An ensemble filter-based heuristic approach for cancerous gene expression classification

Uzma[a,b], and Zahid Halim[a,*]
[a]The Machine Intelligence Research Group (MInG), Faculty of Computer Science and Engineering, Ghulam Ishaq Khan Institute of Engineering Sciences and Technology, Topi, Pakistan, 23460.
[b]Department of Computer Science, University of Wah, Pakistan

***Abstract*—**Gene expression data of cancer has a huge feature set size, making its categorization a challenge for the existing classification methods. It contains redundancy, noise, and irrelevant genes. Therefore, feature selection/reduction plays a crucial role in the classification of such gene expression datasets. This work presents an ensemble of three filter methods, namely, Symmetrical Uncertainty (SU), chi square (X^2), and Relief to reduce the feature dimensions by eliminating redundant and noisy genes. The present work designs a novel heuristic called Local Search-based Feature Selection (LSFS) that further reduces noise generated by the ensemble method. The resulting selected features are then optimized using a genetic algorithm. Afterwards, the optimal set of features is classified using three models; Support Vector Machine (SVM), *k*-NN (*k*-nearest neighbor), and Random Forest (RF) to find cancer relevant genes. Experiments are conducted using six benchmark datasets. The obtained results are compared with five state-of-the-art algorithms based on accuracy, sensitivity, specificity, F-measure, entropy, and precision. Additional experiments are carried out by manipulating the SVM kernel as a fitness value as well as using multiple distance measures and various values of *k* for *k*-NN. Prediction accuracy of the proposed system on the six benchmark datasets is 99%, 90%, 98%, 94%, 98%, and 99%. Significant outcomes obtained from experimental analysis indicate that the proposed approach improves classification of cancerous gene expression data and can be used as a practical tool for the analysis of gene expression data.

***Keywords*—**Cancerous gene, feature selection, classification, ensemble method, evolutionary algorithm

1. Introduction

The term "Cancer" is used for a condition that causes an uncontrolled cell division. It leads to a formation of stuffed mass or lumps known as tumor. This tumor is resultant of unwanted accumulated cells. Cancer cells are immortal and thus they spread rapidly in the vicinity of an origin and influence other systems [1]. To date, more than 100 types of tumors have been identified in the human body which are classified on the basis of their type and origin. Based on the tissue cells, cancer is grouped into six major classes, namely, Carcinoma, Sarcoma, Myeloma, Leukemia, Lymphoma, and a fusion of these. A cancer on the basis of origin is categorized as lung cancer, colon, breast, liver, kidney, prostate, and brain cancer.

The classification of tumor types is crucial for the better understanding and diagnosis of cancer. This also helps in the accurate prediction of tumor and its status that helps in achieving a standard quality outcome as a result of the treatment. This is because the therapy in such cases is more specific and directed rather than targeting the cancer cells blindly when enough information is not available. A limitation of the traditional approaches is their poor classification of tumor types. This issue has to be addressed by designing advanced techniques that could help to differentiate and classify various tumor cells accordingly. Here, the role of DNA (deoxyribonucleic acid) microarrays is critical as it allows to analyze the expression level of genes on a bigger scale simultaneously [2]. This success has also encouraged and widened the scope of computational evaluation and interdisciplinary fields.

These approaches are helpful for cancer prediction [3] and prognosis [4]. They also help in pattern identification and design of the classification model for the gene expression datasets. DNA microarray technology has been studied widely for the prediction of cancer. These techniques have also been proved to be fruitful in pattern extraction and design of a classification model along with cancer prediction and prognosis. However, more work is required to optimize these techniques for better disease prediction, accurate diagnosis, followed by proper medication and its monitored response [5]. Prognosis after the medication should also be obtained.

1.1. DNA microarray

The DNA microarray technology is used by many biologists to monitor the gene expression on a genomic level in a particular organism [6]. It is usually a glass slide where DNA molecules are installed regularly at certain locations, called spots (i.e., features). Microarray contain thousands of spots and each spot has millions of copies of the identical DNA molecules that uniquely fits in with the genes. They act like matrices, with known samples of DNA, cDNA, or oligonucleotides, called probes are combined with the mRNA sequences. The expression level of genes is estimated by the amount of mRNA being paired to an individual probe. The aim is to find either sets of genes that characterize a particular disease state, experimental condition or highly correlated genes that share common biological features. Microarray gene expression data contains information regarding the gene expression levels in a particular tissue. This data serves as a key source of information in different biological studies and analysis. Microarray is therefore useful in the field of oncology and cancerous gene detection.

1.2. Classification challenges in gene expression datasets

The DNA microarray datasets helps in the prediction of cancer and other critical conditions. The machine learning and data mining techniques have been used widely to analyze the gene expression profiles for the identification of these genes [7]. However, the classification of gene expression dataset is still quite challenging for many reasons, especially because of

*Corresponding author.

Email addresses: uzma@giki.edu.pk (Uzma), zahid.halim@giki.edu,pk (Z. Halim)

its small sample size and very large number of features [8]. Most of these features are irrelevant and have no role in the classification. This leads to the curse of dimensionality. Such issues are mainly overcome by the feature selection approaches and hence it makes it relatively convenient to analyze the gene expression datasets [9]. The choice of feature selection technique also depends greatly on the type of microarray data being used as it may result in complex, uneven, and overlapped data that is proved to be problematic at times [8]. Most cancer gene expression datasets are unbalanced as the number of samples belonging to various classes are uneven [10].

1.3. Feature selection

Feature selection, also called attribute/variable selection is a technique to find a subset of relevant attributes that is used for a model construction. According to the various types of data to be analyzed, feature selection can be classified into three categories; supervised, unsupervised or semi-supervised.

The process of feature selection is known as supervised learning, if each data instance in a dataset has known response values. The absence of these response values is known as unsupervised feature learning. While if some data instances in a dataset have responses available and some do not have these available, the problem is known as semi-supervised feature selection. Feature selection is imperative for research projects involving learning from data, especially in the current era of Big Data. Past research utilized varieties of feature selection methods: filters, wrappers embedded, and hybrid methods. Recently, the focus has shifted towards ensemble methods for feature selection [11, 12]. The choice of these algorithms is differentiated by the evaluation metrics.

Filter-based feature selection: Filter-based methods examine the intrinsic properties of features while evaluating the goodness of the gene subset. Filter-based algorithm adopts four types of evaluation principles, i.e., information, consistency, dependency, and distance [13] for measuring the feature characteristics. The solution providers of filter algorithm are generalized for various classifiers due to the independent nature of any learning algorithm. Filter algorithm is efficient, computationally faster and has the ability to scale for high dimensional datasets. However, the features selected by the filter method varies in prediction performance on different learning algorithms. This method also ignores the interaction among features and with the classifiers. The features are evaluated independently and degrade the performance of the learning model due to the lack of features dependency. Individual scores are assigned to each feature without considering its significance in combination with other shared features. Thus, it results in the production of redundant information. Filter-based methods have been previously applied to the microarray data. Relief [14], Fast Correlation-Based Filter (FCBF), and Correlation Feature Selection (CFS) are a few such examples.

Wrapper feature selection: The wrapper method use the subset of evaluator. The subset evaluator creates all possible subsets from feature vectors. Afterwards, it uses the classification algorithm to induce classifiers from the features in each subset. It considers the subset of features with which the classification algorithm performs the best. The wrapper method performs better as compared to the filter method because it does establish an interconnection among features and hence directly influence the learning model. It is computationally more demanding and inflexible to deal with huge datasets [15].

Ensemble feature selection: Ensemble method is a technique that aims to construct a group of feature subsets and then produce an aggregated result out of the group [16]. The ensemble method applies feature selection techniques multiple times and then the results are aggregated. Due to the combination of multiple outcomes, the best performing features and sporadic features will propagate towards the top and bottom, respectively [17]. Thus, the resulted list of features is more stable. The ensemble method consists of two parts. The first part generates multiple features' score list and the second part uses aggregation function to combine all the results obtained from the first part.

Hybrid feature selection: The hybrid method is formed by merging filter and wrapper methods. The combination of filter and wrapper method is a well-known hybrid approach [18]. This technique inherits the advantages of both methods and at different search phase they use diverse evaluation criteria to improve classification performance**.**

1.4. Our contributions and novelty

This work presents a novel design and methodology for the efficient classification of cancer disease and to better analyze the microarray datasets. The classification is done for prediction of the tumor and normal genes. The classification problem in this case has many challenges due to the intrinsic behavior of the dataset. The gene expression data contains a high dimension of genes and a smaller number of samples. There are many features which are redundant and irrelevant, hence they play no role in classification. The learning algorithm cannot work well in such situation. Therefore, to avoid the "curse of dimensionality" feature selection is crucial for the classification problem. The optimal feature subset selection from a vast number of features is a challenging phase for the classification of gene expression datasets. This work proposes an ensemble filter-based feature selection methodology for the classification of cancer gene expression data. Here, a filter with a wrapper-based methodology for an optimal feature subset selection from the cancer gene expression data is presented. The first phase of this work uses an ensemble filter-based feature selection. In this phase the Symmetrical Uncertainty (SU), chi square (X^2), and Relief methods are aggregated using an aggregation function. In the second phase, this work presents a novel heuristic approach that is named as a Local Search-based Feature Selection (LSFS) that is used with Genetic Algorithm (GA) for informative feature selection. The final stage of the proposed work employs the Support Vector Machine (SVM), *k*-Nearest Neighbor (*k*-NN), and Random Forest (RF) classifiers. The performance of the proposed work is evaluated using six cancer gene expression benchmark datasets using accuracy, sensitivity, specificity, F-measure, entropy, and precision as the evaluation metrics. Several experiments are conducted on the proposed methodology for various values of *k* and different distance measures are used for evaluating *k*-NN classifier on the

gene expression datasets. Similarly, the performance of SVM is also tested for linear, quadratic, Radial Basis Function (RBF), and polynomial kernels. The key contributions of this work are listed in the following.

- This work presents a novel evolutionary computing-based methodology for feature selection so that the efficient classification of cancer diseases can be achieved. This will enable to better analyze the microarray data.
- A filter with a wrapper-based methodology is presented for an optimum feature subset selection from the cancer gene expression data.
- For improving the stability of the feature selection techniques, the ensemble of filter methods is used in the first stage.
- The ensemble of multiple filters may generate noise due to the combination of multiple filter approaches. The proposed methodology in the current work reduces this problem by designing a novel heuristic based on Information Gain (IG) for removing this noise.
- The primary novelty of present work lies in the second phase of the proposed methodology. Where, a novel heuristic approach named as a Local Search-Based Feature Subset (LSFS) is presented with the GA for controlling the randomness and informative feature selection.
- Another novelty of this work is to use the second stage of the proposed approach for the global optimization of the first stage results. This phase uses the combination of LSFS and a meta-heuristic algorithm called the GA for optimization.
- The feature subset from the first phase has some chance of having noise due to the combination of relevant and non-redundant features. This limits the performance of the classifiers. Therefore, to remove the noise from feature subset, LSFS is applied.
- The utilization of the feature selection ensemble is yet another novelty of the present work.

The rest of the paper is organized as follows. Section 2 unfolds the reference work in the subject area which is helpful for better understanding of the current situation and challenges related to the proposed work. Section 3 explains the proposed solution. The experiments and their results are covered in Section 4. Section 5 contains discussion on the obtained results. Finally, Section 6 concludes this work.

2. Related works

This section covers the applications of gene expression data and computational techniques that are currently used in this domain. Furthermore, the additional work that has been carried out to analyze the expression of gene expression data is also discussed.

2.1. Background

The DNA microarray technology is capable of analyzing a number of genes simultaneously. This is done by collecting expression data for further evaluation. This data is utilized to explore the related biological mechanism and events involved. This may include the discovery of oncogenes, types of cancer or identification of a particular disease/condition. The expression and evaluation of gene expression data are getting more eye balls in the current era, especially from the fields of precision medicine, machine learning, and pattern recognition [1]. The real-world gene expression data contain many factors that influence the classification performance. This asks for the utility of a feature selection method. Feature selection in gene expression data is also termed as gene selection [19]. The elimination of redundant and irrelevant features from the parent datasets greatly improves the performance of the classification model [20, 21]. A single feature selection technique may produce sub-optimal feature subset for which a training technique compromises on efficiency. Various feature subsets are combined in ensemble-based feature selection technique for the selection of an optimum subset of features using the combination of feature ranking that improves the classification accuracy [22]. The ensemble is divided into two phases. In the first phase, various feature selectors are used, providing a list of features in a sorted order while the second phase practice different aggregation techniques in order to cumulate the list of features [23]. The gene expression data can either be fully labeled, unlabeled, or partially labeled. This leads to the development of supervised, unsupervised or semi-supervised gene selection to discover the biological patterns and class prediction [24]. The supervised datasets contain samples along with their related features. These samples are tagged with a particular label that indicates the information related to the samples. In contrast, the unsupervised datasets contain samples without labels. The presence or absence of these labels in the datasets asks for semi-supervised learning.

