.71 (0) (+)	96.52 ± 3.9 (0.772) (=)	97.64 ± 4.73 (0.001) (+)	69.52 ± 2.33 (0) (+)	52.86 ± 5.71 (0) (+)	93.28 ± 4.34 (0.011) (+)	78.33 ± 10.65 (0.012) (+)	75.80
JMI#50	70.89 ± 19.95 (0) (+)	58.39 ± 15.2 (0) (+)	94.52 ± 5.68 (0.009) (+)	92.64 ± 9.94 (0.039) (+)	92.62 ± 9.79 (0.036) (+)	51.43 ± 12.7 (0) (+)	99.22 ± 1.57 (0.259) (=)	76.44 ± 13.45 (0.1) (+)	79.52
JMI#100	81.07 ± 16.76 (0.003) (+)	66.96 ± 18.44 (0) (+)	96.05 ± 3.73 (0.564) (=)	96.39 ± 5.53 (0.006) (+)	95.48 ± 6.94 (0.003) (+)	67.14 ± 21 (0.024) (+)	98.43 ± 1.92 (0.716) (=)	73.22 ± 13.12 (0.024) (+)	84.34
JMI#200	89.29 ± 12.8 (0.022) (+)	80.89 ± 13.6 (0.003) (+)	96.52 ± 3.9 (0.772) (=)	95.14 ± 5.97 (0.025) (+)	98.57 ± 4.29 (0.001) (+)	74.29 ± 15.05 (0.062) (=)	98.82 ± 1.81 (0.469) (=)	73.33 ± 12.72 (0.022) (+)	88.36
DISR#50	83.75 ± 16.34 (0.008) (+)	76.79 ± 14.55 (0.001) (+)	90.05 ± 7.75 (0.023) (+)	100 ± 0 (1) (=)	100 ± 0 (1) (=)	71.96 ± 17.09 (0.044) (+)	91.34 ± 6.51 (0.01) (+)	74.44 ± 13.52 (0.046) (+)	86.04
DISR#100	88.75 ± 12.42 (0.014) (+)	75.54 ± 17.39 (0.002) (+)	86.55 ± 10.28 (0.009) (+)	100 ± 0 (1) (=)	100 ± 0 (1) (=)	69.29 ± 12.7 (0.008) (+)	94.52 ± 6.06 (0.129) (+)	77.44 ± 13.31 (0.001) (+)	86.51
DISR#200	83.93 ± 16.37 (0.009) (+)	76.79 ± 14.55 (0.001) (+)	88.55 ± 9.27 (0.019) (+)	100 ± 0 (1) (=)	100 ± 0 (1) (=)	76.43 ± 13.87 (0.099) (=)	98.43 ± 2.58 (0.008) (=)	74.44 ± 14.24 (0.056) (=)	87.32
MRMR#50	67.14 ± 15.07 (0) (+)	53.93 ± 22.27 (0) (+)	94.52 ± 5.68 (0.006) (+)	100 ± 0 (1) (=)	98.57 ± 4.29 (0.003) (+)	55.18 ± 22.37 (0.001) (+)	93.32 ± 5.78 (0.039) (+)	65.78 ± 14.3 (0.001) (+)	78.56
MRMR#100	67.32 ± 16.37 (0) (+)	58.57 ± 22.37 (0) (+)	94.55 ± 4.16 (0.001) (+)	100 ± 0 (1) (=)	100 ± 0 (1) (=)	50 ± 23.9 (0.001) (+)	88.97 ± 6.23 (0.001) (+)	73.22 ± 13.12 (0.024) (+)	79.08
MRMR#200	74.29 ± 15.05 (0) (+)	61.43 ± 17.84 (0) (+)	94.57 ± 4.16 (0.003) (+)	100 ± 0 (1) (=)	100 ± 0 (1) (=)	54.29 ± 17.55 (0) (+)	91.72 ± 5.35 (0.005) (+)	72.22 ± 11.3 (0.007) (+)	81.07
Proposed	100 ± 0	97.32 ± 5.37	97.02 ± 3.3	100 ± 0	100 ± 0	87.5 ± 13.1	98.03 ± 2.62	84.67 ± 4.68	95.57
W/T/L	15/6/0	15/6/0	14/7/0	10/11/0	11/10/0	13/8/0	13/8/0	15/6/0	

of Breast dataset, the proposed method attains the maximum performance measures while CMIM and Relief methods provide the subsequent higher performance measures. Thus, it can be inferred that the proposed method reveals interesting outcome and better stability with respect to all the datasets on applying Naïve Bayes classifier.

Subsequently, precision, recall and f-measure values are analysed with respect to SVM classifier. The associated details are presented in Table 7. Firstly, in case of average performance measures, a maximum average precision, recall and f-measure of 0.955, 0.955 and 0.954 is accomplished through SVM classifier when the significant features selected by the proposed feature selection algorithm is given as input. Subsequent to the proposed algorithm, CFS and Relief methods yield the following high average performance measures in the context of non-mutual information based methods while JMI and DISR reveal the next high performance in the category of mutual information based methods. Again, similar to the observations made in the case of Naïve Bayes classifier, CMI yields the least average precision, recall and f-measure.

Further, the performance on individual datasets is investigated. Maximum precision, recall and f-measure value of 1 is achieved by SVM with the features chosen by the proposed algorithm for SRBCT, Lymphoma and Ovarian datasets. In case of Leukemia dataset, subsequent to the proposed method, CFS and Relief methods exhibit noteable performance. When Leukemia_3c dataset is considered, CFS, Relief and the proposed method yield remarkable performance in the descending order. With regard to Lung dataset, the proposed algorithm achieves the maximum performance while CFS and mRMR attain the subsequent higher performance. In case of SRBCT dataset, mRMR, JMI, CFS, and Relief method perform equally with that of the proposed algorithm. For Lymphoma dataset, the proposed algorithm reveals the highest performance and mRMR, JMI, CFS and Relief perform equally with it. In the context of Leukemia_4c dataset, CFS, Relief and the proposed algorithm exhibit comparable performance. In case of Ovarian dataset, the proposed, CFS and mRMR methods yield equally high performance. With regard to the Breast dataset, CFS, mRMR and the proposed algorithm demonstrate comparable performance. Thus, it can be concluded that the proposed feature selection method, when used along with SVM classifier method, provides interesting outcomes and decent performance with respect to all the considered datasets.

Eventually, precision, recall and f-measure metrics are analysed with respect to C4.5 classifier. The related details are presented in Table 9. Firstly, the average performance measures across all the datasets are investigated. C4.5 classifier yields the highest average precision, recall and f-measure values of 0.906, 0.878 and 0.890 when provided with features chosen by the proposed feature selection algorithm. Similar to the observations made from Naïve Bayes and SVM classifiers, following the proposed algorithm, CFS and Relief methods yield the subsequent

Table 4
Classification accuracy and its standard deviation with statistical significance *p*-value achieved by SVM classifier using different feature selection techniques on benchmark datasets.

	leuk	le3c	lung	srbt	lymp	le4c	ovrn	brst	Avg
NOFS	87.5 ± 9.48 (0.001) (+)	73.57 ± 9.51 (0) (+)	90.1 ± 3.94 (0.003) (+)	95 ± 8.29 (0.008) (+)	98.57 ± 4.29 (0.003) (+)	69.46 ± 10.36 (0) (+)	98.83 ± 2.48 (0.001) (+)	65.89 ± 18.57 (0.026) (+)	84.87
CFS	98.57 ± 4.29 (0.331) (=)	97.32 ± 5.37 (0.357) (=)	94.07 ± 3.76 (0.009) (+)	100 ± 0 (1) (=)	100 ± 0 (1) (=)	93.39 ± 8.68 (0.439) (=)	100 ± 0 (1) (=)	86.67 ± 12 (0.6) (=)	96.25
Relief #50	97.14 ± 5.71 (0.151) (=)	95.89 ± 6.29 (0.685) (=)	94.5 ± 4.72 (0.003) (+)	100 ± 0 (1) (=)	100 ± 0 (1) (=)	90.54 ± 8.38 (0.107) (=)	99.2 ± 1.6 (0.001) (+)	81.89 ± 15.31 (0.786) (=)	94.90
Relief #100	97.14 ± 5.71 (0.151) (=)	95.89 ± 6.29 (0.685) (=)	95.02 ± 5 (0.487) (=)	100 ± 0 (1) (=)	100 ± 0 (1) (=)	90.54 ± 8.38 (0.107) (=)	99.6 ± 1.2 (0.331) (=)	82.56 ± 9.99 (0.832) (=)	95.09
Relief #200	97.14 ± 5.71 (0.151) (=)	95.89 ± 6.29 (0.685) (=)	94.52 ± 5.68 (0.003) (+)	100 ± 0 (1) (=)	100 ± 0 (1) (=)	91.79 ± 8.76 (0.221) (=)	99.6 ± 1.2 (0.331) (=)	79.56 ± 12.49 (0.004) (+)	94.81
CMI#50	65.36 ± 6.35 (0) (+)	52.86 ± 5.71 (0) (+)	83.29 ± 3.74 (0) (+)	90.42 ± 6.93 (0.001) (+)	69.52 ± 2.33 (0) (+)	52.86 ± 5.71 (0) (+)	99.6 ± 1.2 (0.331) (=)	60 ± 20.18 (0.007) (+)	71.74
CMI#100	65.36 ± 6.35 (0) (+)	52.86 ± 5.71 (0) (+)	83.29 ± 3.74 (0) (+)	90.42 ± 6.93 (0.001) (+)	69.52 ± 2.33 (0) (+)	52.86 ± 5.71 (0) (+)	99.6 ± 1.2 (0.331) (=)	60 ± 20.18 (0.007) (+)	71.74
CMI#200	65.36 ± 6.35 (0) (+)	52.86 ± 5.71 (0) (+)	83.29 ± 3.74 (0) (+)	90.42 ± 6.93 (0.001) (+)	69.52 ± 2.33 (0) (+)	52.86 ± 5.71 (0) (+)	99.6 ± 1.2 (0.331) (=)	60 ± 20.18 (0.007) (+)	71.74
CMIM#50	65.36 ± 6.35 (0) (+)	52.86 ± 5.71 (0) (+)	94.07 ± 4.87 (0.002) (+)	98.75 ± 3.75 (0.003) (+)	69.52 ± 2.33 (0) (+)	52.86 ± 5.71 (0) (+)	100 ± 0 (1) (=)	76.33 ± 13.34 (0.041) (+)	76.22
CMIM#100	65.36 ± 6.35 (0) (+)	52.86 ± 5.71 (0) (+)	95.05 ± 5.44 (0.516) (=)	98.75 ± 3.75 (0.003) (+)	69.52 ± 2.33 (0) (+)	52.86 ± 5.71 (0) (+)	99.6 ± 1.2 (0.331) (=)	80.33 ± 13.27 (0.004) (+)	76.79
CMIM#200	65.36 ± 6.35 (0) (+)	52.86 ± 5.71 (0) (+)	94.57 ± 5.18 (0.378) (=)	98.75 ± 3.75 (0.003) (+)	69.52 ± 2.33 (0) (+)	52.86 ± 5.71 (0) (+)	99.6 ± 1.2 (0.331) (=)	76.33 ± 11.95 (0.002) (+)	76.23
JMI#50	68.21 ± 9.9 (0) (+)	62.5 ± 8.89 (0) (+)	95.07 ± 4.47 (0.472) (=)	90.69 ± 11.28 (0.024) (+)	94.05 ± 9.72 (0.008) (+)	59.82 ± 6.45 (0) (+)	99.6 ± 1.2 (0.331) (=)	75.22 ± 15.19 (0.002) (+)	80.65
JMI#100	77.86 ± 6.55 (0) (+)	66.61 ± 6.87 (0) (+)	86.25 ± 13.97 (0.048) (+)	95.28 ± 8.06 (0.009) (+)	95.48 ± 6.94 (0.006) (+)	62.5 ± 10.5 (0) (+)	99.62 ± 1.15 (0.331) (=)	76.44 ± 17.91 (0.003) (+)	82.50
JMI#200	81.96 ± 6.21 (0) (+)	72.5 ± 9.82 (0) (+)	94.55 ± 5.68 (0.004) (+)	96.53 ± 5.32 (0.006) (+)	98.57 ± 4.29 (0.003) (+)	66.61 ± 11.35 (0) (+)	99.6 ± 1.2 (0.331) (=)	75.33 ± 14.98 (0.002) (+)	85.71
DISR#50	77.86 ± 6.55 (0) (+)	61.25 ± 14.68 (0) (+)	94.55 ± 4.1 (0.003) (+)	100 ± 0 (1) (=)	100 ± 0 (1) (=)	63.75 ± 7.23 (0) (+)	99.6 ± 1.2 (0.331) (=)	76.44 ± 15.37 (0.027) (+)	84.18
DISR#100	77.68 ± 7.02 (0) (+)	68.21 ± 12.18 (0) (+)	94.05 ± 4.32 (0.002) (+)	100 ± 0 (1) (=)	100 ± 0 (1) (=)	58.21 ± 5.99 (0) (+)	99.22 ± 1.57 (0.001) (+)	73.44 ± 14.85 (0.001) (+)	83.85
DISR#200	80.71 ± 6.23 (0) (+)	66.61 ± 11.35 (0) (+)	93.05 ± 4.02 (0.007) (+)	100 ± 0 (1) (=)	100 ± 0 (1) (=)	63.75 ± 7.23 (0) (+)	98.82 ± 1.81 (0.006) (+)	76.44 ± 13.45 (0.024) (+)	84.92
MRMR#50	69.64 ± 7.32 (0) (+)	62.5 ± 13.86 (0) (+)	94.57 ± 4.73 (0.352) (=)	100 ± 0 (1) (=)	100 ± 0 (1) (=)	53.04 ± 11.35 (0) (+)	100 ± 0 (1) (=)	80.33 ± 12.3 (0.005) (+)	82.51
MRMR#100	68.04 ± 6.21 (0) (+)	65.36 ± 6.35 (0) (+)	96.05 ± 3.73 (0.794) (=)	100 ± 0 (1) (=)	100 ± 0 (1) (=)	54.29 ± 7.95 (0) (+)	100 ± 0 (1) (=)	80.56 ± 15.19 (0.006) (+)	83.04
MRMR#200	69.64 ± 9.72 (0) (+)	61.07 ± 12 (0) (+)	95.55 ± 2.69 (0.544) (=)	100 ± 0 (1) (=)	100 ± 0 (1) (=)	54.29 ± 7.95 (0) (+)	99.6 ± 1.2 (0.331) (=)	81.44 ± 11.7 (0.693) (=)	82.70
Proposed	100 ± 0	94.64 ± 6.59	96.52 ± 3.9	100 ± 0	100 ± 0	96.02 ± 4.89	100 ± 0	83.67 ± 11.84	96.36
W/T/L	16/5/0	16/5/0	13/8/0	10/11/0	10/11/0	16/5/0	4/17/0	16/5/0	

high average performance measures. Furthermore, CMI and CMIM yield the least average precision, recall and f-measure.

Then, the performance measures are examined with respect to the individual datasets. Highest precision, recall and f-measure values of 0.930, 0.916 and 0.916 for Leukemia dataset and 0.936, 0.921 and 0.899 for Leukemia_3c dataset is achieved by C4.5 with features obtained from the proposed feature selection algorithm; subsequent high performance is revealed by CFS and Relief methods for Leukemia dataset and CFS and JMI methods for Leukemia_3c dataset. In the context of Lung dataset, CFS performs the best while the proposed and Relief methods stand next. In case of SRBCT dataset, proposed algorithm attains the highest performance while Relief and JMI take up the subsequent positions. With regard to Lymphoma dataset, proposed algorithm yields the maximum performance while DISR and mRMR rank next. For Leukemia_4c dataset, proposed algorithm achieves the highest performance while Relief and CFS algorithms exhibit the next rank in performance. In the context of Ovarian dataset, proposed algorithm demonstrates maximum performance while Relief and JMI project the subsequent high performance. For Breast dataset, CMIM and CMI methods provide comparable performance with that of the proposed algorithm. Thus, it can be derived that the proposed method gives appealing outcome and superior performance with regard to all the datasets.

Further, the performance of the proposed algorithm is depicted graphically. Fig. 2 projects the classification accuracy obtained by the three classifiers (NB, SVM, and C4.5) on each benchmark dataset with the features obtained through the proposed and other considered feature selection methods. As mentioned earlier, the feature selection methods included for comparison consists of both mutual information based methods and non-mutual information based methods. Fig. 2(a) illustrates the performance comparison of combination of various feature selection methods with the three classifiers on Leukemia dataset. It is evident that the proposed algorithm demonstrates the highest performance with all the three classifiers. It is also revealed that CFS and Relief feature selection methods yield the subsequent highest classification accuracy with respect to Leukemia dataset.

Similar observation is found in case of Leukemia_3c dataset; the proposed method gives the highest classification accuracy with all the three classifiers; the associated details are illustrated in Fig. 2(b). Fig. 2(c) depicts the classification performance of the three classifiers when provided with features obtained from proposed and other considered feature selection methods on Lung dataset. It is displayed that other than CMI, all the feature selection methods exhibit better performance comparable to that of the proposed technique. Further, Fig. 2(d) presents the accuracy performance details pertaining to SRBCT dataset; except CMI, all mutual information-based approaches yield considerably good performance similar to that of the proposed method;

Table 5
Classification accuracy and its standard deviation with statistical significance *p*-value achieved by C4.5 classifier using different feature selection techniques on benchmark datasets.

	leuk	le3c	lung	srbt	lymp	le4c	ovrn	brst	Avg
NOFS	83.04 ± 10.75 (0.088) (=)	86.07 ± 10.65 (0.39) (=)	91.57 ± 5.03 (0.695) (=)	84.17 ± 7.81 (0.003) (+)	86.43 ± 18.4 (0.009) (+)	87.5 ± 9.97 (0.218) (=)	96.42 ± 4.17 (0.041) (+)	54.78 ± 15.16 (0) (+)	83.75
CFS	88.75 ± 5.65 (0.444) (=)	86.25 ± 8.53 (0.357) (=)	93.55 ± 4.99 (0.375) (=)	85.42 ± 9.13 (0.534) (=)	93.81 ± 17.24 (0.552) (=)	87.5 ± 7.66 (0.183) (=)	96.8 ± 3.92 (0.055) (=)	65.33 ± 18.2 (0.019) (+)	87.18
Relief #50	85.71 ± 9.04 (0.191) (=)	87.5 ± 11.43 (0.589) (=)	86.24 ± 6.84 (0.003) (+)	86.67 ± 8.45 (0.744) (=)	83.33 ± 17.11 (0.003) (+)	90.36 ± 11.04 (0.101) (=)	98.8 ± 1.83 (0.288) (=)	64.56 ± 21.16 (0.027) (+)	85.40
Relief #100	87.14 ± 10 (0.342) (=)	86.07 ± 10.65 (0.39) (=)	84.67 ± 8.16 (0.002) (+)	80.69 ± 7.74 (0.006) (+)	80 ± 17.5 (0.002) (+)	91.61 ± 6.87 (0.036) (−)	98.8 ± 1.83 (0.288) (=)	56.56 ± 17.77 (0.001) (+)	83.19
Relief #200	84.82 ± 7.2 (0.102) (=)	81.79 ± 13 (0.13) (=)	85.14 ± 6.78 (0.025) (+)	81.81 ± 10.12 (0.001) (+)	92.14 ± 17.97 (0.193) (=)	86.25 ± 13.97 (0.381) (=)	98.4 ± 1.96 (0.001) (+)	62.89 ± 17.36 (0.007) (+)	84.16
CMI#50	65.36 ± 6.35 (0) (+)	52.86 ± 5.71 (0) (+)	77.29 ± 5.24 (0.004) (+)	83.19 ± 7.94 (0.021) (+)	69.52 ± 2.33 (0) (+)	52.86 ± 5.71 (0) (+)	98.4 ± 1.96 (0.001) (+)	72.56 ± 14.84 (0.088) (=)	71.50
CMI#100	65.36 ± 6.35 (0) (+)	52.86 ± 5.71 (0) (+)	77.79 ± 4.73 (0.005) (+)	80.83 ± 9.66 (0.009) (+)	69.52 ± 2.33 (0) (+)	52.86 ± 5.71 (0) (+)	98.4 ± 1.96 (0.001) (+)	84.56 ± 8.1 (0.847) (=)	72.77
CMI#200	65.36 ± 6.35 (0) (+)	52.86 ± 5.71 (0) (+)	78.76 ± 5.65 (0.011) (+)	83.47 ± 8.79 (0.026) (+)	69.52 ± 2.33 (0) (+)	52.86 ± 5.71 (0) (+)	98.4 ± 1.96 (0.001) (+)	86.78 ± 8.94 (0.518) (=)	73.50
CMIM#50	65.36 ± 6.35 (0) (+)	52.86 ± 5.71 (0) (+)	85.12 ± 7.45 (0.002) (+)	86.39 ± 10.41 (0.725) (=)	69.52 ± 2.33 (0) (+)	52.86 ± 5.71 (0) (+)	98.4 ± 2.65 (0.002) (+)	65.11 ± 20.33 (0.027) (+)	71.95
CMIM#100	65.36 ± 6.35 (0) (+)	52.86 ± 5.71 (0) (+)	80.71 ± 8.59 (0.004) (+)	86.39 ± 11.81 (0.747) (=)	69.52 ± 2.33 (0) (+)	52.86 ± 5.71 (0) (+)	97.62 ± 3.66 (0.0014) (+)	62.67 ± 19.28 (0.011) (+)	71.00
CMIM#200	65.36 ± 6.35 (0) (+)	52.86 ± 5.71 (0) (+)	81.67 ± 11.25 (0.01) (+)	86.67 ± 10.13 (0.77) (=)	69.52 ± 2.33 (0) (+)	52.86 ± 5.71 (0) (+)	97.62 ± 1.95 (0.018) (+)	61.89 ± 14.21 (0.002) (+)	71.05
JMI#50	71.43 ± 13.58 (0.002) (+)	57.32 ± 19.24 (0) (+)	87.64 ± 4.64 (0.042) (+)	82.92 ± 9.9 (0.02) (+)	74.76 ± 18.6 (0.004) (+)	53.04 ± 16.52 (0.001) (+)	98.8 ± 1.83 (0.288) (=)	61.89 ± 12.54 (0.001) (+)	73.48
JMI#100	72.5 ± 14.47 (0.004) (+)	69.82 ± 16.61 (0.005) (+)	87.67 ± 6.4 (0.007) (+)	86.53 ± 10.11 (0.745) (=)	76.67 ± 10.96 (0) (+)	66.79 ± 14.09 (0.045) (+)	99.2 ± 1.6 (0.556) (=)	63 ± 19.47 (0.012) (+)	77.77
JMI#200	86.43 ± 10.2 (0.277) (=)	54.11 ± 18.04 (0) (+)	88.12 ± 7.8 (0.006) (+)	84.31 ± 7.88 (0.03) (+)	92.38 ± 7.62 (0.744) (=)	62.68 ± 16.87 (0.022) (+)	98.4 ± 1.96 (0.135) (=)	59.78 ± 19.31 (0.005) (+)	78.28
DISR#50	84.64 ± 13.53 (0.221) (=)	53.93 ± 17.36 (0) (+)	92.55 ± 5.1 (0.525) (=)	84.17 ± 9.61 (0.036) (+)	96.91 ± 5 (0.288) (=)	58.21 ± 11.27 (0.001) (+)	98 ± 2.68 (0.0012) (+)	61.11 ± 23.65 (0.018) (+)	78.69
DISR#100	72.86 ± 14.57 (0.004) (+)	59.46 ± 13.88 (0) (+)	92.57 ± 3.41 (0.498) (=)	85.42 ± 7.22 (0.481) (=)	96.91 ± 6.21 (0.569) (=)	55.36 ± 16.69 (0.002) (+)	98 ± 2.68 (0.0012) (+)	61 ± 20.66 (0.009) (+)	77.70
DISR#200	85.71 ± 12.78 (0.279) (=)	54.11 ± 14.98 (0) (+)	91.57 ± 4.56 (0.69) (=)	84.44 ± 10.61 (0.0433) (+)	95.24 ± 5 (0.288) (=)	53.93 ± 15.85 (0.001) (+)	98 ± 2.68 (0.0012) (+)	62.67 ± 15.6 (0.004) (+)	78.21
MRMR#50	69.64 ± 19.5 (0.007) (+)	44.64 ± 16.98 (0) (+)	91.57 ± 5.5 (0.7) (=)	84.17 ± 9.61 (0.0368) (+)	95.48 ± 6.67 (0.628) (=)	44.46 ± 18.77 (0) (+)	98 ± 2.68 (0.0012) (+)	65.89 ± 11.44 (0.003) (+)	74.23
MRMR#100	50 ± 16.62 (0) (+)	53.57 ± 15.32 (0) (+)	91.57 ± 5.03 (0.695) (=)	83.06 ± 6.08 (0.0148) (+)	95.24 ± 7.3 (0.947) (=)	46.96 ± 20.36 (0.001) (+)	97.62 ± 1.95 (0.018) (+)	67.33 ± 12.07 (0.008) (+)	73.17
MRMR#200	76.79 ± 15.59 (0.025) (+)	46.79 ± 19.7 (0) (+)	93.07 ± 3.36 (0.418) (=)	81.81 ± 8.44 (0.0122) (+)	95.24 ± 10.67 (1) (=)	64.29 ± 19.17 (0.05) (+)	98.4 ± 2.65 (0.232) (=)	61.11 ± 13.56 (0.001) (+)	77.19
Proposed	91.61 ± 9.38	90.18 ± 9.08	90.02 ± 10.5	87.92 ± 7.51	91.91 ± 7.64	80.54 ± 12.99	99.6 ± 1.2	83.67 ± 10.97	89.43
W/T/L	12/9/0	15/6/0	12/9/0	13/8/0	11/10/0	15/5/1	13/8/0	17/4/0	