2.2. Filter wrapper-based feature selection

Salem et al. [1] present a methodology for the classification of gene expression datasets. Their method, Information Gain (IG)/Standard Genetic Algorithm (SGA), use IG for feature selection first and then the features are reduced by applying the standard GA. They design a procedure for the relationship between the threshold kept for features selection and classifier. The author use a classifier to examine the IG threshold. Various thresholds are tested on the basis of which, only those features having an IG score greater than a predefined threshold value are preferred. Features with comparatively low IG values are eliminated resulting in dimensionality reduction. Afterwards, the classification of cancer gene expression datasets is carried out through Genetic Programming (GP). Pavithra et al. [25] present two types of feature selection technique for the classification of cancer gene expression datasets. The first approach is a filter method for feature selection that involves information gain for optimal feature selection. The other one is a wrapper-based technique. Later in their work, the decision tree (C4.5) is used as a classifier on the feature subsets. Dashtban et al. [26] present a two-phased methodology for the classification of oncogenes. They first reduce the features size by selecting statistically more relevant features by considering Laplacian and Fisher score for feature subset selection. Later, evolutionary approaches on the basis of random restart hill climbing, genetic algorithms, and reinforcement learning are applied for cancer gene classification.

Rouhi et al. [27] present a hybrid algorithm for feature selection of high dimensional microarray datasets. This methodology combines the filter method and meta-heuristic algorithm. This technique first uses the filter method for

reducing the feature's dimensions and then an advanced binary ant colony algorithm is applied on the already reduced subset for informative feature selection. The efficiency of the said work is evaluated using five high dimensional datasets. The quantity of selected features and classification error is used as an evaluation metric. Li et al. [28] design a two-step approach for the classification of DNA microarray datasets. The first step is based on a hybrid approach using Principal Component Analysis (PCA) and GA for optimal feature selection. Whereas, in the second stage, classification is done via Probabilistic Neural Network (PNN) classifier. The topology of PNN is optimized using GA. Experiments in their work are conducted using three different types of datasets.

2.3. Evolutionary computing techniques for gene expression classification

Evolutionary computing is inspired by natural phenomenon of various species for global optimization. These techniques use a particular searching method that is inspired by biological evolution, such as crossover, mutations, and selection operators. A few key technique include ant colony optimization, swarm intelligence, evolutionary algorithms, evolutionary programming, genetic algorithm, and genetic programming [29]. Garro et al. [30] present an evolutionary approach for the classification of gene expression data. The authors use Artificial Bee Colony (ABC) algorithm for reducing the number of genes as a first step. Afterwards, the classification is carried out through training different ANNs over the reduced feature subset. Their results show that ABC perform better for reduced features space as compared to other comparison methods. Authors in [31] present a methodology using Generalize Neuron (GN) for classification of DNA microarray. The methodology is divided into two stages. The first stage uses the ABC algorithm to select the subset of genes that are related to the disease. In the second stage, the obtained subset of features is used to train the generalized neuron with differential evolution. The differential evolution in their work aids in quicker convergence. Similarly, Ayyad et al. [32] introduce an optimization technique for the classification of gene expression data. Their work involves two separate approaches known as Local Mean-based *K*-Nearest Neighbor (LMKNN) method and Smallest Modified *K*-NN (SMKNN) that are used for the classification of high dimensional datasets. Both these techniques are developed on the basis of the basic *k*-NN approach. These are aimed to enhance the efficacy of the classification process. The LMKNN uses the largest circle between the center and the test item. Whereas, SMKNN utilizes the smallest circle between the center and test items. Ludwig et al. [33] present a fuzzy decision tree for the classification of gene expression datasets. Their work is compared to the classical decision tree approaches such as J48, NB, BN, Log, RBF, SMO, BG, RotF, and RanF.

2.4. Ensemble-based feature selection

The standard feature selection algorithm finds the local optimal feature subset in the candidate subset search space. However, the ensemble-based feature selection is superior and has more chances to select the best solution. The ensemble method may have more chances for a reliable outcome by aggregating the output of a number of base selectors [34]. Ghosh et al. [35] design a two-phased novel approach for the classification of cancer gene expression datasets. The first phase is an ensemble of three filter feature selection methods, Relief, chi-square, and symmetrical uncertainty. Union and interactions are used for the aggregation of top features out of the three filter methods. The obtained results from all filter method are combined into a subset and is provided as an input to the GA. Three classifiers, namely, *k*-NN, Multi-Layer Perceptron (MLP), and SVM are used to prove the independent nature of the developed methodology for a particular classifier. A comprehensive survey on various evolutionary computation-based techniques for feature selection can be seen in Xue et al. [36]. Their survey identifies that the GA and Particle Swarm Optimization (PSO) based methods are utilized in the past more frequently to select the optimum feature set for various tasks. The authors identify scalability, computational cost, and representation as a few major challenges for the evolutionary computing-based method to select the optimal feature selection. The work in [37] presents a new evolutionary computation-based approach for feature selection. Traffic sign recognition is adopted as a case study in their work. Their solution is named as Genetic-based Biological Algorithm (GBA). The hyperbolic tangent function is utilized in their work as a mapping technique for nonlinear adaptability. Experiments in their work are performed on German traffic sign recognition benchmark. Their work is compared with the conventional GA and the results suggest lesser computational resources requited by the GBA. The work in [38] presents an evolutionary computation-based solution for feature selection in high dimensional imbalance data. The overall goal of their work is to improve the classification accuracy by selecting the optimal feature set. Their method is named as Interaction Information based Evolutionary Feature subsets Selection (IIEFS) algorithm. Their solution uses interaction information to have higher-level interaction analysis to enhance the search process in the feature space. Likewise the present work, the proposal in [38] also has two phases. In the first stage, candidate features and their pairs are identified using traditional feature weighting methods. Whereas, in the second phase they are evaluated using multivariate interaction information. Classification experiments in their work are performed using three classifiers.

2.5. Metaheuristics for feature selection

In the domain of optimization the metaheuristics serves as a higher-level procedure to explore the available search space for finding a solution that satisfies all or maximum possible constraints. The metaheuristics has the ability to operate with imperfect information or limited computation capacity. In the past, multiple metaheuristic-based solutions have been presented for the feature selection task. A classic work on this can be seen in [39] where the author presents an overview of various feature selection techniques based on various metaheuristics. The author also proposes three metaheuristic strategies to solve the feature selection problem. In their experiments, the simplest version of the problem is considered to avoid overfitting issues. The work in [40] presents a study on feature selection from textual data for sentiment analysis using various metaheuristics. Their work identifies the potential of the GA to be utilized as a feature selection technique combined with measures like, information gain and Minimum

Table 1
Key features of the proposed work and related past contributions

Works	Feature selection	Evolutionary computation techniques	Classification	No. of cancer datasets
Garro et al. [30]	IG	GA	GP	7
Ayyad et al. [32]	IG	-	Modified *k*-nearest neighbor	6
Salem et al. [1]	IG	GA	Genetic programming	7
Uzma et al. [56]	Ensemble of 3 filter method	GA	SVM, k-NN, RF	6
Rani et al. [57]	MI	GA	SVM	3
Ghosh et al. [35]	Ensemble of 3 filter method, SU, Chi square, Relief	GA	MLP, SVM, *k*-NN	5
Current work	Ensemble of 3 filter method	LSFS + GA	SVM, *k*-NN	6

Redundancy Maximum Relevancy (mRMR). Their work also mentions the Ant Colony Optimization (ACO) as a possible solution having the fast convergence ability. The authors in [41] present a novel metaheuristic for the feature selection problem. Their solution is named as chaotic dragonfly algorithm. In their work 10 chaotic maps are employed to adjust the key parameters that control the dragonflies' movements. For the optimization task, they improve upon the basic Dragonfly Algorithm (DA). The experiments in their work suggest that the Gauss chaotic map significantly improves the performance of the DA for the feature selection task. Mafarja et al. [42] presents a hybrid metaheuristic approach for optimizing the selection of appropriate features. Their solution enhances the basic Grey Wolf Optimizer (GWO) and Whale Optimization Algorithm (WOA). Their solution is evaluated on 18 benchmark datasets. Their approach improves the variants that can alleviate the stagnation problems. The authors in [43] present a novel wrapper feature selection algorithm based on iterated greedy metaheuristic. The problem of sentiment classification is considered as a case study in their work. They also introduce a selection procedure that uses pre-computed filter scores for the greedy construction part of the iterated greedy algorithm. For classification performance measurement, multinomial naïve Bayes classifier is utilized in their work. Shukla et al. [44] presents a hybrid metaheuristic method for the gene selection task from the gene expression datasets. Their solution is named as Teaching Learning-Based Gravitational Search Algorithm (TLBOGSA). Their work also incorporates a newer encoding strategy. For the classification accuracy computation, naive Bayes classifier is utilized as a fitness function. Before applying their solution, mRMR is employed to reduce the initial features. This helps in reducing the search spaces as well [45-54]

Xue et al. [45] present an algorithm called NSGA-III based on three objectives for feature selection. Their solution selects reliable features from an incomplete datasets. They construct the missing information using average imputation approach. Each feature is assigned a probability that represents either it being selected or rejected. Afterward, the *k*-NN classifier is used to assess the selected features. Their approach is evaluated by comparing it with four past methods on six incomplete UCI datasets. Xue et al. [46] propose a novel algorithm called SaPSO for large-scale feature selection. It represents the solution into a binary string where each feature's value is compared to a threshold. If it is greater or less, then the corresponding feature value in the solution is set to be 1 or 0, respectively. Experiments on 12 datasets show that their solution reduces feature set by 70% to 80% as compared to the evolutionary computation method. It also provides better results concerning training and test data sets. The authors in [47] design a multi-objective-based algorithm called HMPSOFS for cost-based feature selection. The two operators are combined with the PSO to enhance its performance.

Uzer et al. [48] develop a hybrid approach. They use the Artificial Bee Colony (ABC) algorithm to select the feature subset and then apply SVM for the classification of samples based on the selected feature subset. The ABC algorithm use clustering as an objective function for evaluating the solution. The performance of their work is evaluated through four sets of medical data from the UCI database. The authors in [49] propose a modified version of firefly algorithm to select important features from the massive datasets generated by the intrusion detection system. They use *k*-NN method and addition of extra feature selection to improve the traditional firefly algorithm. The efficiency of their solution is measured using four datasets related to different kinds of attacks. Guan et al. [50] design the search-History-Guided Differential Evolution (HGDE) to select features from large-scale datasets. The HGDE use BSP tree to remember the search history. HGDE is evaluated by comparing it to five algorithms using synthetic data sets. Zhang et al. [51] came up with a new algorithm called MOFS-BDE for feature selection. Three new operators are set up and integrated into MOFS-BDE. These operators increase the algorithm's performance by enhancing the self-learning ability, the convergence of the algorithm, and reducing the computational complexity of the algorithm. Their suggested algorithm is compared with four popular techniques based on 20 datasets. Song et al. [52] develop a new algorithm named VS-CCPSO to select essential features from large-scale data. It first uses SU to find the important features and then divide the search space into low dimensional space using the divide and conquer approach. Then the PSO algorithm is used on each search space to find the optimal subset of features. In their method, the optimal selection of a subset of features depends on a single SU Filter method. However, different filters have varying criteria for selecting the relevant features. Song et al. [53] develop a new algorithm called BBPSO that integrates mutual information to select important features. It combines the filter with the wrapper method. They first find the correlation between the features and the label using mutual information. Then, the PSO-based wrapper method is employed with two newly developed operators. Afterwards, the *k*-NN classifier is applied to categorize the samples according to the selected subset of features. Their solution is compared with eleven algorithms on sixteen datasets. Song et al. [54] designed a three-

step-based hybrid feature selection algorithm called HFS-C-P. First the SU is applied to remove the irrelevant feature; next the clusters are formed on the selected relevant features, after that, the PSO is used for optimization. For the verification of their proposed idea, *k*-NN classifier is used.

2.6. Limitations of the past works addressed in the current proposal

The authors in [1] use the filter-based method of IG for feature selection. They define a threshold list for feature selection and identified the threshold that yields great accuracy for feature subset. The feature subset for various training sets is different, delivering better classification accuracy. However, for stability, the feature subset must be common for a variance of the training set. Therefore, the problem of stability is addressed in the current work by using ensemble filter feature selection approach. The authors in [32,51,53,54] use single filter method for feature selection. Each feature is ranked separately independent of its connection with other attributes. However, a single filter method is not suitable for feature selection of gene expression datasets because of the presence of redundant and irrelevant features. The current work addresses this issue by combining the Relief,SU, and chi square (X^2) feature selection techniques. The work in [35] use ensemble filter feature selection technique followed by the GA for the optimization of feature subsets. Their ensemble of multiple filters may generate noise [55] due to the combination of multiple filter approaches. The proposed methodology in the current work reduces this problem by designing a novel heuristic based on IG for removing the noise.

DNA microarray measures the expression level for a number of genes simultaneously. The unique characteristics of gene expression data makes it more challenging to be analyzed for the prediction of cancer, types of tumors, and many other diseases. The problem of classification of cancer gene expression data has a set of challenges because of its unique behavior. Some of the main challenges in the classification of gene expression datasets involve large dimensions of the features while the relevant features are comparatively quite less. Another challenge is the presence of noise in the datasets collected from multiple sources. These factors influence the performance and quality of the gene expression data analysis. The current work aims to address these challenges. Table 1 lists the key features of the proposed work and related past contributions.

The proposed approach is compared with the five closely related methods that focus on the feature selection techniques for the cancerous gene expression data. These algorithms first use the feature selection techniques to select the appropriate feature subset and afterwards they apply standard classifiers for cancerous sample prediction. These methods include; Ayyad et al. [32], Salem et al. [1], Uzma et al. [56], Rani et al. [57], and Ghosh et al. [35]. Two of these algorithms, i.e., [32] and [1] use IG as a single filter method for the feature selection during their first stage. Whereas, the methods in [56] and [57] use the ensemble of three filter methods and Ghosh et al. [35] utilize the MI method for the feature selection during its first stage. All these five algorithms used here to perform comparison utilize GA in their second stage with an exception of [32]. Key features of these methods are mentioned in Table 1.

3. Proposed solution

This section presents the proposed solution. It starts with an ensemble learning-based feature selection, followed by three filter-based feature selection methods, i.e., symmetrical uncertainty-based, chi square-based, and relief-based approaches. Next, the proposed Local Search-based Feature Selection (LSFS) algorithm is explained.

The proposed method has two phases of feature selection techniques (filter with wrapper techniques) for the classification of cancerous gene expression data. The first phase of the ensemble of the three filter methods uses SU, Relief, and Chi square (X^2). Whereas, the second phase uses the GA for the optimization of the feature subsets generated from the first phase. The GA uses the LSFS for generating the initial solution and removing the noise generated from the first phase. The population is generated randomly. The gene expression datasets suffer from the curse of dimensionality that is a major reason of instability because of a small number of samples and a much larger number of genes. However, the stability of the proposed methodology is handled using multiple (i.e., three) methods. Firstly, the ensemble filter methods are used in the first stage to stabilize the process. Secondly, the stochastic-based feature selection algorithm such as GA uses the local search algorithm (LSFS) to avoid the random seed which is the cause of instability. Thirdly, 5-fold cross-validation is used to improve

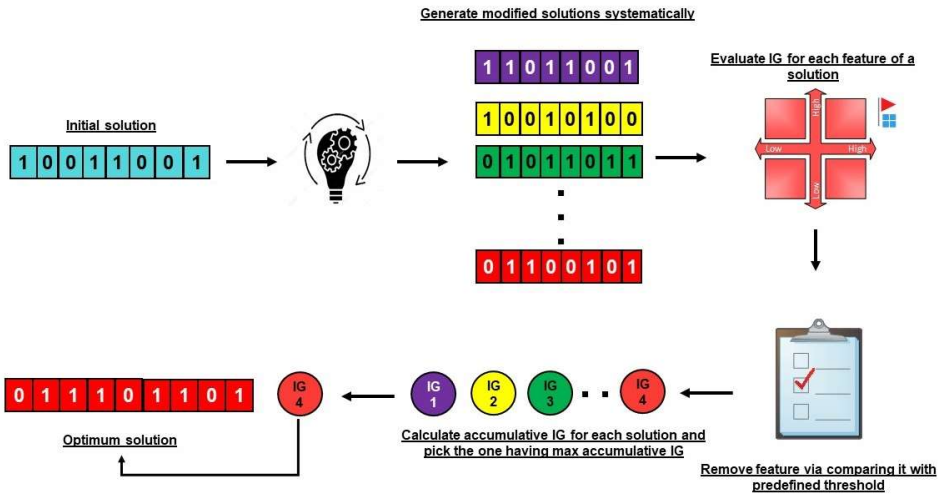

Fig. 1. Feature subset based local search

the stability against the various training datasets. Moreover, redundant and irrelevant features reduce the performance of the learning model. Therefore, feature selection plays a vital role in designing an effective learning model. For various reasons, the feature selection is valuable for classifications, for instance, (1) it trains the learning model faster, even if a costly algorithm is employed, (2) it ensures the model is generalized and reduces the overfitting problem, and (3) it also makes the model conveniently interpretable. Furthermore, in the literature, the GA has been used for the optimization problems. Therefore, the proposed idea here uses an evolutionary algorithm for the optimizations of the feature selection of the gene expression data. Because of the high dimensional gene expression data, the conventional optimization methods cannot efficiently solve the feature selection problem. Hence, the GA has been adopted to compensate it.

3.1. Ensembles learning-based feature selection

Data science techniques play an important role in analyzing the data that is generated from different source. However, increasing size of datasets influence the learning model in terms of both the training and execution time. The feature selection technique removes irrelevant and redundant features while retaining the useful information. Feature selection benefits in terms of speeding up the data mining algorithm, improves the classification performance and understanding the problem while dealing with the most relevant features [58]. The preprocessing step of the classification is feature selection while the goal of feature selection is to increase the classification accuracy. Recently, the stability of feature selection is considered an important issue [59]. Ensemble learning-based feature selection is a newer type of feature selection method [60]. To improve the stability of feature selection, ensemble learning-based feature selection is designed. In this approach diverse feature subset is generated after applying various feature selection techniques. Finally, these subsets are aggregated into a single feature subset.

3.2. Preliminaries

The proposed solution utilizes three filter-based methods. These are explained in the following before going into the details of the proposed solution.

Correlation-based feature selection (CFS): CFS belongs to the class of filter algorithm. It ranks the attributes which are based on the evaluation functioning via the concept of correlation. The function evaluates the feature subset to locate where the features are uncorrelated with each other and correlated with the class label. The CFS removes irrelevant features based on their low correlation with the class. The rest of the features usually have a strong correlation with the class. The subset of u features evaluation function of CFS is expressed in Eq. (1).

$$F_s = \frac{u \quad \overline{v_{ca}}}{\sqrt{u + u(k-1)\overline{v_{aa}}}} \tag{1}$$

where, F_s is the feature subset evaluation criteria, u is the number of attributes in the subset and $\overline{v_{ca}}$ represents the average correlation between the class and attribute. The $\overline{v_{aa}}$ represent the correlation between two features.

Symmetrical Uncertainty (SU) is a variation of Correlation-based Feature Selection (CFS) used as the first measures.

Chi-square: In statistics, the chi-square (X^2) is a test of independence that determines the significant difference between the variables. It calculates the dependence between variables (features) and class.

The X^2 determines the relationship between the feature variable and a class. The feature is discarded if two variables are independent. The formula for X^2 is shown in Eq. (2).

$$X^2 = \sum_{i=1}^{G} \sum_{j=1}^{m} \frac{(S_{ij} - E_{ij})^2}{E_{ij}} \tag{2}$$

The number of samples and expected frequency is represented by S_{ij} and E_{ij} respectively within an interval j belonging to the class i.

Relief algorithm: The Relief algorithm is known for fast implementation that deals accurately with dependent features and noisy data. The algorithm assigns a weight to each feature to show its significance. The weight is calculated by finding two nearest neighbors of randomly chosen sample: one is taken from the same class (called nearest hit) and the other one is taken from the opposite class known as the nearest miss. After assigning the weight to each feature, the top N features based on a particular threshold are selected. Relief is designed for two class problem. It has been modified to expand its behavior. The modification is done by incorporating two significant ideas. The first one is that Relief searches n nearest neighbors making it less sensitive towards noise (it also manages the missing data). Second, it examines the multiclass problem and assigns normalized weight to the features based on their probability of the class. Where, *Diff* in Relief is computed using Eq. (3).

$$Diff^{(i,X,nI)} = V_{alue}(i,x) - V_{alue}(i,nI) \tag{3}$$

Information Gain: In information theory, the Information Gain (IG) calculates the difference between two probability distributions. It measures the quantity of the gained information by the feature with respect to the class. However, the irrelevant feature should be given no information. The IG can be calculated using Eq. (4).

$$\mathrm{IG}(S, a) = \mathrm{H}(S) - \mathrm{H}(S \mid a) \tag{4}$$

where, IG is the information gain for the data in S for the variable a, H(S) is the entropy of the data before change, and H($S \mid a$) is the conditional entropy for the data given variable a [9].

3.3. Local Search-based Feature Selection

The proposed novel local search algorithm called the Local Search-based Feature Selection (LSFS) utilizes the 2-opt operator. The assemble filter method generates noise due to the combination of various subsets of features generated using multiple filter methods. Therefore, it is possible to select an irrelevant feature, which may reduce the performance of the classification. Consequently, the proposed heuristic. i.e., LSFS is used which employed the 2-opt operator. It utilizes a systematic method to remove the irrelevant features or select the relevant feature based on IG value. Therefore, after the first phase of the proposed solution gets executed, this serves as a filter to select the most relevant features. LSFS is designed to optimize the solution for selecting optimum features as presented in Algorithm-1. The proposed heuristic uses IG for identifying the important features and cumulative IG value is

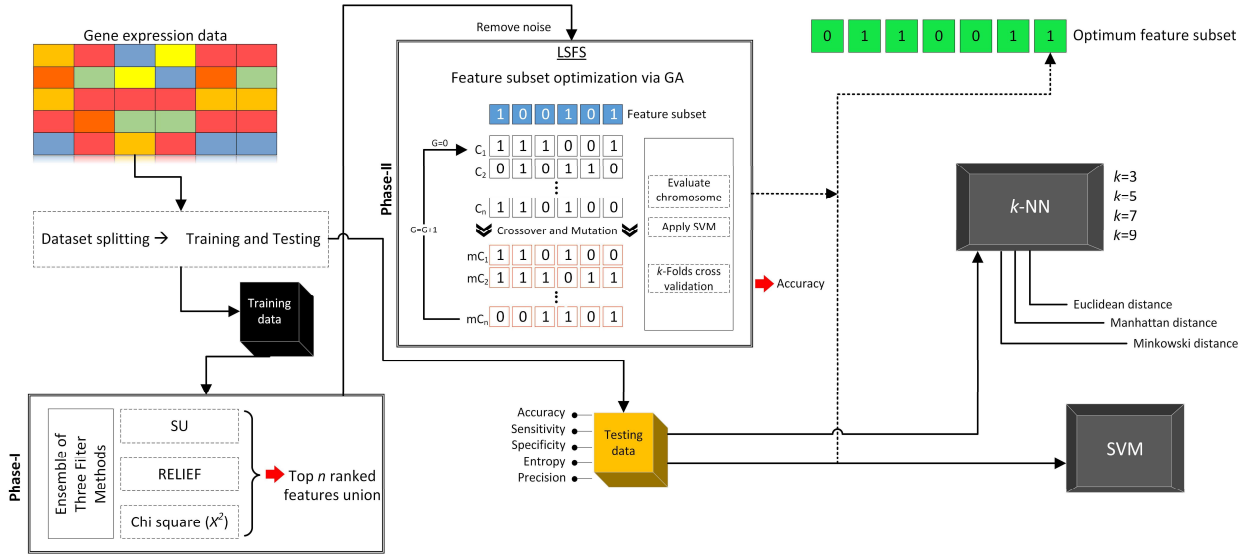

Fig. 2. Overall working of the proposed

used for the solution evaluation. The algorithm takes the initial solution (the feature subset generated from the ensemble filter-based feature selection) as an input. The initial solution is a binary string where 1 and 0 show the selected and unselected features, respectively. Later it applies operations to modify the initial solution for searching the optimum points as shown in Fig. 1. The solution is modified in terms of adding or removing a feature based on their IG value. First, the parameter *s_rate* is defined for the LSFS which shows the number of modified solutions generated from the initial solution. The algorithm initially selects two positions randomly in the initial solution. Say, *i* and *j*, it then selects the subset between these two positions and reverses them in the initial solution as shown in step 8 of the Algorithm-1. After this, it calculates the IG value for each feature of the modified solution in step 10. The calculated IG value is compared with the predefined threshold in step 11. If the IG value of a feature is greater than the predefined threshold, it is selected, otherwise that feature is discarded. In step 12, the accumulative IG value is stored for each modified solution. The accumulative IG value is the summation of the IG value of the individual feature in the solution. The new solution is selected in step 18 which is generated via systematically having maximum accumulative IG value. The proposed method uses the evolutionary algorithm for the optimizations of the feature selection of gene expression data. Because of the high dimensional gene expression data, the conventional optimization method cannot efficiently solve the feature selection problem. Hence, the GA has been adopted to compensate.

The proposed work is a novel filter-wrapper based method. Various filter methods are available for feature selection. Each filter-based method assigns different ranking to the same feature; therefore, utility of a single selection method is not recommended for feature selection tasks. To overcome the challenge in features selection of gene expression datasets, the proposed solution has adopted a two-phased methodology. The first phase uses an ensemble-based filter approach. This phase combines the information from two filter methods, SU, chi square, and Relief-based feature selection. The reason of using these filter methods is to select the most relevant features from the gene expression datasets to classify the cancerous samples. The classification problem of cancer gene expression data has a set of challenges because of its unique behavior. The internal view of the gene expression dataset makes it challenging to process. The huge dimension of gene expression data contains noise, redundancy, and irrelevant items that make it difficult to analyze. The purpose of using SU is to remove irrelevant features. It ignores the features that are independent of each other but have low correlation with the class. The Relief deals with the noise and redundant features. The chi square is used to select the features that are highly dependent on the class label. This phase combines the essential features given by the two approaches. If one approach ignores the essential features, there is a chance that the other would have picked it. The feature subset selected using the ensemble method is more robust. The top_n ranked features are picked from both methods and combined together. The union is done through an aggregation function for combining the top_n ranked features. The final subset earned by union aggregation is passed to the second stage. The second stage is used for the global optimization of the first stage. This phase uses the combination of a LSFS algorithm and a meta-heuristic algorithm called the GA for optimization. The feature subset from the first phase has chances of having noise [55] due to the combination of relevant and non-redundant features. This noise limits the classification performance. Therefore, LSFS is applied to remove the noise from the feature subset. The proposed methodology in the current work reduce this problem by designing a novel heuristic based on IG for removing the noise. Therefore, the major novelty of the present work lies in the second phase. It presents a novel heuristic approach named as LSFS that is used with genetic algorithm for controlling the randomness and informative feature selection. Hence, the second stage is used for the global optimization of the first stage. This phase uses the combination of a feature subset-based local search algorithm (LSFS) and a meta-heuristic algorithm called the GA for optimization.