it is to be noted that an accuracy of 100% is achieved when the significant features obtained from the proposed algorithm is provided to Naïve Bayes and SVM classifiers. Then, Fig. 2(e) projects the classification performance in terms of accuracy with respect to Lymphoma dataset. It is noticeable that when features selected through CMI and CMIM methods are presented to the classifiers, all the three classifiers yield the least performance. Excluding these two algorithms, the other feature selection methods demonstrate considerable results. Further, when the features selected through the proposed method are supplied to Naïve Bayes and SVM classifiers, an accuracy of 100% is attained. After that, Fig. 2(f) depicts classification accuracy obtained for Leukemia_4c dataset. It can be inferred that CFS and Relief feature selection methods attain maximum accuracy while other existing methods yield only low performance. However, the proposed technique yields decent performance with Naïve Bayes and C4.5 and highest accuracy with SVM classifier. Subsequently, the classification accuracy with regard to Ovarian dataset is presented in Fig. 2(g). It is deduced that the proposed algorithm achieves the highest performance; when features from proposed feature selection method is given to SVM and C4.5, an accuracy of 100% and 99.6% is achieved respectively. Also, the other feature selection techniques report comparable performance with that of the proposed technique. After that, Fig. 2(h) portrays the classification accuracy pertaining to Breast dataset. When considering breast dataset, proposed algorithm yields high and stable performance with all the three classifiers. It is also noticed that CMI gives high accuracy with C4.5 but low performance with NB and SVM. It is to be noted that the imbalance ratio for Breast dataset is 0.90 signifying that it is more or less a balanced dataset. Even then, the CMI method does not provide stable outcomes with all the classifiers. Following the illustrations of classification accuracies with respect to individual datasets, the average classification accuracy across all the datasets is projected in Fig. 2(i).

Further, ROC curve is plotted to analyse the performance of NB, SVM and C4.5 classifiers when presented with features selected from proposed feature selection algorithm with respect to each benchmark dataset. Fig. 3, Figs. 4 and 5 shows the ROC curves of Naïve Bayes, SVM and C4.5 classifier (with features from proposed feature selection method) for all the datasets used in the experimental work. Based on the analysis from the ROC graph, the proposed method exhibits good performance with all the three classifiers. It is to be highlighted that the area under ROC of Naïve Bayes classifier is 1 for Leukemia and SRBCT datasets signifying the highest performance. Further, ROC(area) of SVM is 1 for Leukemia, SRBCT, Lymphoma and Ovarian datasets signalling the high efficiency of the proposed technique. The performance of the proposed algorithm is good with all the three classifier as no classifier characterizes an ROC(area) of less than 0.5. Thus, the classification performance of the proposed technique is justified.

Friedman test is conducted in order to analyse and order the considered algorithms based on the ranking of the obtained

Table 6
Precision, recall and f-measure of Naïve Bayes classifier with various feature selection algorithms on benchmark datasets.

		leuk	le3c	lung	srbt	lymp	le4c	ovrn	brst	Avg
	Pre	0.988	0.945	0.881	0.992	0.888	0.866	0.915	0.411	0.861
NOFS	Rec	0.975	0.925	0.813	0.988	0.867	0.842	0.896	0.478	0.848
	F-Me	0.984	0.950	0.876	0.987	0.910	0.889	0.898	0.436	0.866
	Pre	0.989	0.943	0.952	1.000	1.000	0.899	1.000	0.324	0.888
CFS	Rec	0.988	0.925	0.913	1.000	1.000	0.875	1.000	0.520	0.903
	F-Me	0.986	0.948	0.946	1.000	1.000	0.909	1.000	0.400	0.899
	Pre	0.968	0.938	0.897	1.000	1.000	0.849	0.985	0.824	0.933
Relief #50	Rec	0.953	0.908	0.840	1.000	1.000	0.773	0.980	0.803	0.907
	F-Me	0.956	0.934	0.857	1.000	1.000	0.850	0.984	0.789	0.921
	Pre	0.968	0.970	0.923	1.000	1.000	0.924	0.975	0.832	0.949
Relief #100	Rec	0.953	0.958	0.860	1.000	1.000	0.908	0.966	0.812	0.932
	F-Me	0.956	0.964	0.890	1.000	1.000	0.930	0.972	0.801	0.939
	Pre	0.989	0.964	0.943	1.000	1.000	0.924	0.985	0.817	0.953
Relief #200	Rec	0.988	0.950	0.910	1.000	1.000	0.908	0.980	0.802	0.942
	F-Me	0.986	0.969	0.921	1.000	1.000	0.930	0.984	0.791	0.948
	Pre	0.431	0.283	0.821	0.917	0.484	0.283	0.993	0.624	0.605
CMI#50	Rec	0.500	0.350	0.691	0.913	0.350	0.308	0.991	0.613	0.590
	F-Me	0.518	0.367	0.823	0.897	0.570	0.367	0.992	0.600	0.642
	Pre	0.431	0.283	0.821	0.917	0.484	0.283	0.993	0.624	0.605
CMI#100	Rec	0.500	0.350	0.691	0.913	0.350	0.308	0.991	0.613	0.590
	F-Me	0.518	0.367	0.823	0.897	0.570	0.367	0.992	0.600	0.642
	Pre	0.431	0.283	0.821	0.917	0.484	0.283	0.993	0.624	0.605
CMI#200	Rec	0.500	0.350	0.691	0.913	0.350	0.308	0.991	0.613	0.590
	F-Me	0.518	0.367	0.823	0.897	0.570	0.367	0.992	0.600	0.642
	Pre	0.431	0.283	0.953	0.982	0.484	0.283	0.971	0.802	0.649
CMIM#50	Rec	0.500	0.350	0.911	0.979	0.350	0.308	0.963	0.789	0.644
	F-Me	0.518	0.367	0.950	0.974	0.570	0.367	0.965	0.782	0.687
	Pre	0.431	0.283	0.955	0.992	0.484	0.283	0.943	0.796	0.646
CMIM#100	Rec	0.500	0.350	0.903	0.992	0.350	0.308	0.923	0.787	0.639
	F-Me	0.518	0.367	0.951	0.988	0.570	0.367	0.928	0.781	0.684
	Pre	0.431	0.283	0.959	0.983	0.484	0.283	0.942	0.796	0.645
CMIM#200	Rec	0.500	0.350	0.906	0.979	0.350	0.308	0.926	0.787	0.638
	F-Me	0.518	0.367	0.959	0.975	0.570	0.367	0.932	0.781	0.684
	Pre	0.761	0.675	0.941	0.949	0.869	0.575	0.993	0.777	0.817
JMI#50	Rec	0.723	0.522	0.881	0.938	0.833	0.432	0.992	0.767	0.761
	F-Me	0.707	0.573	0.938	0.926	0.894	0.505	0.992	0.762	0.787
	Pre	0.813	0.652	0.962	0.975	0.918	0.626	0.985	0.748	0.835
JMI#100	Rec	0.777	0.600	0.923	0.967	0.900	0.560	0.980	0.734	0.805
	F-Me	0.792	0.647	0.957	0.962	0.934	0.633	0.984	0.728	0.830
	Pre	0.899	0.744	0.962	0.953	0.974	0.649	0.989	0.748	0.865
JMI#200	Rec	0.866	0.686	0.915	0.942	0.967	0.569	0.986	0.734	0.833
	F-Me	0.877	0.769	0.960	0.944	0.979	0.679	0.988	0.729	0.866
	Pre	0.887	0.774	0.940	1.000	1.000	0.631	0.934	0.767	0.866
DISR#50	Rec	0.849	0.724	0.902	1.000	1.000	0.569	0.930	0.751	0.841
	F-Me	0.836	0.746	0.906	1.000	1.000	0.667	0.915	0.741	0.851
	Pre	0.910	0.756	0.913	1.000	1.000	0.615	0.957	0.792	0.868
DISR#100	Rec	0.881	0.724	0.882	1.000	1.000	0.574	0.955	0.779	0.849
	F-Me	0.888	0.736	0.875	1.000	1.000	0.636	0.946	0.771	0.856
	Pre	0.875	0.778	0.931	1.000	1.000	0.735	0.986	0.765	0.884
DISR#200	Rec	0.822	0.754	0.881	1.000	1.000	0.687	0.983	0.749	0.859
	F-Me	0.836	0.754	0.894	1.000	1.000	0.731	0.984	0.739	0.867
	Pre	0.703	0.519	0.959	1.000	0.974	0.524	0.949	0.662	0.786
MRMR#50	Rec	0.638	0.482	0.937	1.000	0.967	0.458	0.946	0.640	0.758
	F-Me	0.658	0.508	0.945	1.000	0.979	0.518	0.934	0.601	0.768
	Pre	0.656	0.527	0.958	1.000	1.000	0.401	0.918	0.807	0.783
MRMR#100	Rec	0.593	0.461	0.914	1.000	1.000	0.365	0.912	0.723	0.746
	F-Me	0.647	0.535	0.946	1.000	1.000	0.425	0.892	0.700	0.768
	Pre	0.767	0.590	0.959	1.000	1.000	0.456	0.933	0.794	0.812
MRMR#200	Rec	0.701	0.549	0.918	1.000	1.000	0.394	0.931	0.710	0.775
	F-Me	0.734	0.574	0.947	1.000	1.000	0.472	0.919	0.693	0.792
	Pre	1.000	0.970	0.973	1.000	1.000	0.857	0.983	0.868	0.956
Proposed	Rec	1.000	0.958	0.953	1.000	1.000	0.779	0.980	0.849	0.940
	F-Me	1.000	0.969	0.968	1.000	1.000	0.857	0.980	0.845	0.952

classification accuracies [81–83]. The average ranking results obtained from the Friedman test for the proposed algorithm and the state-of-the-art methods are recorded in Table 10. It is observed that the proposed method attains superior ranking when compared to that of the other methods with respect to all the classifiers. The ranking is assigned in the increasing order from lower(best) to higher(worst) for each of the techniques. Subsequently Wilcoxon signed rank test is performed in order to demonstrate the significance and effectiveness of the proposed method as against the other methods used for comparison. Table 11 records the results obtained from the Wilcoxon test for proposed technique as against other comparison techniques in terms of p-value for the Naïve Bayes, SVM and C4.5 classifier. In Table 11, the values represented with * indicates that the proposed method rejects the particular methods at a significance level of 0.05 and values without * projects the proposed method has accepts the considered feature selection algorithms.