Input: Initial solution
Output: Modified solution M_S (in a systematic way)

```
1. S←generate initial solution
2. Repeat for the defined number of solution (s_rate)
3.     F_s←∅
4.           cum_ig ←∅
5.           s_rate ←n
6.      for k=0 to s_rate do
7.         Randomly select two position i and j & Reverse the subset between two pos.
8.                        R_s ← Reverse(s, i, j)
9.                        for j=0 to len(R_s) do
10.                           Calculate IG_v of feature j
11.                           if IG_v> threshold
12.                                    cum_ig ←cum_ig U<j , IG_v>
13.                           end if
14.                       end for
15.                       S_v =sum (cum_ig)
16.          F_S ←F_S U<k, S_v>
17.           end repeat
18.     M_S ← MAX(F_S)
19.   return  M_S
```

Algorithm-1. Local Search-based Feature Selection

The local search algorithm uses the obtained feature subset of the first phase as an input. The LSFS uses a systematic approach for the optimization of feature subset. Next, the solution of the local search algorithm is passed to the metaheuristic GA. The GA uses the solution obtained from the LSFS as an initial solution, instead of creating a random initial solution. The final feature subset obtained via GA is used for classification. The proposed model uses SVM, *k*-NN, and RF classifiers to show the independence of the classifier and the proposed model. Once the optimal set of features is extracted using the proposed approach, various experiments using the SVM and *k*-NN classifiers have been performed. The choice of these classifiers is made based on their superior performance is such tasks in the past [61] and also their suitability for the binary classification problems [62]. The SVM classifier has been reported to perform better for the binary classification problem provided it gets a suitable feature set, therefore it has been utilized in this work. Additionally, the SVM is effective in high dimensional spaces, making it suitable for the present case study. Another advantage of SVM is it being memory efficient. The proposed approach being an evolutionary computing-based method consumes more memory, therefore integrating SVM into the solution helps in the conservation of the same. The workflow of the proposed model is shown in Fig. 2.

A GA is used in this work for further optimizing the so far selected features. Collection of chromosomes forms a population. An individual in the GA population is represented by a binary string. Where, 1s and 0s at the positions *i* and *j* denote the selected and dropped features at i^{th} and j^{th} slot, respectively. The population is generated randomly. The solution obtained via the LSFS algorithm is to act as an initial solution to the GA. The initial solution is to represent a binary string. Each cell of the chromosomes shows the position of the selected or dropped feature. The binary string consisting of 1s and 0s indicate the selected and dropped features due to LSFS algorithm (see Fig. 3).

Once the initial solution for the genetic algorithm is selected, it signals for the population to be created randomly. Therefore, for the population of *m* chromosomes, *m* lists of order of some length *n* are generated. Next, swaps are applied to the initial solution, where *n* swaps are applied on each of the initial solution to form a final chromosome. The same process is repeated to generate all *m* chromosomes (individuals in the population).

Reproduction operators: The genetic operator of uniform crossover is used in the current work. This uniform crossover combines both the genes of chromosome (X and Y) to maintain the uniformity. This process treats each gene separately, by generating a random number (0 or 1) which decides either the gene is selected from the first or second parent is transferred into the offspring. For example, if the random number is 1, it means that the offspring selects a gene from the second parent otherwise it is selected from the first one. The same process is repeated for selecting genes of the offspring. The process of uniform crossover is represented in Fig. 3.

The proposed work use the swap mutation as a genetic operator. The swap mutation randomly selects two genes in the chromosome and interchanges their position. A chromosome C_1

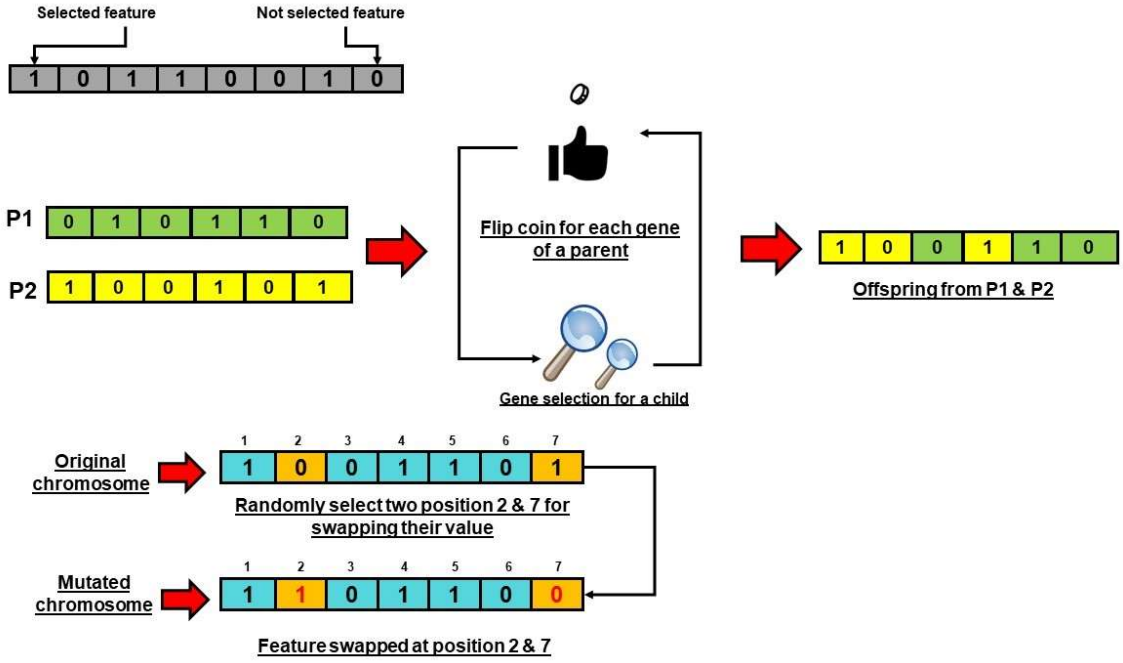

Fig. 3. Reproduction operators

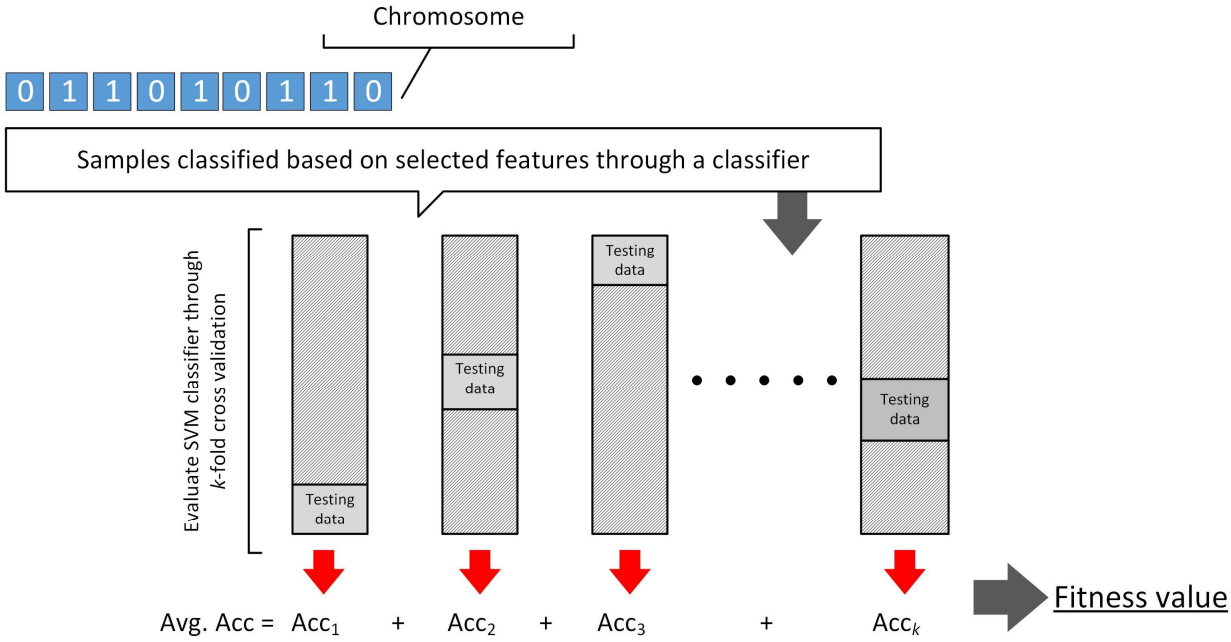

Fig. 4. Fitness value evaluation

of binary string having length n, selects two random positions followed by swapping the position of randomly selected two genes. For example, consider a chromosome C_1 of length 7. The algorithm selects two positions, i.e., 2 and 7 randomly and interchanges their elements. The resultant is a new chromosome having the element at position 3 swapped with an element at position 5 as illustrated in Fig. 3.

Fitness function: A chromosome shows a feature subset in the form of a binary string. A classifier is used to assess the fitness value of the chromosome. The proposed work uses various classifiers for the fitness value evaluation of the chromosomes. Each chromosome represents a feature subset S_f. The dataset d is extracted from the original data set D on the basis of S_f as given in Eq. (5). Then the SVM classifier is executed. The performance of the classifiers is evaluated using n -fold cross-validation. The average accuracy of the classifier represents the fitness value of the chromosome, as shown in Eq. (6), which determines whether the chromosome is fit for the next generation, where the diagrammatic representation of the fitness value evaluation is given in Fig. 4.

$$d = S_f(D) \tag{5}$$

$$F_{value} = \frac{1}{n}\sum_{i=1}^{n} SVM_{(d)} \tag{6}$$

The suggested method uses 10-folds cross-validation for accessing the performance of the classifier. The simulation of the proposed model is repeated for an average of 10 runs. Overall simulations performed for each dataset is 100.

This work presents a novel filter-wrapper-based method for feature selection. It is based on the two-phased feature selection approach (i.e., filter with wrapper technique) for the classification of cancer gene expression data. The purpose of using filter with wrapper method is to overcome its limitations. Therefore, the first phase of the ensemble here uses SU, Relief, and chi square (X^2). Whereas, the second phase uses the GA for the optimization of the feature subsets generated from the first phase. The GA utilizes LSFS for generating the initial solution and for removing the noise generated in the first phase. The gene expression datasets suffer from the curse of dimensionality. This is a major reason of instability because of the smaller number of samples and a large number of genes. However, the stability of the proposed methodology is handled with three methods. Firstly, the ensemble filter methods are used in the first stage to stabilize the process. Secondly, the stochastic feature selection algorithm, such as, GA uses the local search algorithm (LSFS) to avoid the random seed which is the reason of instability. Thirdly, the 5-fold cross-validation is used to improve the stability against various training data.
The proposed approach takes into account those attributes having IG value greater than a given threshold. Each attribute has same evaluation time that is calculated by the IG method. This makes the evaluation factor constant, i.e., *O(1)*. The time complexity of the proposed approach mainly depends on three factos, where n denotes the number of samples, f denotes the dimension in the data, and p represents the population size. The evolutionary algorithm will iterate over the population g times. Thus, the proposed approach's overall worst-case time complexity becomes $O(n \times f \times p \times g)$, where n denotes the number of samples, f denotes the dimension in the data, and p denotes the population size.

4. Experiments and results

This section contains the experiments conducted using the proposed model. For this, six benchmark cancer gene expression datasets are utilized. The results of the proposed model are compared with three state-of-the-art algorithms for evaluation purposes. Various types of experiments are conducted using two classifiers, considering various fitness

functions, a number of variation of *k* for the *k*-NN and distance measures.

4.1. Performance metrics

The performance metrics utilized for reporting results include: accuracy, sensitivity, specificity, precision, F-measures and entropy. Prerequisite to these metrics is the computation of the confusion matrix. Each row of the matrix shows the observations of the predicted class and each column visualizes the actual value. On the basis of event observation the positive class represents the positive events and vice versa. The term True Positive (TP) means that observation is positive along with a positive predication made by the classification model. The term False Negative (FN) signifies that the prediction is negative, but the observation still remains positive. The term False Positive (FP) means that the observation is negative, but the model prediction is positive. The negative observation and prediction is indicated by the term True Negative (TN).

Accuracy: Accuracy is the measure to evaluate a classifier's performance. It is defined as a ratio of correctly identified observations to the total observations. The higher the accuracy, better are the results. Its value ranges from 0 (worst) to 1 (best). Eq. (7) shows the computation of accuracy. Where, TP shows positive instances predicted as positive, FP indicates negative instances predicted as positive, FN are the positive instances predicted as negative, and TN represent negative instances predicted as negative.

$$Accuracy = \frac{(TP+T\ )}{(TP+FP+FN+T\ )} \tag{7}$$

Table 2
Datasets detail and Parameter settings

Name	Samples	Features	Per class instances
Prostate	136	12600	Class-1=77, Class-2=59
Lung cancer	181	12534	Class-1=31, Class-2=150
Leukemia	72	3571	Class-1=47, Class-2=25
Central nervous system	61	7129	Class-1=21, Class-2=40
Colon cancer	62	2000	Class-1=40, Class-2=22
DBLCL	77	7070	Class-1=58, Class-2=19
Parameter settings		Population size	100
		Mutation rate	10%
		Crossover rate	50
		Fitness function	SVM linear
		s_rate	20
		Reproduction operations	Uniform crossover & swap mutations

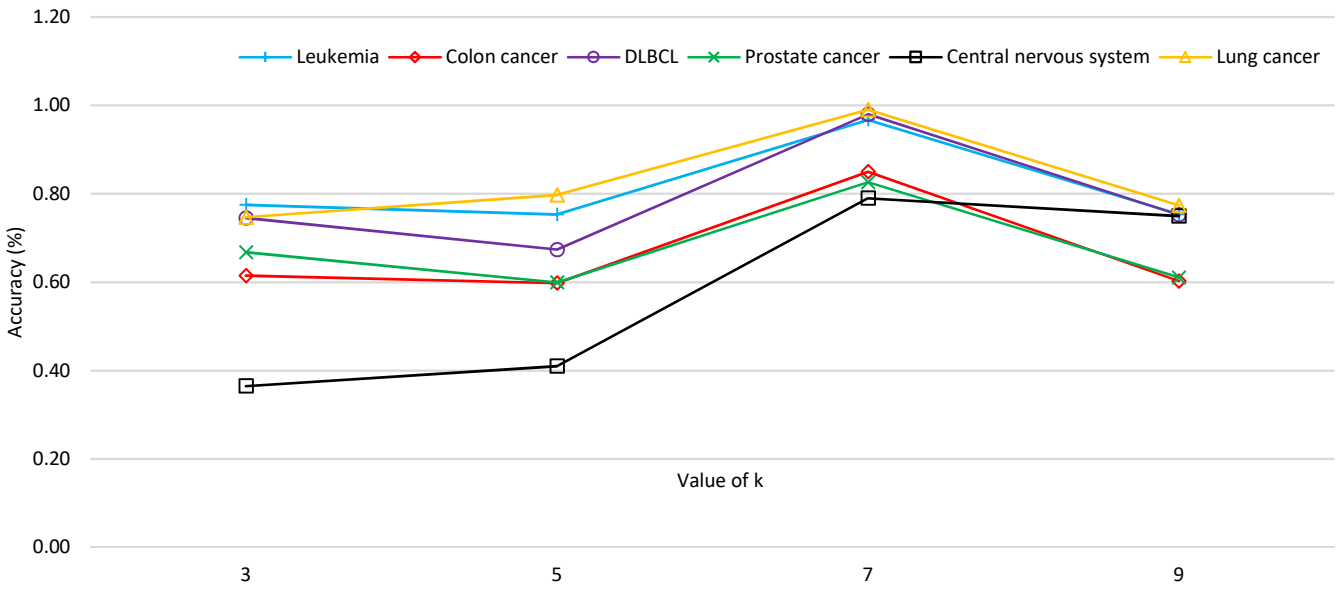

Fig. 5. Accuracies obtained for four values of *k*

Table 3
Results for SVM plugged into the proposed framework

Dataset	Linear kernel						Sigmoid kernel					
	Accuracy	Sensitivity	Specificity	Precision	F measure	Entropy	Accuracy	Sensitivity	Specificity	Precision	F measure	Entropy
Leukemia	0.992	**0.996**	0.981	0.993	**0.994**	0.006	0.984	0.985	0.976	0.993	**0.988**	0.006
Colon cancer	0.901	0.925	0.858	0.926	0.925	**0.071**	0.817	0.136	0.826	0.016	0.029	0.066
DLBCL	0.981	0.996	0.929	0.979	0.987	0.020	0.774	0.774	0.690	0.999	0.872	0.001
Prostate cancer	0.947	0.948	0.948	0.958	0.950	0.040	0.603	0.627	0.832	0.900	0.739	**0.094**
Central nervous system	0.989	0.998	0.968	**0.988**	0.990	0.011	0.719	0.719	0.732	**1.000**	0.830	0.000
Lung cancer	**0.994**	0.986	**0.995**	0.979	0.982	0.020	**0.994**	**0.986**	**0.995**	0.979	0.982	0.020
	Polynomial kernel						RBF kernel					
Leukemia	0.678	0.672	0.683	0.953	0.780	0.045	0.979	0.984	0.973	0.986	**0.984**	0.013
Colon cancer	0.838	0.864	0.807	0.890	0.876	**0.104**	0.664	0.665	0.950	0.998	0.798	0.001
DLBCL	0.748	0.748	0.702	**1.000**	0.855	0.000	0.751	0.751	0.723	**1.000**	0.857	0.000
Prostate cancer	0.927	0.941	0.908	0.935	0.937	0.062	0.591	0.585	0.614	0.480	0.104	0.000
Central nervous system	0.875	0.891	0.866	0.915	0.900	0.080	**0.981**	**0.985**	**0.975**	0.988	0.986	0.011
Lung cancer	**0.998**	**0.991**	**0.999**	0.996	**0.990**	0.003	0.805	0.799	0.805	0.800	0.799	**0.178**

Table 4
Results of the proposed solution by plugging-in *k*-NN and Linear kernel as a fitness function

Dataset	Accuracy	Sensitivity	Specificity	Precision	F measure	Entropy
	Euclidean distance					
Leukemia	0.97	0.97	0.97	**0.98**	**0.97**	0.02
Colon cancer	0.82	0.83	0.77	0.92	0.87	0.08
DLBCL	0.90	0.98	0.78	0.87	0.92	0.12
Prostate cancer	0.83	0.86	0.80	0.85	0.85	**0.14**
Central nervous system	0.79	0.78	0.85	0.95	0.85	0.05
Lung cancer	**0.99**	**0.99**	**0.99**	0.96	**0.97**	0.04
	Minkowski distance					
Leukemia	0.94	0.94	0.96	0.98	0.95	0.02
Colon cancer	1.00	0.58	**1.00**	**1.00**	0.73	0.00
DLBCL	0.88	0.93	0.76	0.91	0.92	0.09
Prostate cancer	0.83	0.89	0.76	0.80	0.84	**0.18**
Central nervous system	0.70	0.80	0.46	0.81	0.80	0.17
Lung cancer	**0.99**	**1.00**	0.99	0.95	**0.97**	0.05
	Manhattan distance					
Leukemia	**0.97**	**0.96**	**0.97**	**0.98**	**0.96**	0.02
Colon cancer	0.87	0.79	0.87	0.93	0.85	0.07
DLBCL	0.91	0.84	0.91	0.89	0.86	0.10
Prostate cancer	0.81	0.84	0.78	0.82	0.83	0.16
Central nervous system	0.65	0.73	0.44	0.79	0.76	**0.19**
Lung cancer	0.93	0.90	0.92	0.81	0.85	0.17

Sensitivity: Sensitivity (also called recall) (Eq. (8)) is the ratio of correct positive predictions and the total number of positive samples. The highest value of sensitivity is 1 and the lowest, i.e., worst is 0. At times, it is also called True Positive Rate (TPR) or recall.

$$Sensitivity = \frac{TP}{(TP+FN)} \quad (8)$$

Specificity: Specificity (Eq. (9)) is defined as the number of correct negative predictions divided by the total number of negatives. Specificity is also known as True Negative Rate. (TNR). The ideal value of specificity is 1, whereas, the worst value is 0.

$$Specificity = \frac{TN}{(TN+FP)} \quad (9)$$

Precision: Precision, also known as Positive Predictive Value (PPV), is defined as the ratio of total correct positive predictions and all positive predictions. The best value of specificity is 1 and its worst value is 0. Eq. (10) shows the computation of precision.

$$Precision = \frac{TP}{(TP+FP)} \quad (10)$$

F-measures: The F-measures is the harmonic mean of precision and recall. It is sometimes called F-scores and mathematically as shown in Eq. (11).

$$\text{F} - \text{measures} = 2 * \frac{(\text{precision} * \text{recall})}{(\text{precision} + \text{rec} \quad)} \quad (11)$$

The good F-measures value is represented by low values of FP and FN has low value. The best value for this measure is 1, and the worst value is 0 in case of the lowest precision and recall value.

Entropy: Entropy is a metric that measures the uncertainty or disorder of the target class. It is mathematically shown in Eq. (12).

$$\text{Entropy} \ = \sum_{j=1}^{c} - p_j \, log_2 p_j \quad (12)$$

where, c represent the number of classes, and p_j is the probability of the class j. The value of entropy lies between 0 and 1. The 1 shows the high level of uncertainty means low level of purity of the class distribution.

Table 5
Results of the proposed solution by plugging-in *k*-NN and polynomial kernel as a fitness function

Dataset	Accuracy	Sensitivity	Specificity	Precision	F measure	Entropy
			Euclidean distance			
Leukemia	0.91	0.91	0.96	**0.98**	0.94	0.01
Colon cancer	0.83	0.81	0.92	0.97	0.88	0.02
DLBCL	0.84	0.98	0.67	0.79	0.87	**0.19**
Prostate cancer	0.81	0.83	0.79	0.87	0.84	0.12
Central nervous system	0.67	0.74	0.53	0.79	0.76	0.18
Lung cancer	**0.99**	**1.00**	**0.98**	0.92	**0.95**	0.08
			Minkowski distance			
Leukemia	0.89	0.90	0.88	0.93	0.91	0.07
Colon cancer	0.78	0.75	0.85	0.93	0.83	0.07
DLBCL	0.94	0.98	0.84	0.95	0.96	0.05
Prostate cancer	0.82	0.86	0.77	0.82	0.84	0.16
Central nervous system	0.61	0.70	0.50	0.76	0.73	**0.21**
Lung cancer	**0.99**	**1.00**	**0.99**	**0.96**	**0.98**	0.03
			Manhattan distance			
Leukemia	0.92	0.92	0.92	0.96	**0.94**	0.04
Colon cancer	0.79	0.79	0.83	0.91	0.84	0.09
DLBCL	0.90	0.99	0.77	0.88	0.93	0.11
Prostate cancer	0.87	0.92	0.81	0.85	0.88	0.14
Central nervous system	0.61	0.70	0.50	0.76	0.73	**0.21**
Lung cancer	**0.99**	**1.00**	**0.98**	**0.93**	0.96	0.07

4.2. Datasets and parameters

The proposed model is evaluated using six benchmark datasets. These two classed datasets are leukemia, colon cancer, lung cancer, central nervous system, Diffuse Large B Cell Lymphoma (DLBCL), and prostate cancer. The detail of these datasets is listed in Table 2. The leukemia dataset contains 72 samples and 3571 features. There are 47 samples belonging to the class of Acute Lymphoblastic Leukemia (ALL) the remaining 25 belongs to the class of Acute Myeloid Leukemia (AML). The total number of samples for Colon cancer are 62 and the number of genes is 2000. The number of samples associated with cancer and normal class is 40 and 22, respectively. Next, the gene expression of lung cancer has 181 samples and 12534 is its features dimension. The information about the number of samples relate to two classes is such as; Mesothelioma having 31 number of samples and the remaining 150 belongs to the class of ADCA. The prostate cancer has 136 samples, where 77 belong to the tumor, and 59 samples belong to normal class. The number of features in prostate cancer dataset are 12600. Center nervous system dataset has 61 samples. Where, 21 belong to class-1 and remaining 40 belong to class-2. The total number of features in this dataset are 7129. The gene expression dataset of DBLCL contains 77 samples having 7070 features. The two classes of DBLCL have 58 and 19 samples each.

To select the parameters for the *k*-NN classifiers various experiments are carry out. These experiments are based on the multiple values of *k* and the distance measures. The variable *k* show the number of selected nearest neighbors for the *k*-NN. The *k*-NN classifier is evaluated for the combination of five values of *k*, i.e., 3, 5, 7, and 9 with distance measures setting as Euclidean, Minkowski, and Manhattan. These experiments suggest that the proposed methodology performs better for sitting the value of *k* at 7 and Euclidean as a distance measure for the *k*-NN classifiers.

Whereas, the key parameters for the SVM is penalty factor C. The value of C effects the complexity for the SVM classification model and the outcome of the feature selection. The proposed model use linear SVM for C having value 1 as a classifier and 5-fold cross-validation is used to select the best value. The average value of 10 run is reported by executing it on each dataset. Where, the key parameters for the SVM are penalty factor C. The value of C affects the complexity of the SVM classification model and the outcome of the feature selection. The parameters for the GA used in the proposed work are tuned based on trial and error process to select suitable parameters. To select the suitable value of population size (P), various population size such as 20, 40, 60, 80, 100, 120, 140, and 160 are set. The performance of the proposed work changes from 20 to 100 population size. However, after that, i.e., 100 population size, the performance remains almost constant. Similarly, for the mutation rate various values are evaluated, i.e., 2 %, 4%, 8%, 10%, 12%, 14%, and 16%. There is no effect on the performance of the proposed work below 10% mutation rate, and above 10%, the performance is degraded. Based on the population size the crossover rate is set at P/2. This results in half of the population being selected at each iteration for the reproduction operations. The parameters of the GA are shown in Table 2.

4.3. Plugging-in classifiers

Various experiments are conducted using the proposed model with SVM, *k*-NN, and RF. This set of experiment is divided into two main categories: the first category uses the SVM as a classifier while the second category uses *k*-NN. Table 3 lists results of the experiment with SVM plugged into the proposed framework. Here, the proposed model uses four kernels of SVM for classification. These include linear, polynomial, sigmoid, and RBF kernels. The fitness function also varies with SVM classifier.