Table 7
Precision, recall and f-measure of SVM classifier with various feature selection algorithms on benchmark datasets.

		leuk	le3c	lung	srbt	lymp	le4c	ovrn	brst	Avg
	Pre	0.874	0.656	0.869	0.944	0.974	0.553	0.990	0.673	0.817
NOFS	Rec	0.800	0.558	0.752	0.942	0.967	0.476	0.989	0.662	0.768
	F-Me	0.851	0.664	0.876	0.943	0.979	0.597	0.988	0.654	0.819
	Pre	0.990	0.967	0.930	1.000	1.000	0.889	1.000	0.881	0.957
CFS	Rec	0.990	0.950	0.862	1.000	1.000	0.867	1.000	0.868	0.942
	F-Me	0.986	0.966	0.930	1.000	1.000	0.908	1.000	0.864	0.957
	Pre	0.977	0.957	0.948	1.000	1.000	0.873	0.992	0.833	0.948
Relief #50	Rec	0.963	0.942	0.899	1.000	1.000	0.842	0.989	0.823	0.932
	F-Me	0.970	0.952	0.939	1.000	1.000	0.879	0.992	0.812	0.943
	Pre	0.977	0.957	0.951	1.000	1.000	0.873	0.996	0.847	0.950
Relief #100	Rec	0.963	0.942	0.900	1.000	1.000	0.842	0.994	0.832	0.934
	F-Me	0.970	0.952	0.944	1.000	1.000	0.879	0.996	0.823	0.946
	Pre	0.977	0.957	0.942	1.000	1.000	0.888	0.996	0.814	0.947
Relief #200	Rec	0.963	0.942	0.899	1.000	1.000	0.867	0.994	0.802	0.933
	F-Me	0.970	0.952	0.937	1.000	1.000	0.895	0.996	0.794	0.943
	Pre	0.431	0.283	0.789	0.937	0.484	0.283	0.996	0.647	0.606
CMI#50	Rec	0.500	0.350	0.610	0.925	0.350	0.308	0.994	0.603	0.580
	F-Me	0.518	0.367	0.796	0.900	0.570	0.367	0.996	0.577	0.637
	Pre	0.431	0.283	0.789	0.937	0.484	0.283	0.996	0.647	0.606
CMI#100	Rec	0.500	0.350	0.610	0.925	0.350	0.308	0.994	0.603	0.580
	F-Me	0.518	0.367	0.796	0.900	0.570	0.367	0.996	0.577	0.637
	Pre	0.431	0.283	0.789	0.937	0.484	0.283	0.996	0.647	0.606
CMI#200	Rec	0.500	0.350	0.610	0.925	0.350	0.308	0.994	0.603	0.580
	F-Me	0.518	0.367	0.796	0.900	0.570	0.367	0.996	0.577	0.637
	Pre	0.431	0.283	0.936	0.991	0.484	0.283	1.000	0.780	0.648
CMIM#50	Rec	0.500	0.350	0.861	0.988	0.350	0.308	1.000	0.768	0.641
	F-Me	0.518	0.367	0.931	0.986	0.570	0.367	1.000	0.762	0.688
	Pre	0.431	0.283	0.939	0.991	0.484	0.283	0.996	0.824	0.654
CMIM#100	Rec	0.500	0.350	0.866	0.988	0.350	0.308	0.994	0.809	0.646
	F-Me	0.518	0.367	0.939	0.986	0.570	0.367	0.996	0.802	0.693
	Pre	0.431	0.283	0.934	0.991	0.484	0.283	0.996	0.786	0.648
CMIM#200	Rec	0.500	0.350	0.858	0.988	0.350	0.308	0.994	0.771	0.640
	F-Me	0.518	0.367	0.934	0.986	0.570	0.367	0.996	0.761	0.688
	Pre	0.577	0.507	0.942	0.916	0.895	0.489	0.996	0.762	0.760
JMI#50	Rec	0.563	0.458	0.876	0.925	0.867	0.399	0.994	0.753	0.730
	F-Me	0.597	0.545	0.941	0.897	0.915	0.508	0.996	0.748	0.768
	Pre	0.774	0.577	0.475	0.959	0.918	0.475	0.997	0.779	0.744
JMI#100	Rec	0.682	0.489	0.407	0.963	0.900	0.407	0.997	0.771	0.702
	F-Me	0.736	0.591	0.510	0.952	0.934	0.510	0.996	0.762	0.749
	Pre	0.804	0.676	0.936	0.976	0.974	0.561	0.996	0.777	0.837
JMI#200	Rec	0.725	0.556	0.866	0.971	0.967	0.461	0.994	0.761	0.788
	F-Me	0.784	0.667	0.935	0.964	0.979	0.579	0.996	0.749	0.832
	Pre	0.764	0.528	0.950	1.000	1.000	0.513	0.996	0.783	0.817
DISR#50	Rec	0.686	0.458	0.907	1.000	1.000	0.422	0.997	0.771	0.780
	F-Me	0.740	0.538	0.942	1.000	1.000	0.546	0.996	0.762	0.816
	Pre	0.751	0.619	0.948	1.000	1.000	0.415	0.993	0.751	0.810
DISR#100	Rec	0.667	0.522	0.905	1.000	1.000	0.365	0.994	0.741	0.774
	F-Me	0.724	0.618	0.937	1.000	1.000	0.460	0.992	0.732	0.808
	Pre	0.819	0.579	0.935	1.000	1.000	0.516	0.989	0.783	0.828
DISR#200	Rec	0.723	0.497	0.876	1.000	1.000	0.422	0.988	0.771	0.785
	F-Me	0.779	0.589	0.925	1.000	1.000	0.543	0.988	0.762	0.823
	Pre	0.586	0.554	0.935	1.000	1.000	0.343	1.000	0.822	0.780
MRMR#50	Rec	0.565	0.464	0.866	1.000	1.000	0.327	1.000	0.809	0.754
	F-Me	0.606	0.553	0.935	1.000	1.000	0.401	1.000	0.802	0.787
	Pre	0.523	0.581	0.963	1.000	1.000	0.329	1.000	0.832	0.778
MRMR#100	Rec	0.533	0.483	0.935	1.000	1.000	0.331	1.000	0.812	0.762
	F-Me	0.567	0.573	0.957	1.000	1.000	0.400	1.000	0.801	0.787
	Pre	0.554	0.548	0.963	1.000	1.000	0.317	0.996	0.831	0.776
MRMR#200	Rec	0.567	0.440	0.930	1.000	1.000	0.325	0.997	0.819	0.760
	F-Me	0.593	0.531	0.953	1.000	1.000	0.391	0.996	0.813	0.785
	Pre	1.000	0.921	0.966	1.000	1.000	0.954	1.000	0.857	0.962
Proposed	Rec	1.000	0.892	0.936	1.000	1.000	0.905	1.000	0.838	0.946
	F-Me	1.000	0.930	0.962	1.000	1.000	0.953	1.000	0.833	0.960

Further in order to assess the normality condition of the proposed technique as against the considered existing feature selection methods, Shapiro Wilks test is employed [82] with the significance level set to 0.05 over all the datasets with Naïve Bayes, SVM and C4.5 classifiers. Tables 12–14 show the results obtained from the Shapiro Wilks test with respect to the classification results of Naïve Bayes, SVM and C4.5 classifiers on all the datasets over the feature selection algorithms respectively. The values in the bracket represents the *p*-value and the symbol * indicates that the normality condition is not satisfied on the *p*-value.

Followingly, in order to test the significant differences between the proposed method and the rest of the considered feature selection techniques Holm procedure adopted [82,83]. Table 15 records the p-values for each of the comparison made between the proposed method and considered existing feature selection techniques. Then the feature selection techniques are

Table 8
Precision, recall and f-measure of C4.5 classifier with various feature selection algorithms on benchmark datasets.

		leuk	le3c	lung	srbt	lymp	le4c	ovrn	brst	Avg
	Pre	0.857	0.902	0.907	0.833	0.889	0.894	0.965	0.572	0.853
NOFS	Rec	0.798	0.854	0.864	0.842	0.865	0.835	0.955	0.546	0.820
	F-Me	0.827	0.853	0.904	0.817	0.873	0.871	0.964	0.538	0.831
	Pre	0.910	0.863	0.941	0.869	0.921	0.849	0.970	0.662	0.873
CFS	Rec	0.848	0.815	0.898	0.858	0.905	0.790	0.963	0.654	0.841
	F-Me	0.880	0.844	0.932	0.834	0.907	0.852	0.968	0.648	0.858
	Pre	0.902	0.894	0.863	0.882	0.906	0.898	0.989	0.657	0.874
Relief #50	Rec	0.848	0.854	0.665	0.867	0.898	0.867	0.986	0.650	0.829
	F-Me	0.851	0.862	0.854	0.852	0.890	0.891	0.988	0.640	0.853
	Pre	0.910	0.904	0.857	0.838	0.863	0.917	0.989	0.563	0.855
Relief #100	Rec	0.853	0.854	0.710	0.808	0.832	0.896	0.988	0.561	0.813
	F-Me	0.864	0.854	0.837	0.792	0.852	0.904	0.988	0.560	0.831
	Pre	0.871	0.781	0.873	0.816	0.879	0.860	0.985	0.636	0.838
Relief #200	Rec	0.831	0.725	0.755	0.792	0.832	0.788	0.985	0.630	0.792
	F-Me	0.845	0.774	0.855	0.794	0.860	0.848	0.984	0.621	0.823
	Pre	0.431	0.283	0.785	0.872	0.484	0.283	0.985	0.752	0.609
CMI#50	Rec	0.500	0.350	0.635	0.863	0.350	0.308	0.985	0.718	0.589
	F-Me	0.518	0.367	0.767	0.828	0.570	0.367	0.984	0.712	0.639
	Pre	0.431	0.283	0.791	0.865	0.484	0.283	0.985	0.871	0.624
CMI#100	Rec	0.500	0.350	0.655	0.850	0.350	0.308	0.985	0.848	0.606
	F-Me	0.518	0.367	0.773	0.802	0.570	0.367	0.984	0.843	0.653
	Pre	0.431	0.283	0.809	0.880	0.484	0.283	0.985	0.888	0.630
CMI#200	Rec	0.500	0.350	0.654	0.871	0.350	0.308	0.985	0.870	0.611
	F-Me	0.518	0.367	0.786	0.830	0.570	0.367	0.984	0.866	0.661
	Pre	0.431	0.283	0.845	0.849	0.484	0.283	0.985	0.666	0.603
CMIM#50	Rec	0.500	0.350	0.699	0.863	0.350	0.308	0.983	0.656	0.589
	F-Me	0.518	0.367	0.839	0.835	0.570	0.367	0.984	0.638	0.640
	Pre	0.431	0.283	0.814	0.842	0.484	0.283	0.980	0.640	0.594
CMIM#100	Rec	0.500	0.350	0.630	0.846	0.350	0.308	0.979	0.626	0.574
	F-Me	0.518	0.367	0.795	0.842	0.570	0.367	0.976	0.620	0.632
	Pre	0.431	0.283	0.813	0.859	0.484	0.283	0.978	0.623	0.594
CMIM#200	Rec	0.500	0.350	0.667	0.850	0.350	0.308	0.974	0.620	0.577
	F-Me	0.518	0.367	0.805	0.845	0.570	0.367	0.976	0.616	0.633
	Pre	0.712	0.579	0.858	0.822	0.666	0.489	0.989	0.619	0.717
JMI#50	Rec	0.658	0.492	0.770	0.829	0.562	0.404	0.988	0.620	0.665
	F-Me	0.687	0.556	0.859	0.806	0.691	0.496	0.988	0.605	0.711
	Pre	0.726	0.680	0.888	0.872	0.737	0.660	0.993	0.649	0.776
JMI#100	Rec	0.682	0.650	0.711	0.858	0.633	0.595	0.994	0.634	0.720
	F-Me	0.708	0.669	0.871	0.848	0.716	0.645	0.992	0.623	0.759
	Pre	0.898	0.573	0.869	0.845	0.931	0.608	0.985	0.620	0.791
JMI#200	Rec	0.852	0.425	0.767	0.846	0.920	0.527	0.983	0.593	0.739
	F-Me	0.860	0.522	0.868	0.825	0.928	0.604	0.984	0.588	0.772
	Pre	0.841	0.547	0.936	0.851	0.992	0.602	0.982	0.634	0.798
DISR#50	Rec	0.795	0.513	0.881	0.829	0.992	0.449	0.977	0.614	0.756
	F-Me	0.833	0.517	0.922	0.824	0.985	0.561	0.980	0.603	0.778
	Pre	0.778	0.564	0.941	0.872	0.970	0.507	0.982	0.648	0.783
DISR#100	Rec	0.717	0.606	0.912	0.854	0.958	0.408	0.977	0.615	0.756
	F-Me	0.723	0.557	0.925	0.838	0.966	0.514	0.980	0.596	0.762
	Pre	0.878	0.535	0.913	0.856	0.992	0.588	0.981	0.635	0.797
DISR#200	Rec	0.824	0.453	0.871	0.833	0.992	0.378	0.980	0.626	0.745
	F-Me	0.854	0.519	0.909	0.823	0.985	0.534	0.980	0.620	0.778
	Pre	0.742	0.411	0.936	0.860	0.967	0.451	0.982	0.677	0.753
MRMR#50	Rec	0.675	0.375	0.863	0.829	0.958	0.321	0.977	0.666	0.708
	F-Me	0.689	0.420	0.915	0.829	0.963	0.436	0.980	0.653	0.736
	Pre	0.515	0.541	0.924	0.840	0.945	0.468	0.978	0.698	0.739
MRMR#100	Rec	0.469	0.479	0.853	0.817	0.925	0.418	0.974	0.674	0.701
	F-Me	0.487	0.520	0.911	0.809	0.944	0.447	0.976	0.662	0.720
	Pre	0.816	0.471	0.932	0.837	0.947	0.592	0.985	0.627	0.776
MRMR#200	Rec	0.733	0.456	0.868	0.804	0.925	0.536	0.983	0.614	0.740
	F-Me	0.747	0.448	0.923	0.800	0.943	0.595	0.984	0.589	0.754
	Pre	0.930	0.936	0.902	0.882	0.937	0.805	0.996	0.861	0.906
Proposed	Rec	0.916	0.921	0.831	0.871	0.925	0.725	0.997	0.840	0.878
	F-Me	0.916	0.899	0.895	0.863	0.939	0.783	0.996	0.833	0.890

ordered based on the *pHolm* value obtained from the Holm test for Naïve Bayes, SVM and C4.5 classifiers respectively.