Experiments are conducted for different values of *k*, i.e., 3, 5, 7, and 9. Fig. 5 shows the accuracies obtained for four values of *k*. Where, it can be seen that optimum accuracy is obtained for *k*=7. Therefore, the remaining experiments are conducted with this setting for *k*-NN. Experiments are also conducted for three distance measures, namely, Euclidean, Minkowski, and Manhattan distances. Table 4 lists the results using *k*-NN classifier and linear kernel as the fitness function. Performance variation can be observed in case of the distance measure. By plugging *k*-NN as a classifier and polynomial as a fitness

Table 6
Results of the proposed solution by plugging-in *k*-NN and sigmoid kernel as a fitness function

Dataset	Euclidean distance						Minkowski distance						Manhattan distance					
	Accuracy	Sensitivity	Specificity	Precision	F measure	Entropy	Accuracy	Sensitivity	Specificity	Precision	F measure	Entropy	Accuracy	Sensitivity	Specificity	Precision	F measure	Entropy
Leukemia	**0.94**	**0.93**	**0.97**	0.93	**0.94**	0.02	**0.97**	0.97	**0.96**	**0.96**	**0.97**	**0.29**	0.92	0.90	0.97	0.90	0.90	0.09
Colon cancer	0.75	0.75	0.64	**0.96**	0.84	0.03	0.81	0.80	0.83	0.95	0.81	0.15	0.78	0.76	0.86	0.95	0.84	0.05
DLBCL	0.84	0.93	0.63	0.87	0.89	0.12	0.90	**0.99**	0.74	0.87	0.92	0.12	0.89	0.96	0.70	0.92	0.93	0.07
Prostate cancer	0.81	0.88	0.74	0.79	0.83	0.19	0.78	0.84	0.74	0.75	0.79	0.21	0.79	0.83	0.76	0.80	0.84	0.17
Central nervous system	0.67	0.73	0.54	0.83	0.77	0.15	0.55	0.64	0.47	0.69	0.66	0.25	0.70	0.75	0.64	0.81	0.77	0.17
Lung cancer	0.92	0.81	0.94	0.68	0.73	**0.26**	0.94	0.91	0.95	0.72	0.82	0.23	**0.99**	**0.98**	**0.99**	**0.98**	**0.98**	**0.19**

Table 7
Results of the proposed solution by plugging-in *k*-NN and RBF kernel as a fitness function

Dataset	Euclidean distance						Minkowski distance						Manhattan distance					
	Accuracy	Sensitivity	Specificity	Precision	F measure	Entropy	Accuracy	Sensitivity	Specificity	Precision	F measure	Entropy	Accuracy	Sensitivity	Specificity	Precision	F measure	Entropy
Leukemia	**0.97**	**0.97**	**0.96**	**0.98**	**0.97**	0.02	0.97	0.97	0.96	0.97	0.97	0.03	0.97	0.98	0.97	**0.98**	**0.98**	0.02
Colon cancer	0.76	0.75	0.79	0.92	0.82	0.08	0.81	0.80	0.81	0.94	0.80	0.17	0.79	0.83	0.78	0.54	0.65	**0.33**
DLBCL	0.85	0.91	0.71	0.90	0.79	0.20	0.88	0.92	0.81	0.93	0.92	0.07	0.85	0.90	0.76	0.91	0.90	0.09
Prostate cancer	0.80	0.83	0.77	0.80	0.81	0.18	0.79	0.84	0.75	0.78	0.80	**0.19**	0.81	0.84	0.78	0.81	0.82	0.17
Central nervous system	0.64	0.70	0.45	0.83	0.75	0.15	0.84	0.88	0.81	0.85	0.86	0.13	0.60	0.71	0.22	0.78	0.74	0.19
Lung cancer	0.56	0.68	0.46	0.67	0.67	**0.26**	**0.99**	**0.99**	**1.00**	**0.98**	**0.98**	0.02	**0.99**	**1.00**	**0.99**	0.97	**0.98**	0.03

function, the performance of the proposed work on all the datasets for using Euclidean, Minkowski, and Manhattan distance measures is mentioned in Table 5. For the Euclidean distance and prostrate dataset, which is the largest datasets in terms of the number of features, this work yields an accuracy of 81%, 83% sensitivity, 79% specificity, and 87% precision. The average performance of the proposed work for the Euclidean distance measure is 84%, 87%, 80%, and 88% accuracy, sensitivity, specificity, and precision, respectively. The performance based on Minkowski as a distance measure on the prostrate dataset shows an accuracy of 83%, 86% sensitivity, 77% specificity, and 82% precision. The average performance on the Minkowski measures is 83%, 86%, 80%, and 89% as accuracy, sensitivity, specificity, and precision, respectively. The proposed idea archives 87% accuracy, 92% sensitivity, 81% specificity, and 85% precision, respectively for prostate cancer using Manhattan as a distance measure. Table 6 lists the results using *k*-NN classifier and sigmoid kernel as the fitness function. This experiment gives 81% accuracy, 88% sensitivity, 74% specificity, and 79% precision on the prostrate dataset for the Euclidean distance measure. Considering all datasets, the average accuracy, sensitivity, specificity, and precision is 83%, 74%, 84%, and 84%, respectively. The proposed idea on the largest dataset has 78% accuracy, 84% sensitivity, 74% specificity, and 75% precision using Minkowski as distance measures. For this measure, 82%, 85%, 78%, and 82% are the average accuracy, sensitivity, specificity, and precision, respectively on six datasets. Using Manhattan as a distance measure the average performance on all datasets is 86% accuracy, 82% sensitivity, 87% specificity, and 89% precision. However, for the large dataset, it gives 79% accuracy, 83% sensitivity, 76% specificity, and 80% precision.

Table 7 lists the results using *k*-NN classifier and RBF kernel as the fitness function. The average accuracy, sensitivity, specificity, and precision for the Euclidean measure is 76%, 80%, 69%, and 80%, respectively. The results for the largest datasets for this measure is 80% accuracy, 83% sensitivity, 77% specificity, and 80% precision. Whereas, using Minkowski as a distance measure the accuracy, sensitivity, specificity, and precision obtained on the prostrate dataset are 79%, 84%, 75%, and 78%, respectively. For this measure, the average accuracy, sensitivity, specificity, and precision is 88%, 90%, 85%, and 90%, respectively. The average results give 83% accuracy, 87% sensitivity, 75% specificity, and 83% precision on all datasets for the Manhattan measure. From the results in Tables 3-7 it can be seen that the proposed work is applied to six benchmark gene expression datasets to confirm its effectiveness. The size of the prostate cancer dataset in terms of the number of samples and features is the largest. The experiments represent that the accuracy of the proposed model on the prostate dataset is 94% and 83% by plugging SVM and *k*-NN classifier, respectively. The proposed model gives 99% accuracy on lung cancer data by plugging in SVM classifier and also for *k*-NN for the values of k=3 and k=7 (and polynomial kernel as a fitness function). The accuracy of the proposed model on colon cancer data is 90% with SVM over linear kernel as a fitness value. Whereas, using *k*-NN accuracy reaches up to 82% with Euclidean distance and linear functions as a fitness value. Using SVM as a classifier and linear function as a fitness value, the proposed algorithm gives 98% of the accuracy on DLBCL dataset. The accuracy while using *k*-NN (for k=7) and Euclidean distance plus linear as a fitness value is 91%. For the CNS dataset the proposed model has 99% accuracy using the SVM classifier. However, using *k*-NN (for k=7) with Euclidean as a distance measure and linear as a fitness value, the accuracy obtained is 80%. The leukemia dataset gives 99% accuracy of the proposed model by using the linear function as a fitness value. However, the *k*-NN gives 97% accuracy.

It is therefore extracted from these results that the proposed model works at its optimum for *k*-NN classifier with the value of k being 7 and Euclidean as a distance measure plus linear function as a fitness function. However, comparing the results

Table 8
Comparison of the proposed approach with five state-of-the-art methods

Methods	Metrics	Leukemia	DLBCL	Lung cancer	Colon cancer	Prostate cancer	Central nervous system
(Ayyad et al., 2019) [32]	Accuracy	0.7907	0.6697	0.8634	0.6086	0.6023	0.8361
	Sensitivity	0.8303	0.8990	0.7647	0.7597	0.7272	0.8013
	Specificity	0.7306	0.4617	0.8909	0.5115	0.4556	0.1961
	Precision	0.8582	0.6339	0.4397	0.4438	0.5014	0.5162
	F-measure	0.8440	0.7436	0.5584	0.6422	0.5916	0.6033
	Entropy	0.0570	**0.1255**	**0.1569**	0.1417	**0.1507**	0.1526
(Saleem et al., 2017) [1]	Accuracy	0.7333	0.8750	0.3784	0.6923	0.5714	0.8462
	Sensitivity	**1.0000**	**1.0000**	0.5000	0.6250	0.9412	1.0000
	Specificity	0.5000	**1.0000**	0.7857	0.5714	0.5000	1.0000
	Precision	0.7333	0.8667	0.1304	0.8333	0.5926	0.8000
	F-measure	0.8462	0.9286	0.2069	0.7143	0.7273	0.8889
	Entropy	0.0988	0.0539	0.1154	0.0660	0.1347	0.0775
(Uzma et al., 2020) [56]	Accuracy	0.9000	0.9750	0.9946	0.8462	**0.9400**	0.8462
	Sensitivity	0.8333	**1.0000**	**1.0000**	**1.0000**	**0.9556**	0.7500
	Specificity	**1.0000**	0.9636	0.9714	0.8333	0.9400	0.8889
	Precision	**1.0000**	0.9333	0.9935	0.3333	0.9400	0.7500
	F-measure	0.9000	0.9636	0.9967	0.5000	**0.9400**	0.7500
	Entropy	0.0000	0.0608	0.0063	**0.3662**	0.0000	0.2158
(Rani et al., 2019) [57]	Accuracy	0.9333	**1.0000**	**1.0000**	0.8462	0.9375	0.6923
	Sensitivity	0.8889	**1.0000**	**1.0000**	**1.0000**	0.8889	0.5500
	Specificity	**1.0000**	**1.0000**	**1.0000**	0.8333	**0.9605**	0.7556
	Precision	**1.0000**	**1.0000**	**1.0000**	0.3333	0.9194	0.5000
	F-measure	0.9412	**1.0000**	**1.0000**	0.5000	0.9000	0.5067
	Entropy	0.0000	0.0000	0.0000	**0.3662**	0.0761	**0.3466**
(Ghosh et al., 2019) [35]	Accuracy	**1.0000**	0.9375	**1.0000**	0.7846	0.9000	0.7538
	Sensitivity	**1.0000**	0.9833	**1.0000**	0.7718	0.8926	0.7586
	Specificity	**1.0000**	0.8048	**1.0000**	0.7648	0.9204	0.8850
	Precision	**1.0000**	0.9349	**1.0000**	0.8150	0.9308	0.8893
	F-measure	**1.0000**	0.9585	**1.0000**	0.7928	0.9113	0.8188
	Entropy	0.0000	0.0273	0.0000	0.0724	0.0290	0.0453
LSFS (Proposed)	Accuracy	0.9900	0.9800	0.9900	**0.9000**	**0.9400**	**0.9800**
	Sensitivity	0.9900	0.9400	0.9800	0.9200	0.9400	**0.9900**
	Specificity	0.9800	0.9400	0.9900	**0.8500**	0.9400	**0.9600**
	Precision	0.9900	0.9500	0.9700	**0.9200**	**0.9500**	**0.9800**
	F-measure	0.9900	0.9400	0.9700	0.9200	**0.9400**	**0.9800**
	Entropy	0.0090	0.0400	0.0200	0.0700	0.0400	0.0190

Table 9
Optimal features count selected after ensemble filter, LSFS, and GA

Datasets	Ensemble filter	LSFS	GA
Leukemia	16	16	18.6
Colon cancer	19.8	19.8	21.7
DLBCL	19	19	21.9
Prostate cancer	20	20	30.6
Central nervous system	19	19	25.4
Lung cancer	19	19	20.2

obtained via SVM and *k*-NN classifiers suggest that SVM with a linear function as a fitness function performs better. Table 8 represent the comparison between the proposed work and five state-of-the-art methods for the same task. The table shows accuracies of the six competing methods on six datasets. Where, the proposed work performs better than others for the CNS, prostate cancer, and DLBCL datasets. It performs second best over the leukemia dataset.

Table 8 also lists the accuracy, sensitivity, specificity, precision, F-measures, and entropy measures of the six competing algorithms. The results suggest that the proposed algorithm performs better for the colon cancer dataset in the case of the sensitivity, specificity, precision, F-measure and entropy, which have values 0.925, 0.857, 0.92549, 0.92, and 0.07, respectively. The comparison using the colon cancer dataset suggest that for the metrics of sensitivity, specificity, precision, F-measure, and entropy the proposed work performs better than other five approaches. The performance of the proposed work on high dimensional dataset, i.e., prostate is also better based on sensitivity, specificity, precision, F-measure and entropy metrics. The current work also perform better for CNS and lung cancer datasets based on sensitivity, specificity, precision, F-measure and entropy.

The proposed model selects the optimum features after applying ensemble-based filter, LSFS, and GA. The detail about the number of selected features using these feature selection techniques are shown in Table 9. The selected optimal features for smaller datasets is 18 and for the larger datasets it is 30 as shown in Fig. 6. As shown in Fig. 7 the average classification time of the proposed model is comparatively lesser on all the datasets. The computational time of the classification model combined with the feature selection process is comparatively less when paralleled with the rest of the datasets as shown in Fig. 8. The proposed model selects a small number of optimal features, yielding better accuracy with less computational timing as compared to other algorithms for all the datasets.

4.4. Experiments on influence of feature selection ensemble and LSFS

The two major contributions of this work are the feature selection ensemble and the LSFS method. An experiment has been performed to see the effect of these on the obtained results.

Table 10 shows the effect of using the ensemble filter method, results after applying the LSFS heuristic, and also the obtained results after using the proposed GA-based solution for the optimization purpose. The table shows that after applying the first phase, the classifiers obtain and average accuracy, sensitivity, specificity, precision, F-measure, and entropy of 86%, 74%, 95%, 91%, 80%, and 0.06%, respectively on all six datasets. An enhanced performance is obtained using the proposed where an average accuracy, sensitivity, specificity, precision, F-measure, and entropy of 96%, 96%, 94%, 86%, 95%, and 0.03% is obtained respectively on all six datasets. The table also lists the performance after applying the LSFS heuristic.

4.5. Statistical significance

In order to show the statistical significance of the proposed approach in comparison to the five state-of-art algorithms, the paired sample *t*-test is conducted. For this, first the null (H_0) and alternate hypothesis (H_A) are defined. These are listed in Table 11. The performance of the proposed approach is evaluated against the competing algorithms based on the confusion matrix. The probability to reject the null hypothesis, which is called the level of significance (i.e., α), is set to 5%. Whereas, the probability to accept the null hypothesis, called the confidence level (1- α), is 95%. The degree of freedom (*df*) shows the number of datasets and it is set to 6. The *p*-value is the probability value that determines the evidence to reject the null hypothesis in favor of an alternative value. A smaller *p* value shows strong evidence to reject the null hypothesis.

The *t*-statistical test is performed based on the performance metrics, such as, accuracy, sensitivity, specificity, and precision to show the significance of the proposed idea. Therefore, first the scores are assigned based on the abovementioned four performance metrics to each algorithm for all datasets as shown in Table 12. The scores represent the ratio of the number of performance metrics for which the algorithm performs best out of the total performance metrics. Using the data mentioned in Table 12 the result of the of the paired sample *t*-test are shown at Table 13. The paired sample *t*-test shows a significant difference between the proposed method and (Ayyad et al., 2019) [$t(5) = 8.907279194, p < 0.05$)], (Salem et al., 2017) [$t(5) = 4.706184093$, $p < 0.05$)], (Uzma et al., 2020) [$t(5)=2.7643$, $p<0.05$], (Rani et al., 2019) [$t(5)= 2.869193781$, $p<0.05$] and (Ghosh et al., 2019) [$t(5) = 3.299572848$, $p < 0.05$)]. Thus, these analysis conclude that there is a significant difference between the groups based on the *p*-value. Hence, the null hypothesis (H_0) is rejected by obtaining small *p*-value in favor of the alternative hypothesis.

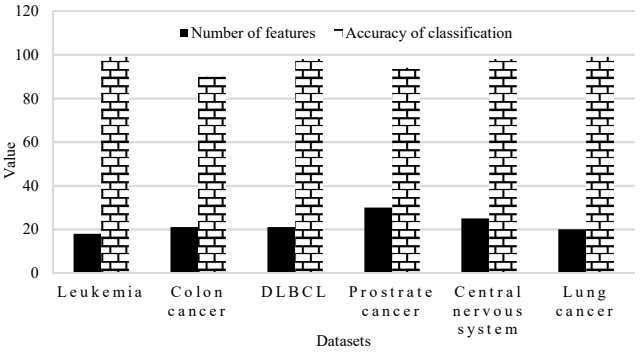

Fig. 6. Optimum number of features per dataset

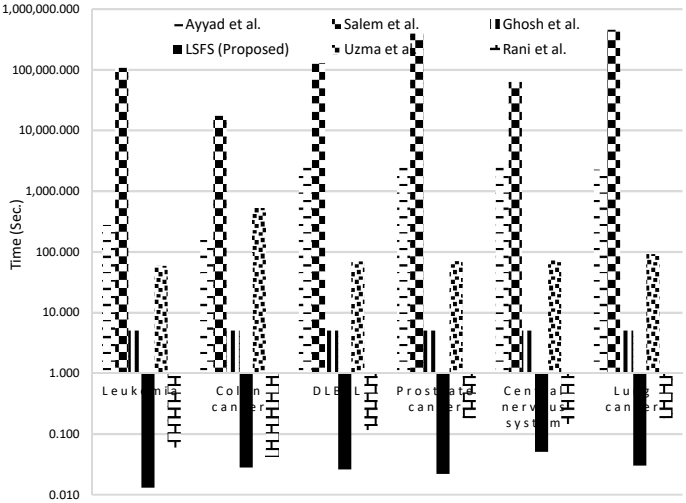

Fig. 7. Average classification time

4.6. Comparison with other evolutionary feature selection methods

An experiment is performed to compare the proposed approach with two newer evolutionary computing-based

Table 10
The effect of filter method and LSFS

After applying	Datasets	Accuracy	Sensitivity	Specificity	Precision	F measure	Entropy
Filter method	Leukemia	0.93	0.86	**1.00**	**1.00**	0.92	0.00
	Colon cancer	0.77	0.40	**1.00**	**1.00**	0.57	0.00
	DLBCL	1.00	**1.00**	**1.00**	**1.00**	**1.00**	0.00
	Prostate cancer	0.89	0.79	**1.00**	**1.00**	0.88	0.00
	Central nervous system	0.62	0.40	0.75	0.50	0.44	0.35
	Lung cancer	1.00	**0.99**	**1.00**	**1.00**	**1.00**	0.00
LSFS	Leukemia	1.00	**1.00**	**1.00**	**1.00**	**0.99**	0.00
	Colon cancer	0.77	0.71	0.83	0.83	0.77	**0.15**
	DLBCL	0.94	**1.00**	0.92	0.80	0.88	**0.18**
	Prostate cancer	0.86	**1.00**	0.76	0.73	0.85	**0.23**
	Central nervous system	0.62	0.33	0.86	0.67	0.44	**0.27**
	Lung cancer	1.00	1.00	**1.00**	**1.00**	1.00	0.00
Proposed solution	Leukemia	**0.99**	0.99	**1.00**	**1.00**	**0.99**	**0.01**
	Colon cancer	**0.98**	**0.94**	**1.00**	0.95	**0.94**	0.04
	DLBCL	**0.99**	0.98	**1.00**	**1.00**	0.97	0.02
	Prostate cancer	**0.90**	0.92	0.85	0.92	**0.92**	0.07
	Central nervous system	**0.94**	**0.94**	**0.94**	0.95	**0.94**	0.04
	Lung cancer	**0.98**	**0.99**	0.96	**1.00**	0.98	0.02

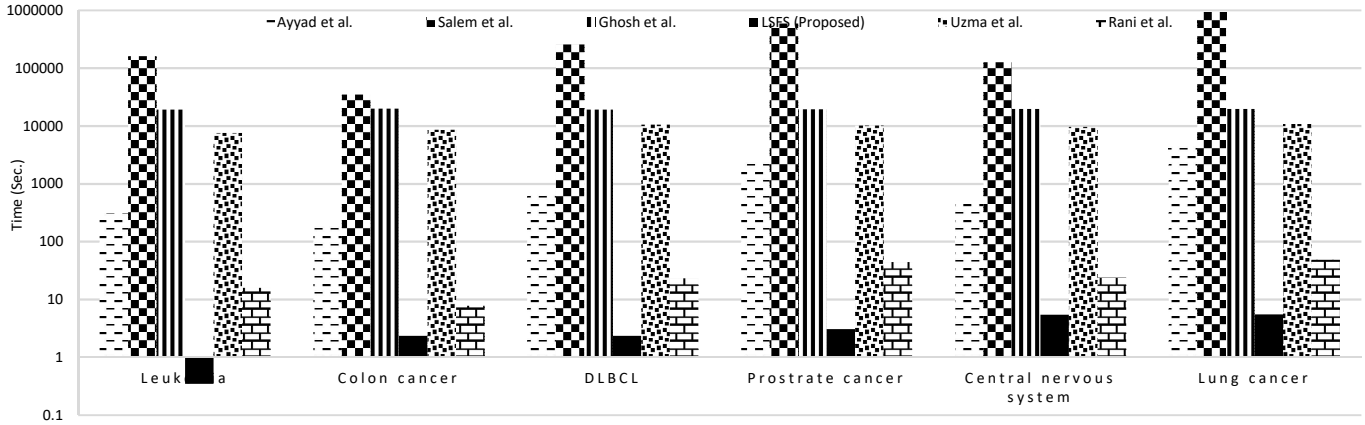

Fig. 8. Average classification time combined including the feature selection process

feature selection methods that target better classification accuracy. For this, GBA [37] and IIEFS algorithm [38] are used. The GBA is a new evolutionary computation-based approach for feature selection. Whereas, IIEFS is also an evolutionary computing approach that uses interaction information to have higher-level interaction analysis to enhance the search process in the feature space. For this experiment, the three competing methods are executed on the six benchmark datasets. The results of this experiment are shown in Fig. 9. The obtained results suggest better performance of the present work in majority of the cases. The average accuracy of the proposed approach on the six datasets is 98%, whereas the IIEFS methods has performed the second best by achieving an average accuracy of 97%. The accuracy obtained on all the datasets for the present work in better than the other two methods with an exception on leukemia and prostate cancer dataset, where IIEFS performed better. The performance of the GBA is towards the lower side in comparison to the other two methods on the six datasets. For the sake of a fair comparison, all the competing methods are restricted to extract the same number of features and the classification accuracies reported here are based on an average of 10 runs using the SVM classifier's earlier obtained optimum configuration.

5. Discussion

The gene expression analysis is of significance in the medical sciences and other domains due to the "mystery" of biological systems. Therefore, one needs to understand gene expression data and extract important information. The DNA microarray technology can identify the expression level of hundreds of genes simultaneously. However, the internal view of the gene expression dataset makes it challenging to process**.** The classification task of cancer gene expression data has a set of challenges because of its unique behavior. A main challenge in the classification of gene expression datasets involves large dimensions of the features while the relevant features are quite less comparatively. Therefore, in the literature, feature selection techniques are used to reduce the dimensions of the data for better gene expression analysis. The key concern of the past

Table 11
Null hypotheses with its alternate

Null hypotheses	Alternate hypothesis
H_0: The proposed approach does not perform better based on the six performance metrics	H_A: The proposed approach performs better based on the six performance metrics

Table 12
Ranking of the competing methods

Datasets	Proposed	Ayyad et al., 2019	Salem et al., 2017	Uzma et al.,2020	Rani et al., 2019	Ghosh et al, 2019
Leukemia	1.00	0.00	0.17	0.33	0.33	0.17
DLBCL	0.50	0.00	0.33	0.33	0.33	0.33
Lung cancer	1.00	0.00	0.00	0.33	0.67	0.33
Colon cancer	0.67	0.00	0.17	0.17	0.17	0.00
Prostrate cancer	0.67	0.00	0.00	0.67	0.00	0.17
Central nervous system	1.00	0.00	0.17	0.00	0.00	0.00
Avg.	0.81	0.00	0.14	0.31	0.25	0.17
Std.dev	0.22	0.00	0.13	0.22	0.25	0.15

Table 13
Paired sample *t*-test

	Paired differences					T	df	Sig. (2-tailed)
	Mean	Std. deviation	Std. error mean	95% confidence interval of the difference				
Pairs				Lower	Upper			
Proposed.-Ayyad et al.	0.806	0.222	0.090	0.806	0.806	8.907	5.000	0.000
Proposed-Salem et al.	0.667	0.096	0.039	0.667	0.667	4.706	5.000	0.005
Proposed-Uzma et al.	0.500	0.000	0.000	0.500	0.500	2.764	5.000	0.040
Proposed-Rani et al.	0.556	-0.031	-0.013	0.556	0.556	2.869	5.000	0.035
Proposed- Ghosh et al.	0.639	0.072	0.030	0.639	0.639	3.300	5.000	0.021

works is to find the optimal feature subset from a high dimensional feature set that efficiently classifies the cancerous samples. The use of a two-phased technique, such as, filters with the wrapper method has shown promising results in the past [1, 37, 38, 35]. The current work aimed to address this problem by designing two phases of feature selection techniques. It presented a filter with a wrapper-based methodology for an optimal feature subset selection from the cancer gene expression data. For improving the stability of the feature selection, the ensemble of filter methods was used in the first stage. The first stage of the proposed work combined the three filter methods, i.e., SU, X^2, and Relief. These filters measures used different criteria for measuring the importance of features. Therefore, the current work combined the essential features in three ways. The feature subset selected via the ensemble method is more robust. The top N ranked features were picked from multiple methods and combined together. The final subset earned by union aggregation was passed to the second stage. The second stage was used for the global optimization of the first stage. This phase used the combination of a local search-based feature subset (LSFS) and a GA for optimization. A heuristic may produce quality results, but it can stuck in the local optima. Whereas, using the population-based metaheuristic, such as, GA has a tendency to find the global optimal. Hence the combination of heuristic algorithm and GA performs well as they are able to take advantage of their respective strengths while suppressing their individual shortcomings. Therefore, in the proposed work, the GA is initialized with the proposed heuristic, i.e., LSFS to produce efficient solutions by avoiding random population generations. The feature subset from the first phase had some chances of having noise due to the combination of relevant and non-redundant features. To remove the noise from feature subset LSBFS was applied. The performance of each subset was evaluated using the SVM classifier with 5-fold cross-validation. The GA led to the selection of the feature subset that provided better classification accuracy.