5.5. Discussion

The proposed method outperforms the existing feature selection method, in terms of the chosen features and the classification accuracy, with regard to majority of the considered benchmark datasets. In order to justify the effectiveness of the proposed algorithm, performance comparison is made with both mutual information (MI) and non-mutual information state-of-art feature selection methods. Among the feature selection methods that are compared for performance, CMI, CMIM, JMI, DISR and mRMR belong to the category of MI based feature selection methods while CFS and Relief belong to non-mutual information-based feature selection techniques. In this work, experiments are conducted on eight high dimensional benchmark datasets. The datasets are chosen such that they encompass imbalanced and balanced data and involve binary and multi-class so that the potential of the proposed algorithm is tested in all scenarios. Tables 3–5 present

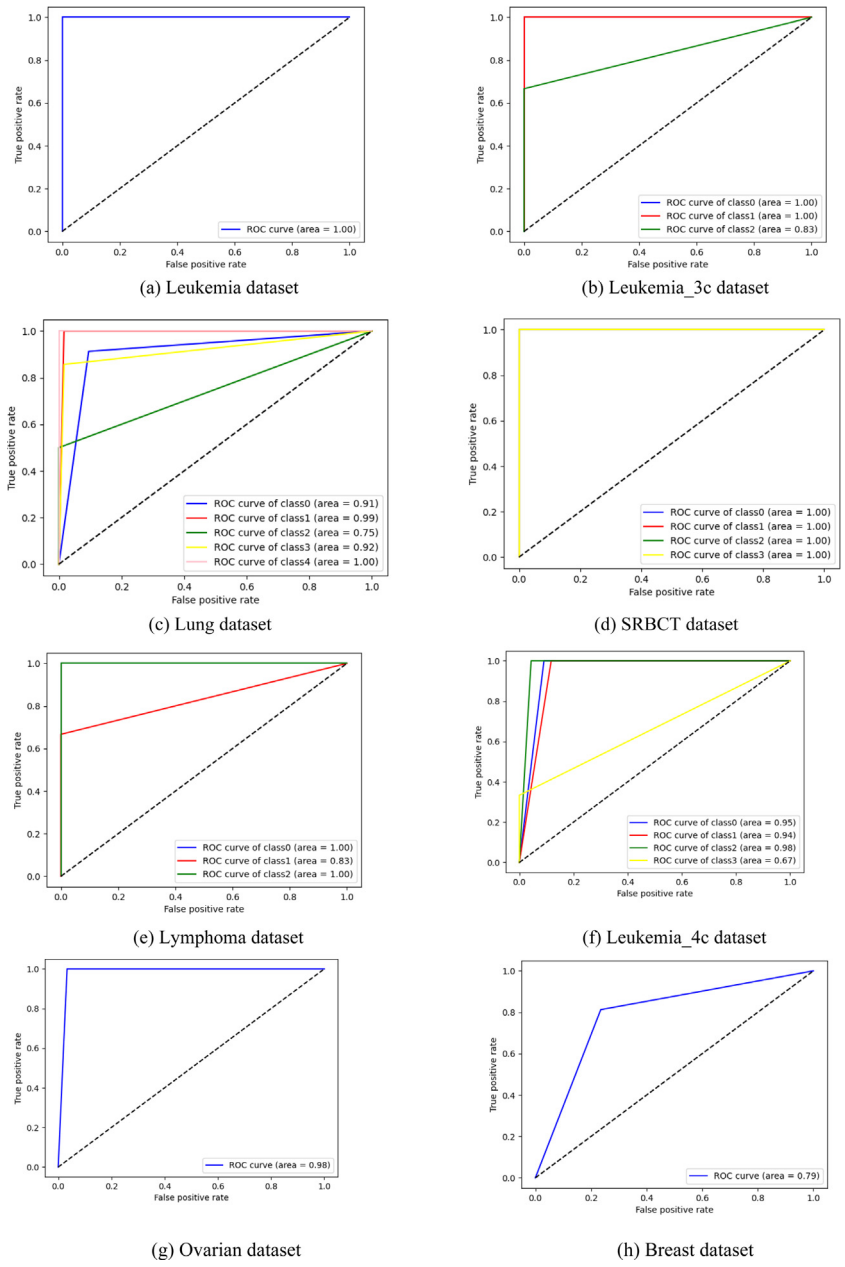


Fig. 3. ROC curve of Naïve Bayes classifier when presented with features from proposed method for each benchmark dataset.

the classification accuracies obtained by NB, SVM and C4.5 classifier with 10-fold cross validation when provided with the features selected from the proposed and the considered existing feature selection methods on each benchmark dataset.

It is evident that the highest average accuracy across all the datasets is achieved by all the three classifiers only when it is supplied with the features chosen by the proposed feature selection method. Further, the subsequent higher average accuracies are attained when features selected through the non-mutual information based methods namely CFS and Relief are given to the considered classifiers. On comparing the proposed feature selection method with MI based methods, features selected by JMI and DISR give higher average accuracies next to the accuracy yielded by the features selected through the proposed method. Further, it is noticeable that CMI and CMIM perform the least with majority of the cases. The least performance may be owed to the fact that CMI and CMIM methods rely only on relevance measures to identify significant features. Excluding these two methods, JMI, DISR and mRMR perform considerably well. Though the existing feature selection algorithms exhibit good performance in terms of average classification accuracy, some of them yield lower performance on individual datasets due to the imbalanced nature of the datasets. For instance, in case of Lung dataset that characterizes an imbalance ratio of 23.6, Naïve Bayes and C4.5 yield low performance when provided with features selected through CMI method and SVM attains lower accuracy when presented with features selected by DISR method. Yet another instance wherein Leukemia_4c dataset characterizes an imbalance ratio of 9.50, CMI, CMIM, mRMR in combination with Naïve Bayes and SVM classifier and CMI, CMIM, JMI, DISR and mRMR in combination with C4.5 yield poor performance. A large deviation in the classification performance is also noticed in these cases (Tables 3–5). Further, it is identified that the proposed method yields remarkable as well as stable accuracies on all the datasets with all the classifiers. Thus, it can be concluded that the

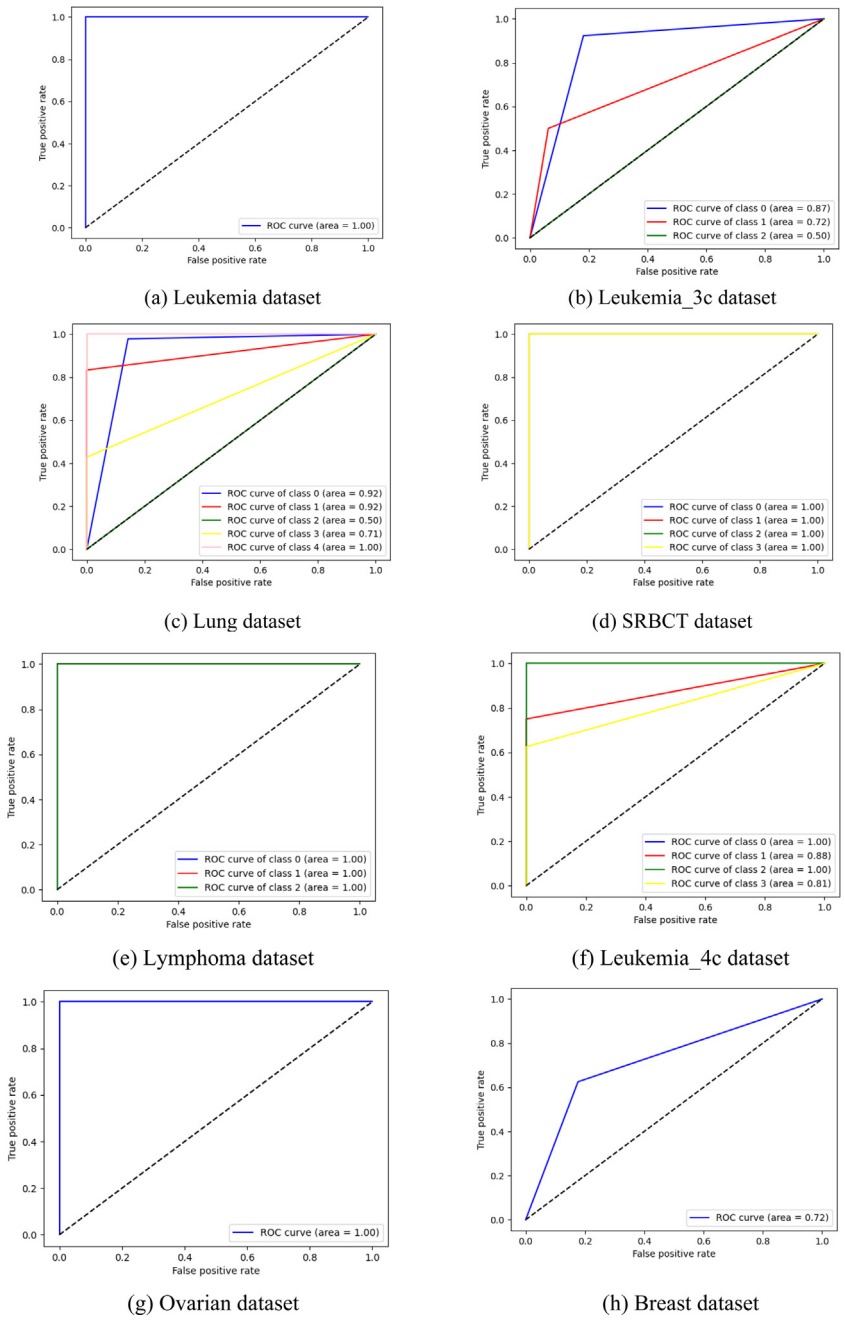


Fig. 4. ROC curve of SVM classifier when presented with features from proposed method for each benchmark dataset.

proposed method exhibits appreciable performance and handles the class imbalance issue as well.

6. A real-life application

Recently many researchers focus on analysing gene expression microarray data for early diagnosis of cancer so that a proper treatment plan can be derived. A major issue concerned with analysing gene expression data is its high dimensionality, since the microarray data characterizes large number of features (genes) with smaller number of samples. Further, in majority of microarray datasets, only few genes contribute towards prediction of cancer while the remaining features are either redundant or irrelevant in the context of cancer prediction. When these irrelevant and redundant genes are also included for building classification model to predict cancer, the classification performance may be affected. This may lead to misclassification of the types of the cancer and thereby difficulty in differentiating the tissues as contaminated or non-contaminated. From the large number of features in the microarray data, finding the relevant features to cancer prediction could provide successful treatment plans, the challenge with this is high dimensionality of data with small number of observations and the presence of irrelevant and redundant features. In this way, finding important features to build efficient classification model is still a challenging task. In this context, feature selection techniques provide solutions to find out the relevant genes/features from high dimensional data. However, even though being analysed for decades, feature selection is still a difficult and challenging task in the case of high dimensional data due to the high search space necessity. The search space increases exponentially with increase in number

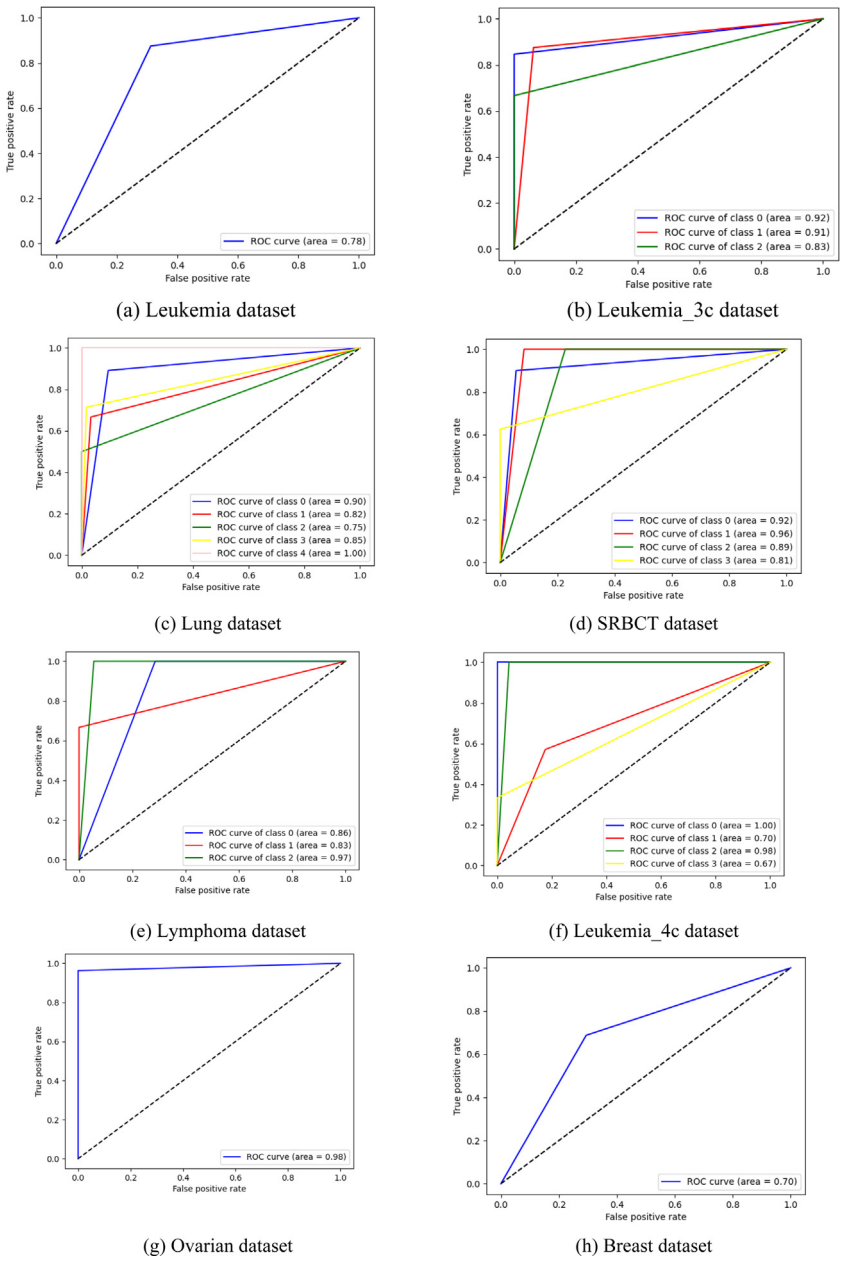


Fig. 5. ROC curve of C4.5 classifier when presented with features from proposed method for each benchmark dataset.

of features (denoted by n) as it requires exploring 2^n possible combinations. Hence, it is referred to as the combinatorial optimization problem. With this consideration, in this work, eight microarray datasets relevant to cancer prediction are taken for experimental analysis. The predominant features are identified by ignoring the irrelevant and redundant features. These features that derive meaningful interpretation, when presented to well-known classifiers result in justifiable performance. Further, the proposed technique is proposed to be enhanced and applied on text mining and image processing based applications with the view to identify strongly relevant features.

7. Conclusion

In this research work, a new feature selection strategy based on the approximate Markov blanket in fusion with the concept of mutual information and Monte Carlo tree search is proposed for selecting prominent features from high dimensional data. This method is designed to find relevant features and to eradicate the problems of choosing irrelevant and redundant features on high dimensional datasets. Initially, this method finds out irrelevant and the redundant features by grouping the features. In this way, the redundant and irrelevant features are identified and removed. Sometimes, the redundant features will be misjudged as relevant features. In order to avoid this situation, the proposed method based on the Monte Carlo tree search technique gives the best features with complete eradication of the redundant features. The proposed technique is compared with seven other mutual information based and non-mutual information based feature selection methods such as Correlation based Feature Selection (CFS), Relief, Joint Mutual Information (JMI), Conditional Mutual Information Maximization (CMIM), Double Input Symmetrical Relevance (DISR), Conditional Mutual Information (CMI) and minimal Redundancy Maximal Relevance (mRMR) with regard to identifying significant features that characterize high distinctive power and thereby contribute to efficient

Table 9
Average classification accuracies of NB, SVM and C4.5 for all the feature selection methods.

	NB	SVM	C4.5	Average
CFS	92.16	96.25	87.18	91.86
Relief50	92.62	94.90	85.40	90.97
Relief100	94.21	95.09	83.19	90.83
Relief200	95.04	94.81	84.16	91.34
CMI#50	71.98	71.74	71.50	71.74
CMI#100	71.98	71.74	72.77	72.16
CMI#200	71.98	71.74	73.50	72.41
CMIM#50	76.06	76.22	71.95	74.75
CMIM#100	75.76	76.79	71.00	74.52
CMIM#200	75.80	76.23	71.05	74.36
JMI#50	79.52	80.65	73.48	77.88
JMI#100	84.34	82.50	77.77	81.54
JMI#200	88.36	85.71	78.28	84.11
DISR#50	86.04	84.18	78.69	82.97
DISR#100	86.51	83.85	77.70	82.69
DISR#200	87.32	84.92	78.21	83.48
MRMR#50	78.56	82.51	74.23	78.43
MRMR#100	79.08	83.04	73.17	78.43
MRMR#200	81.07	82.70	77.19	80.32
Proposed	95.57	96.36	89.43	93.78

Table 10
Average rankings of the algorithms based on Friedman test.

Algorithm	Ranking		
	NB	SVM	C4.5
CFS	6.125(4)	4.3125(2)	6.3125(3)
Relief50	6.875(5)	7.3125(5)	6(2)
Relief100	6.0625(3)	5.5625(3)	9.875(10)
Relief200	5.25(2)	6.75(4)	9.4375(7)
CMI50	16(18)	17.375(18)	14(17)
CMI100	16(18)	17.375(18)	14.25(18)
CMI200	16(18)	17.375(18)	13.25(16)
CMIM50	13.25(15)	14.125(16)	12.6875(15)
CMIM100	13.5625(17)	13.4375(15)	14.75(20)
CMIM200	13.25(15)	14.5(17)	14.5(19)
JMI50	12.0625(12)	12.1875(14)	11.4375(12)
JMI100	10.375(10)	11.125(12)	7.75(4)
JMI200	8.375(7)	10.1875(10)	9.5625(9)
DISR50	9.8125(9)	9.1875(9)	9.5(8)
DISR100	9.3125(8)	11.25(13)	8.9375(5)
DISR200	8.25(6)	10.25(11)	9.125(6)
MRMR50	12.25(13)	8.375(7)	11.75(13)
MRMR100	12.375(14)	7.4375(6)	12.5(14)
MRMR200	10.8125(11)	8.75(8)	10.625(11)
Proposed	**4(1)**	**3.125(1)**	**3.75(1)**

Table 11
Results obtained from Wilcoxon test for the proposed algorithm against comparisons techniques with Naïve Bayes, SVM and C4.5 classifiers.

Algorithms	NB			SVM			C4.5		
	R+	R-	Exact *P*-value	R+	R-	Exact *P*-value	R+	R-	Exact *P*-value
CFS	21.5	14.5	0.2	13.5	14.5	0.2	23	13	0.2
Relief50	27.5	8.5	0.59766	30.5	5.5	0.09375	29	7	0.14844
Relief100	27.5	8.5	0.59766	29.5	6.5	0.12891	30	6	0.10938
Relief200	16.5	11.5	0.2	30.5	5.5	0.09375	31	5	0.07812
CMI50	35	1	0.015626*	36	0	0.007812*	36	0	0.007812*
CMI100	35	1	0.015626*	36	0	0.007812*	35	1	0.015626*
CMI200	35	1	0.015626*	36	0	0.007812*	34	2	0.02344*
CMIM50	36	0	0.007812*	28	0	0.015626*	36	0	0.007812*
CMIM100	36	0	0.007812*	36	0	0.007812*	36	0	0.007812*
CMIM200	36	0	0.007812*	36	0	0.007812*	36	0	0.007812*
JMI50	35	1	0.015626*	36	0	0.007812*	36	0	0.007812*
JMI100	35	1	0.015626*	36	0	0.007812*	36	0	0.007812*
JMI200	34	2	0.02344*	36	0	0.007812*	35	1	0.015626*
DISR50	34.5	1.5	0.019533*	34.5	1.5	0.019533*	30	6	0.10938
DISR100	34.5	1.5	0.019533*	34.5	1.5	0.019533*	29	7	0.14844
DISR200	31.5	4.5	0.0664*	34.5	1.5	0.019533*	32	4	0.05468
MRMR50	28	0	0.015626*	26.5	1.5	0.03907*	32	4	0.05468
MRMR100	34.5	1.5	0.019533*	26.5	1.5	0.03907*	32	4	0.05468
MRMR200	34.5	1.5	0.019533*	34.5	1.5	0.019533*	31	5	0.07812

Table 12
Shapiro–Wilk test for normality condition on classification accuracies of Naïve Bayes classifier.

	leuk	le3c	lung	srbt	lymp	le4c	ovrn	brst
CFS	* (0.001)	* (0.001)	(0.05)	(1)	(1)	* (0.001)	(1)	* (0.005)
Relief50	* (0.001)	* (0.001)	* (0.332)	(1)	(1)	(0.097)	* (0.001)	* (0.015)
Relief100	* (0.001)	* (0.001)	(0.687)	(1)	(1)	* (0.001)	* (0.02)	* (0.007)
Relief200	* (0.001)	* (0.001)	* (0.018)	(1)	(1)	* (0.001)	* (0.001)	* (0.017)
CMI#50	* (0.004)	* (0.002)	* (0.001)	* (0.007)	* (0.001)	* (0.002)	* (0.001)	(0.21)
CMI#100	* (0.004)	* (0.002)	* (0.001)	* (0.007)	* (0.001)	* (0.002)	* (0.001)	(0.21)
CMI#200	* (0.004)	* (0.002)	* (0.001)	* (0.007)	* (0.001)	* (0.002)	* (0.001)	(0.21)
CMIM#50	* (0.004)	* (0.002)	* (0.043)	* (0.001)	* (0.001)	* (0.002)	* (0.003)	* (0.031)
CMIM#100	* (0.004)	* (0.002)	* (0.024)	* (0.001)	* (0.001)	* (0.002)	* (0.003)	(0.13)
CMIM#200	* (0.004)	* (0.002)	* (0.01)	* (0.001)	* (0.001)	* (0.002)	(0.174)	(0.13)
JMI#50	(0.564)	(0.929)	* (0.029)	* (0.002)	* (0.004)	(0.488)	* (0.001)	(0.131)
JMI#100	(0.211)	(0.595)	* (0.029)	* (0.001)	* (0.001)	(0.147)	* (0.001)	(0.261)
JMI#200	* (0.015)	(0.376)	* (0.01)	* (0.001)	* (0.001)	(0.242)	* (0.001)	(0.317)
DISR#50	* (0.013)	(0.22)	(0.071)	(1)	(1)	(0.612)	(0.141)	(0.162)
DISR#100	* (0.005)	(0.254)	(0.355)	(1)	(1)	(0.266)	(0.053)	(0.051)
DISR#200	* (0.022)	(0.22)	(0.131)	(1)	(1)	(0.289)	* (0.001)	(0.118)
MRMR#50	(0.132)	(0.21)	* (0.029)	(1)	(0.001)	(0.715)	* (0.004)	(0.09)
MRMR#100	* (0.013)	(0.062)	* (0.033)	(1)	(1)	(0.422)	(0.951)	(0.087)
MRMR#200	(0.242)	(0.544)	* (0.033)	(1)	(1)	(0.061)	* (0.014)	(0.476)

Table 13
Shapiro–Wilk test for normality condition on classification accuracies of SVM classifier.

	leuk	le3c	lung	srbt	lymp	le4c	ovrn	brst
CFS	* (0.001)	* (0.001)	* (0.029)	(1)	(1)	* (0.003)	(1)	(0.176)
Relief50	* (0.001)	* (0.001)	(0.153)	(1)	(1)	* (0.023)	* (0.001)	(0.154)
Relief100	* (0.001)	* (0.001)	(0.076)	(1)	(1)	* (0.023)	* (0.001)	* (0.006)
Relief200	* (0.001)	* (0.001)	* (0.029)	(1)	(1)	* (0.008)	* (0.001)	* (0.019)
CMI#50	* (0.004)	* (0.002)	(0.382)	* (0.019)	* (0.001)	* (0.002)	* (0.001)	(0.144)
CMI#100	* (0.004)	* (0.002)	(0.382)	* (0.019)	* (0.001)	* (0.002)	* (0.001)	(0.144)
CMI#200	* (0.004)	* (0.002)	(0.382)	* (0.019)	* (0.001)	* (0.002)	* (0.001)	(0.144)
CMIM#50	* (0.004)	* (0.002)	(0.235)	* (0.001)	* (0.001)	* (0.002)	(1)	(0.586)
CMIM#100	* (0.004)	* (0.002)	* (0.017)	* (0.001)	* (0.001)	* (0.002)	* (0.001)	* (0.005)
CMIM#200	* (0.004)	* (0.002)	(0.081)	* (0.001)	* (0.001)	* (0.002)	* (0.001)	(0.265)
JMI#50	(0.141)	(0.057)	(0.05)	* (0.013)	* (0.001)	* (0.012)	* (0.001)	* (0.047)
JMI#100	* (0.003)	* (0.007)	(0.096)	* (0.001)	* (0.001)	* (0.01)	* (0.001)	(0.285)
JMI#200	* (0.001)	(0.107)	* (0.027)	* (0.001)	* (0.001)	* (0.008)	* (0.001)	(0.814)
DISR#50	* (0.003)	(0.485)	* (0.023)	(1)	(1)	* (0.006)	* (0.001)	(0.253)
DISR#100	* (0.002)	(0.374)	(0.161)	(1)	(1)	* (0.001)	* (0.001)	(0.078)
DISR#200	* (0.001)	(0.45)	(0.212)	(1)	(1)	* (0.006)	* (0.001)	(0.131)
MRMR#50	* (0.048)	* (0.045)	(0.154)	(1)	(1)	(0.195)	(1)	* (0.041)
MRMR#100	* (0.003)	* (0.004)	* (0.029)	(1)	(1)	(0.108)	(1)	(0.325)
MRMR#200	(0.127)	* (0.043)	* (0.007)	(1)	(1)	(0.108)	* (0.001)	* (0.002)

Table 14
Shapiro–Wilk test for normality condition on classification accuracies of C4.5 classifier.

	leuk	le3c	lung	srbt	lymp	le4c	ovrn	brst
CFS	* (0.001)	(0.075)	(0.14)	(0.071)	* (0.001)	* (0.016)	* (0.012)	(0.249)
Relief50	* (0.022)	* (0.032)	(0.321)	(0.104)	* (0.001)	* (0.009)	* (0.001)	* (0.017)
Relief100	* (0.036)	(0.068)	(0.673)	* (0.019)	* (0.006)	* (0.001)	* (0.001)	(0.396)
Relief200	* (0.02)	* (0.002)	(0.175)	* (0.001)	* (0.007)	(0.096)	* (0.001)	(0.284)
CMI#50	* (0.004)	* (0.002)	(0.357)	(0.061)	* (0.001)	* (0.002)	* (0.001)	(0.583)
CMI#100	* (0.004)	* (0.002)	(0.252)	(0.477)	* (0.001)	* (0.002)	* (0.001)	(0.555)
CMI#200	* (0.004)	* (0.002)	(0.174)	(0.275)	* (0.001)	* (0.002)	* (0.001)	(0.39)
CMIM#50	* (0.004)	* (0.002)	(0.541)	* (0.02)	* (0.001)	* (0.002)	* (0.001)	* (0.022)
CMIM#100	* (0.004)	* (0.002)	* (0.013)	(0.175)	* (0.001)	* (0.002)	* (0.001)	(0.646)
CMIM#200	* (0.004)	* (0.002)	(0.969)	* (0.039)	* (0.001)	* (0.002)	* (0.001)	(0.221)
JMI#50	* (0.014)	(0.165)	(0.17)	(0.322)	(0.696)	(0.167)	* (0.001)	(0.716)
JMI#100	(0.697)	(0.902)	(0.336)	* (0.033)	(0.096)	(0.142)	* (0.001)	(0.349)
JMI#200	(0.084)	* (0.04)	(0.355)	(0.072)	* (0.001)	(0.3)	* (0.001)	(0.419)
DISR#50	(0.175)	(0.149)	(0.295)	(0.076)	* (0.001)	(0.053)	* (0.003)	(0.19)
DISR#100	(0.176)	(0.116)	* (0.004)	(0.051)	* (0.001)	(0.127)	* (0.003)	(0.406)
DISR#200	* (0.046)	(0.055)	(0.276)	(0.291)	* (0.001)	(0.264)	* (0.003)	(0.343)
MRMR#50	(0.364)	* (0.043)	(0.371)	(0.076)	* (0.001)	(0.292)	* (0.003)	* (0.023)
MRMR#100	(0.382)	(0.163)	(0.124)	* (0.002)	* (0.001)	(0.834)	* (0.001)	*(0.756)
MRMR#200	(0.533)	(0.185)	* (0.001)	* (0.009)	* (0.001)	(0.169)	* (0.001)	*(0.117)

classification. To measure the performance and effectiveness of the proposed approach, experiments are conducted on the eight microarray datasets namely Leukemia, Leukemia_3c, Lung, Lymphoma, Leukemia_4c, Ovarian and Breast using three classifiers namely Naïve Bayes, Support Vector Machine and C4.5. Further, the proposed method is compared with existing feature selection approaches in terms of number of features selected, classification accuracy, recall, precision, F-measure, standard deviation, statistical validations. From the obtained results and experimental analysis, it is concluded that the proposed method performs better in selecting the discriminative features without redundancies thereby aiding in attaining better classification accuracy.

Table 15
Results of Holm test for proposed algorithm as against the considered state-of-the-art feature selection algorithms with Naïve Bayes, SVM and C4.5 classifiers.

Naïve Bayes			SVM			C4.5		
Algorithm	unadjusted *p*	*pHolm*	Algorithm	unadjusted textitp	*pHolm*	Algorithm	unadjusted *p*	*pHolm*
CMI50	0.00005	0.000945	CMI50	0.000001	0.000028	CMIM100	0.0002	0.003805
CMI100	0.00005	0.000945	CMI100	0.000001	0.000028	CMIM200	0.000279	0.00502
CMI200	0.00005	0.000945	CMI200	0.000001	0.000028	CMI100	0.000386	0.006558
CMIM100	0.001226	0.019619	CMIM200	0.00012	0.001925	CMI50	0.00053	0.00848
CMIM50	0.001766	0.026484	CMIM50	0.0002	0.003004	CMI200	0.00132	0.019801
CMIM200	0.001766	0.026484	CMIM100	0.00049	0.006858	CMIM50	0.002516	0.035222
MRMR100	0.004636	0.060273	JMI50	0.002186	0.028422	MRMR100	0.003096	0.040248
MRMR50	0.005287	0.063444	DISR100	0.006019	0.072227	MRMR50	0.006841	0.08209
JMI50	0.006418	0.070598	JMI100	0.006841	0.075249	JMI50	0.009354	0.10289
MRMR200	0.021276	0.212763	DISR200	0.01601	0.160099	MRMR200	0.020116	0.201162
JMI100	0.031151	0.280356	JMI200	0.016961	0.160099	Relief100	0.038394	0.345542
DISR50	0.049416	0.395329	DISR50	0.040413	0.323307	JMI200	0.049416	0.395329
DISR100	0.072502	0.507514	MRMR200	0.057224	0.400567	DISR50	0.051913	0.395329
JMI200	0.139135	0.83481	MRMR50	0.075927	0.455562	Relief200	0.054514	0.395329
DISR200	0.150786	0.83481	MRMR100	0.144871	0.724353	DISR200	0.069205	0.395329
Relief50	0.331087	1	Relief50	0.156883	0.724353	DISR100	0.079483	0.395329
CFS	0.472522	1	Relief200	0.220397	0.724353	JMI100	0.176296	0.528889
Relief100	0.485645	1	Relief100	0.409925	0.81985	CFS	0.386335	0.77267
Relief200	0.672604	1	CFS	0.68809	0.81985	Relief50	0.446873	0.77267

In future, it is of high interest to adopt other evaluation measures to assess the feature importance and classify the features as strongly relevant features, irrelevant features, weakly relevant features and redundant features based on the new criteria. In the context of application domain, the proposed technique will be useful in selecting relevant features and thereby improving accuracy of any classification task. The method can also be adapted to find its application in other domains such as text classification, image processing and so on.

CRediT authorship contribution statement

G. Manikandan: Conceptualization, Methodology, Software, Validation, Investigation, Writing - original draft, Writing - review & editing, Visualization. **S. Abirami:** Conceptualization, Supervision, Writing - review & editing.

Declaration of competing interest

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

Acknowledgements

We thank the editors and anonymous reviewer's valuable comments, guidance and suggestions for improving the earlier version of this manuscript.

References

[1] Kou Gang, Pei Yang, Yi Peng, Feng Xiao, Yang Chen, Fawaz E. Alsaadi, Evaluation of feature selection methods for text classification with small datasets using multiple criteria decision-making methods, Appl. Soft Comput. 86 (2020) 105836.
[2] Remeseiro Beatriz, Veronica Bolon-Canedo, A review of feature selection methods in medical applications, Comput. Biol. Med. 112 (2019) 103375.
[3] Gao Longwen, Shuigeng Zhou, Jihong Guan, Effectively classifying short texts by structured sparse representation with dictionary filtering, Inform. Sci. 323 (2015) 130–142.
[4] Hu Liang, Wanfu Gao, Kuo Zhao, Ping Zhang, Feng Wang, Feature selection considering two types of feature relevancy and feature interdependency, Expert Syst. Appl. 93 (2018) 423–434.
[5] Manbari Zhaleh, Fardin AkhlaghianTab, Chiman Salavati, Hybrid fast unsupervised feature selection for high-dimensional data, Expert Syst. Appl. 124 (2019) 97–118.
[6] Jović. Alan, Karla. Brkić, Nikola. Bogunović, A review of feature selection methods with applications, in: In 2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Ieee, 2015, pp. 1200–1205.
[7] Ambusaidi, A. Mohammed, Xiangjian He, Priyadarsi Nanda, Zhiyuan Tan, Building an intrusion detection system using a filter-based feature selection algorithm, IEEE Trans. Comput. 65 (10) (2016) 2986–2998.
[8] Zhang Rui, Feiping Nie, Xuelong Li, Xian Wei, Feature selection with multi-view data: A survey, Inf. Fusion 50 (2019) 158–167.
[9] Gao Wanfu, Liang Hu, Ping Zhang, Feature redundancy term variation for mutual information-based feature selection, Appl. Intell. 50 (4) (2020) 1272–1288.
[10] Wang Aiguo, Ning An, Guilin Chen, Lian Li, Gil Alterovitz, Improving PLS–RFE based gene selection for microarray data classification, Comput. Biol. Med. 62 (2015) 14–24.
[11] Chuang Li-Yeh, Cheng-Huei Yang, Kuo-Chuan Wu, Cheng-Hong Yang, A hybrid feature selection method for DNA microarray data, Comput. Biol. Med. 41 (4) (2011) 228–237.
[12] Cui Yan, Chun-Hou Zheng, Jian Yang, Wen Sha, Sparse maximum margin discriminant analysis for feature extraction and gene selection on gene expression data, Comput. Biol. Med. 43 (7) (2013) 933–941.
[13] Sharma Aman, Rinkle Rani, C-HMOSHSSA: Gene selection for cancer classification using multi-objective meta-heuristic and machine learning methods, Comput. Methods Programs Biomed. 178 (2019) 219–235.
[14] Remeseiro Beatriz, Veronica Bolon-Canedo, A review of feature selection methods in medical applications, Comput. Biol. Med. 112 (2019) 103375.
[15] García-Torres Miguel, Francisco Gómez-Vela, Belén Melián-Batista, J. Marcos Moreno-Vega, High-dimensional feature selection via feature grouping: A variable neighborhood search approach, Inform. Sci. 326 (2016) 102–118.
[16] Ma Zhen, João Manuel R.S. Tavares, Effective features to classify skin lesions in dermoscopic images, Expert Syst. Appl. 84 (2017) 92–101.
[17] Ang Jun Chin, Andri Mirzal, Habibollah Haron, Haza Nuzly Abdull Hamed, Supervised unsupervised and semi-supervised feature selection: a review on gene selection, IEEE/ACM Trans. Comput. Biol. Bioinform. 13 (5) (2015) 971–989.
[18] Jun Chin Ang, Andri Mirzal, Habibollah Haron, Haza Nuzly Abdull Hamed, Supervised, unsupervised, and semi-supervised feature selection: a review on gene selection, IEEE/ACM Trans. Comput. Biol. Bioinform. 13 (5) (2015) 971–989.
[19] Bolón-Canedo Verónica, Noelia Sánchez-Maroño, Amparo Alonso-Betanzos, Feature selection for high-dimensional data, Prog. Artif. Intell. 5 (2) (2016) 65–75.
[20] Ayesha Shaeela, Muhammad Kashif Hanif, Ramzan Talib, Overview and comparative study of dimensionality reduction techniques for high dimensional data, Inf. Fusion 59 (2020) 44–58.
[21] Peng. Hanchuan, Fuhui. Long, Chris. Ding, Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy, IEEE Trans. Pattern Anal. Mach. Intell. 27 (8) (2005) 1226–1238.
[22] Bommert Andrea, Xudong Sun, Bernd Bischl, Jörg Rahnenführer, Michel. Lang, Benchmark for filter methods for feature selection in high-dimensional classification data, Comput. Statist. Data Anal. 143 (2020) 106839.

[23] Lyu Hongqiang, Mingxi Wan, Jiuqiang Han, Ruiling Liu, Cheng Wang, A filter feature selection method based on the maximal information coefficient and Gram–Schmidt orthogonalization for biomedical data mining, Comput. Biol. Med. 89 (2017) 264–274.
[24] T. Vivekanandan, N. Ch Sriman Narayana Iyengar, Optimal feature selection using a modified differential evolution algorithm and its effectiveness for prediction of heart disease, Comput. Biol. Med. 90 (2017) 125–136.
[25] Guyon Isabelle, André Elisseeff, An introduction to variable and feature selection, J. Mach. Learn. Res. 3 (Mar) (2003) 1157–1182.
[26] Solorio-Fernández. Saúl, J. Ariel Carrasco-Ochoa, José Fco Martínez-Trinidad, A systematic evaluation of filter unsupervised feature selection methods, Expert Syst. Appl. 162 (2020) 113745.
[27] Liu Luying, Jianchu Kang, Jing Yu, Zhongliang Wang, A comparative study on unsupervised feature selection methods for text clustering, in: International Conference on Natural Language Processing and Knowledge Engineering, 2005.
[28] Varshavsky Roy, Assaf Gottlieb, Michal Linial, David Horn, Novel unsupervised feature filtering of biological data, Bioinformatics 22 (14) (2006) e507–e513.
[29] Zhao Zheng, Huan Liu, Spectral feature selection for supervised and unsupervised learning, in: Proceedings of the 24th international conference on Machine learning, 2007, pp. 1151-1157.
[30] Cai Deng, Chiyuan Zhang, Xiaofei He, Unsupervised feature selection for multi-cluster data, in: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, 2010, pp. 333-342.
[31] Öztürk Celal, Mustafa Tarım, Sibel Arslan, Feature selection and classification of metabolomics data using artificial bee colony programming (ABCP), Int. J. Data Min. Bioinform. 23 (2) (2020) 101–118.
[32] Luo Lin-Kai, Deng-Feng Huang, Ling-Jun Ye, Qi-Feng Zhou, Gui-Fang Shao, Hong Peng, Improving the computational efficiency of recursive cluster elimination for gene selection, IEEE/ACM Trans. Comput. Biol. Bioinform. 8 (1) (2010) 122–129.
[33] Li Hong-Dong, Yi-Zeng Liang, Qing-Song Xu, Dong-Sheng Cao, Bin-Bin Tan, Bai-Chuan Deng, Chen-Chen Lin, Recipe for uncovering predictive genes using support vector machines based on model population analysis, IEEE/ACM Trans. Comput. Biol. Bioinform. 8 (6) (2011) 1633–1641.
[34] Lu. Meng, Embedded feature selection accounting for unknown data heterogeneity, Expert Syst. Appl. 119 (2019) 350–361.
[35] Hu Qinghua, Wei Pan, Shuang An, Peijun Ma, Jinmao Wei, An efficient gene selection technique for cancer recognition based on neighborhood mutual information, Int. J. Mach. Learn. Cybern. 1 (1–4) (2010) 63–74.
[36] Saengsiri Patharawut, Phayung Meesad, Sageemas Na Wichian, Unger Herwig, Comparison of hybrid feature selection models on gene expression data, in: 2010 Eighth International Conference on ICT and Knowledge Engineering, IEEE, 2010, pp. 13–18.
[37] Lee Chien-Pang, Yungho Leu, A novel hybrid feature selection method for microarray data analysis, Appl. Soft Comput. 11 (1) (2011) 208–213.
[38] Kaur Kiranpreet, Nagamma Patil, A fast and novel approach based on grouping and weighted mRMR for feature selection and classification of protein sequence data, Int. J. Data Min. Bioinform. 23 (1) (2020) 47–61.
[39] Pes Barbara, Nicoletta Dessì, Marta Angioni, Exploiting the ensemble paradigm for stable feature selection: a case study on high-dimensional genomic data, Inf. Fusion 35 (2017) 132–147.
[40] Rubul Kumar Bania, Anindya Halder, R-Ensembler: A greedy rough set based ensemble attribute selection algorithm with kNN imputation for classification of medical data, Comput. Methods Programs Biomed. 184 (2020) 105122.
[41] Borja Seijo-Pardo, Verónica Bolón-Canedo, Amparo Alonso-Betanzos, On developing an automatic threshold applied to feature selection ensembles, Inf. Fusion 45 (2019) 227–245.
[42] Huawen Liu, Lei. Liu, Huijie. Zhang, Ensemble gene selection by grouping for microarray data classification, J. Biomed. Inform. 43 (1) (2010) 81–87.
[43] Pengyi Yang, Bing B. Zhou, Zili. Zhang, Albert Y. Zomaya, Zomaya a multi-filter enhanced genetic ensemble system for gene selection and sample classification of microarray data, BMC Bioinformatics 11 (1) (2010) 1–12.
[44] HongFang Zhou, Yao Zhang, YingJie Zhang, HongJiang Liu, Feature selection based on conditional mutual information: minimum conditional relevance and minimum conditional redundancy, Appl. Intell. 49 (3) (2019) 883–896.
[45] HongFang Zhou, Jing Wen, Dynamic feature selection method with minimum redundancy information for linear data, Appl. Intell. 50 (11) (2020) 3660–3677.
[46] Zhi-Chao Sha, Zhang-Meng Liu, Chen Ma, Jun Chen, Feature selection for multi-label classification by maximizing full-dimensional conditional mutual information, Appl. Intell. 51 (1) (2021) 326–340.
[47] Jundong Li, Kewei Cheng, Suhang Wang, Fred Morstatter, Robert P. Trevino, Jiliang Tang, Huan Liu, Feature selection: A data perspective, ACM Comput. Surv. 50 (6) (2017) 1–45.
[48] Gavin. Brown, Adam Pocock, Ming-Jie Zhao, Mikel Luján, Conditional likelihood maximisation: a unifying framework for information theoretic feature selection, J. Mach. Learn. Res. 13 (1) (2012) 27–66.
[49] Sungyoung Lee, Young-Tack. Park, Brian J. d'Auriol, A novel feature selection method based on normalized mutual information, Appl. Intell. 37 (1) (2012) 100–120.
[50] Sung-Nien Yu, Ming-Yuan Lee, Conditional mutual information-based feature selection for congestive heart failure recognition using heart rate variability, Comput. Methods Programs Biomed. 108 (1) (2012) 299–309.
[51] Mohamed Bennasar, Yulia Hicks, Rossitza Setchi, Feature selection using joint mutual information maximisation, Expert Syst. Appl. 42 (22) (2015) 8520–8532.
[52] Zhongsheng Hua, Jian Zhou, Ye Hua, Wei Zhang, Strong approximate Markov blanket and its application on filter-based feature selection, Appl. Soft Comput. 87 (2020) 105957.
[53] Bolón-Canedo Verónica, Noelia Sánchez-Marono, Amparo Alonso-Betanzos, José Manuel Benítez, Francisco. Herrera, A review of microarray datasets and applied feature selection methods, Inform. Sci. 282 (2014) 111–135.
[54] M. Hall, Correlation-based feature selection for discrete and numeric class machine learning, in: ICML '00 Proceedings of the Seventeenth International Conference on Machine Learning. Morgan Kaufmann, 2000, pp. 359–366.
[55] K. Kira, L. Rendell, The feature selection problem: Traditional methods and a new algorithm, in: Association for the Advancement of Artificial Intelligence, AAAI Press and MIT Press, Cambridge, MA, USA, 1992, pp. 129–134.
[56] D.D. Lewis, Feature selection and feature extraction for text categorization, in: Proceedings of the Workshop on Speech and Natural Language, Association for Computational Linguistics, 1992, pp. 212–217.
[57] Hanchuan Peng, Fuhui Long, Chris Ding, Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy, IEEE Trans. Pattern Anal. Mach. Intell. 27 (8) (2005) 1226–1238.
[58] Yang Howard Hua, John E. Moody, Data visualization and feature selection: new algorithms for nongaussian data, in: NIPS, Vol. 12, 1999.
[59] Brown. Gavin, Adam Pocock, Ming-Jie. Zhao, Mikel. Luján, Conditional likelihood maximisation: a unifying framework for information theoretic feature selection, J. Mach. Learn. Res. 13 (1) (2012) 27–66.
[60] Fleuret François, Fast binary feature selection with conditional mutual information, J. Mach. Learn. Res. 5 (9) (2004).
[61] Meyer Patrick Emmanuel, Colas Schretter, Gianluca Bontempi, Information-theoretic feature selection in microarray data using variable complementarity, IEEE J. Sel. Top. Sign. Proces. 2 (3) (2008) 261–274.
[62] Ping Zhang, Wanfu. Gao, Feature selection considering uncertainty change ratio of the class label, Appl. Soft Comput. 95 (2020) 106537.
[63] G. Wei, J. Zhao, Y. Feng, et al., A novel hybrid feature selection method based on dynamic feature importance, Appl. Soft Comput. 93 (2020) 106337.
[64] Junghye Lee, In Young Choi, Chi-HyuckJun, An efficient multivariate feature ranking method for gene selection in high-dimensional microarray data, Expert Syst. Appl. 166 (2021) 113971.
[65] Zeng Zilin, Hongjun Zhang, Rui Zhang, Chengxiang. Yin, A novel feature selection method considering feature interaction, Pattern Recognit. 48 (8) (2015) 2656–2666.
[66] Drotár Peter, Juraj Gazda, Zdenek Smékal, An experimental comparison of feature selection methods on two-class biomedical datasets, Comput. Biol. Med. 66 (2015) 1–10.
[67] Adámek. Jiří, Stefan Milius, Jiří Velebil, Bases for parametrized iterativity, Inform. and Comput. 206 (8) (2008) 966–1002.
[68] in: Proceedings of the International Conference on Applied Economics and Finance (ICOAEF IV 2018) & Extended with Social Sciences, 2018,
[69] Yu Lei, Huan Liu, Efficient feature selection via analysis of relevance and redundancy, J. Mach. Learn. Res. 5 (2004) 1205–1224.
[70] García-Torres Miguel, Francisco Gómez-Vela, Belén Melián-Batista, J. Marcos Moreno-Vega, High-dimensional feature selection via feature grouping: A variable neighborhood search approach, Inform. Sci. 326 (2016) 102–118.
[71] Chaudhry. Muhammad Umar, Jee-Hyong Lee, Feature selection for high dimensional data using Monte Carlo tree search, IEEE Access 6 (2018) 76036–76048.
[72] Dramiński Michał, Alvaro Rada-Iglesias, Stefan Enroth, Claes Wadelius, Jacek Koronacki, Jan Komorowski, Monte Carlo Feature selection for supervised classification, Bioinformatics 24 (1) (2008) 110–117.
[73] Park Chan Hee, Seoung Bum Kim, Sequential random k-nearest neighbor feature selection for high-dimensional data, Expert Syst. Appl. 42 (5) (2015) 2336–2342.
[74] Dramiński Michał, Jacek Koronacki, Rmcfs: an r package for Monte Carlo feature selection and interdependency discovery, J. Stat. Softw. 85 (1) (2018) 1–28.
[75] Amin. Adnan, Sajid Anwar, Awais Adnan, Muhammad Nawaz, Newton Howard, Junaid Qadir, Ahmad Hawalah, Amir Hussain, Comparing oversampling techniques to handle the class imbalance problem: A customer churn prediction case study, IEEE Access 4 (2016) 7940–7957.

[76] Shahee Shaukat Ali, Usha Ananthakumar, An effective distance based feature selection approach for imbalanced data, Appl. Intell. 50 (3) (2020) 717–745.
[77] Finch Holmes, Comparison of the performance of nonparametric and parametric MANOVA test statistics when assumptions are violated, Methodology 1 (1) (2005) 27–38.
[78] Zimmerman, W. Donald, Bruno D. Zumbo, Relative power of the wilcoxon test, the friedman test, and repeated-measures ANOVA on ranks, J. Exp. Educ. 62 (1) (1993) 75–86.
[79] Ashok Kumar, S. Abirami, Aspect-based opinion ranking framework for product reviews using a Spearman's rank correlation coefficient method, Inform. Sci. 460 (2018) 23–41.
[80] S. Sreejith, H. Khanna Nehemiah, Arputharaj Kannan, A classification framework using a diverse intensified strawberry optimized neural network (DISON) for clinical decision-making, Cogn. Syst. Res. 64 (2020) 98–116.
[81] Demšar Janez, Statistical comparisons of classifiers over multiple data sets, J. Mach. Learn. Res. 7 (2006) 1–30.
[82] García Salvador, Alberto Fernández, Julián Luengo, Francisco Herrera, A study of statistical techniques and performance measures for genetics-based machine learning: accuracy and interpretability, Soft Comput. 13 (10) (2009) 959.
[83] Alcalá-Fdez Jesús, Alberto Fernández, Julián Luengo, Joaquín Derrac, Salvador García, Luciano Sánchez, Francisco Herrera, Keel data-mining software tool: data set repository, integration of algorithms and experimental analysis framework, J. Mult.-Valued Logic Soft Comput. 17 (2011).