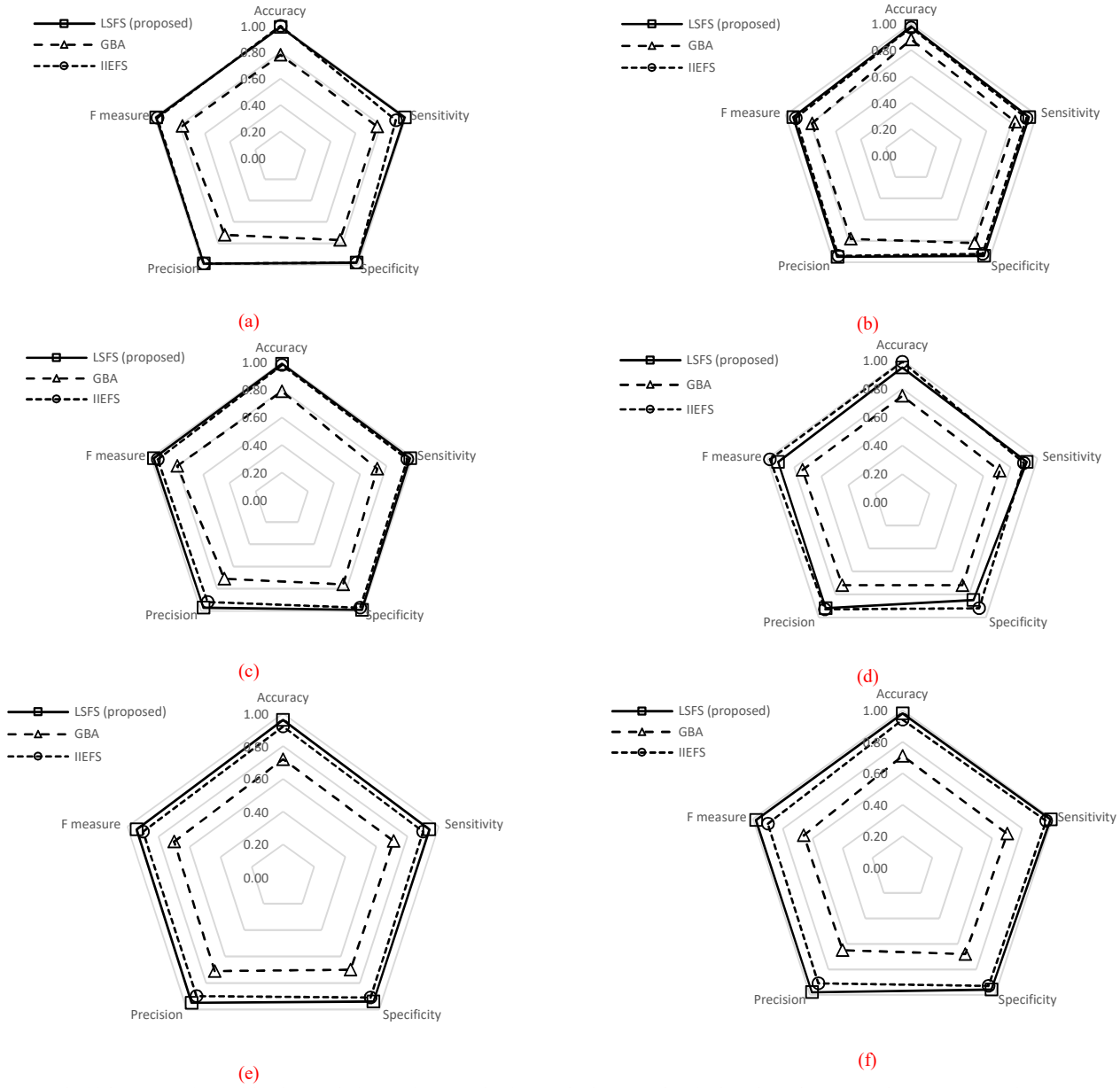

Fig. 9. Results if the Comparison with other evolutionary feature selection methods (a) Leukemia data (b) Colon cancer data (c) DLBCL data () Prostate cancer data (e) Central nervous system data (f) Lung cancer data

Various experiments were performed for the classification of the cancerous samples based on the selected feature subsets. These experiments were categorized by using the various number of classifiers. The three classifiers utilized for the evaluation of the proposed solution included SVM, *k*-NN, and RF. The SVM was used with three types of the kernels (Table 3). Using linear kernel the average performance of the proposed work was: 96% accuracy, 97% sensitivity, 94% specificity, 97% precision, 97% F-measures, and 0.028 entropy on all six datasets. However, for the sigmoid kernel, the average accuracy, sensitivity, specificity, precision, F-measure, and entropy was 81%, 70%, 87%, 81%, 73%, and 0.03, respectively. The average performance of the proposed work using polynomial kernel was: 84% accuracy, 85% sensitivity, 82% specificity, 94% precision, 88% F-measure, and 0.04 entropy on all datasets. While using RBF as the SVM kernel the average performance of the proposed work on six datasets showed 79% accuracy, 79% sensitivity, 80% specificity, 84% precision, 92% F-measures, and 0.033 entropy. The comparisons of four types of kernels identified that the linear kernel performed better for the proposed work. By plugging *k*-NN as a classifier and polynomial/Sigmoid/RBF as a fitness functions, the performance of the proposed work on all the datasets for using Euclidean, Minkowski, and Manhattan as a distance was also observed (Tables 4, 5 and 6).

The proposed solution performed better with SVM and the linear kernel. However, the current work also performed better by using *k*-NN with Euclidean as distance measures and for *k* equals 7. The results on the proposed work suggest that comparatively, SVM performs better when plugged into the proposed solution. The proposed approach was also compared with five closely related methods. These results are shown in Table 8. The accuracy of the proposed work on leukemia, DLBCL, lung cancer, colon cancer, prostate, and CNS dataset was 99%, 98%, 99%, 90%, 94%, and 98%, respectively. Based on accuracy, the proposed approach performs better than others in the majority of the cases. Considering the same metric, the methods Rani et al. [47-57] and Ghosh et al. [35] perform better in two cases each thus becomes the second best solution. The three methods, i.e., present proposal, [57], and [35] perform close to each other. The reason being their utilization of multiple filter methods as an enable. However, the current proposal has an added advantage of using the combination of LSFS and GA for further optimizing the obtained result. The assemble filter method generates noise due to the combination of various subsets of features generated using various filter methods. It is, therefore, possible to select an irrelevant feature, which may reduce the performance of the classification. Consequently, the proposed heuristic (LSFS) is used, which employ the 2-opt operator. It uses a systematic method to remove the irrelevant features or select the relevant feature based on IG value. After the first phase of the proposed idea, this serves as a filter to choose the most relevant features. The results showed that current work performs better for colon cancer, prostate cancer, and central nervous system datasets. The accuracy of the proposed idea on the smallest and largest dataset is better than the competing algorithms.

Like any contribution, other than the novelty and multiple strengths of the proposed work, there are a few limitations of this work as well. A limitation of the proposed LSFS algorithm is that it depends on the parameter *s_rate* that generates the number of solutions in a systematic way. As the number of solutions increases, the time consumed by LSFS also surges. Therefore, limitation of this work is dependency of the optimal feature subset on the number of candidate solutions. However, to overcome this issue in the simulations, the parameter *s_rate* is set to a limited value.

6. Conclusion

The analysis of gene expression datasets is quite challenging because of various reasons such as; (a) much larger feature dimensions than the number of samples, (b) irrelevant, redundant, and noisy features (c) very few pertinent features being relevant for the identification of cancer biomarkers. The current work presented a two-phased optimization technique for the feature selection of cancer gene expression datasets. The proposed work reported a novel strategy for the optimization of existing approaches that are used for the identification of genes associated with cancer. The first phase of the proposed work used the ensemble filter-based feature selection. This phase was the combination of three filter-based methods for feature selection, namely, Symmetrical Uncertainty (SU), chi square (X^2), and Relief. Each filter method had its own criteria for the selection of features, giving multiple feature subsets using the same datasets. Therefore, the ensemble method is more robust than a single filter-based feature selection for the removal of irrelevant and redundant features. The second phase of the proposed methodology involved the combination of a newly designed heuristic approach called Local Search-based Feature Selection (LSFS) followed by the Genetic Algorithm (GA). This technique, i.e., LSFS removed the noise generated by the ensemble filter method. The results obtained via LSFS were used as an initial solution of the GA. The Support Vector Machine (SVM) kernel was used for the fitness evaluation of the chromosome. The roulette wheel selection, uniform cross over and swap mutation were used as the genetic operators. The performance of the proposed work was evaluated by comparing it with the five state-of-the-art algorithms for the same problem. Six evaluation metrics, namely, accuracy, sensitivity, specificity, precision, F-measure, and entropy were used. The classification was done using two classifiers, i.e., SVM and *k*-NN. The post-experimental analysis revealed that the accuracy of the proposed model was better while using SVM as a classifier and linear kernel as a fitness function. The SVM gave 99% accuracy as compared to the *k*-NN on the proposed methodology for all datasets. The proposed work selected a small number of features, giving higher accuracy as compared to other competing algorithms. The number of features selected for a smaller dataset was 18 while for a larger dataset these were 30.

This work has multiple possible future directions. In the future, a computational intelligence-based framework can be designed for the optimization and predication of biological information. To predict the protein secondary structure and for the enzyme function classification, utilization of various deep learning methods can also be a future undertaking. Another future extension can be to use the gene expression data for neurodegenerative diseases on which popular data science techniques can be applied that could help in the identification of genes associated with a particular neuronal condition. From

the machine learning perspective, the current work can be extended in two ways, i.e., feature selection for supervised and unsupervised methods. For the supervised feature selection, the deep learning can be utilized by incorporating it in an optimization algorithm as a fitness function. Whereas, in case of the unsupervised method, a self-organizing map-based genetic algorithm can be employed to select the optimal feature subset.

Acknowledgement

The authors are indebted to the editor and anonymous reviewers for their helpful comments and suggestions. The authors wish to thank GIK Institute for providing research facilities. This work was sponsored by the GIK Institute graduate research fund under GA-F scheme.

References

[1]H. Salem, G. Attiya, and N. El-Fishawy, "Classification of human cancer diseases by gene expression profiles," *Applied Soft Computing*, vol. 50, pp.124-134, 2017.

[2]H. M. Alshamlan, G. H. Badr and Y. A. Alohali, "Genetic Bee Colony (GBC) Algorithm: A New Gene Selection Method for Microarray Cancer Classification," *Computational Biology and Chemistry*, vol. 56, pp.49–60, 2015.

[3]E. Bard and W. Hu, "Identification of a 12-Gene Signature for Lung Cancer Prognosis through Machine Learning," *Journal of Cancer Therapy*, vol. 2, pp. 148-156, 2011.

[4] J. A. Cruz and D. S. Wishart, "Applications of Machine Learning in Cancer Prediction and Prognosis," US National Library of Medicine National Institutes of Health, Cancer Informatics, vol. 2, pp. 59–77, 2006.

[5] G. Chakraborty and B. Chakraborty, "Multi-objective Optimization Using Pareto GA for Gene-Selection from Microarray Data for Disease Classification," IEEE International Conference Systems, Man, and Cybernetics (SMC), pp. 2629 – 2634, 2013.

[6] I. Guyon, & A. Elisseeff, "An introduction to variable and feature selection," *Journal of machine learning research*, vol.3, pp. 1157-1182, 2003.

[7] T.R. Golub, D.K. Slonim, P.Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, and , C.D. Bloomfield, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, vol. 286*, no.* 5439, pp.531-537, 1999.

[8] V. Bolón-Canedo, N. Sánchez-Marono, A. Alonso-Betanzos, J.M. Benítez, and F. Herrera, " A review of microarray datasets and applied feature selection methods," *Information Sciences*, vol. *282*, pp. 111-135,2014.

[9] I. Guyon, S. Gunn, M. Nikravesh, and L.A. eds. Zadeh, *Feature extraction: foundations and applications,* vol. 207, Springer, 2008.

[10] H. Yu, J. Ni, Y. Dan, and S. Xu, "Mining and integrating reliable decision rules for imbalanced cancer gene expression data sets," *Tsinghua Science and technology*, vol. 17, no. 6, pp. 666-673, 2012.

[11] D. Lavanya, and K.U. Rani, "Ensemble decision tree classifier for breast cancer data," *International Journal of Information Technology Convergence and Services*, vol. 2, no. 1, p. 17, 2012

[12] A. Rouhi, and H. Nizamabad-pour, "A hybrid method for dimensionality reduction in microarray data based on advanced binary ant colony algorithm," 1st Conference on Swarm Intelligence and Evolutionary Computation (CSIEC), pp. 70-75, Bam, Iran, 2016. doi: 10.1109/CSIEC.2016.7482124 .

[13] M. Dash, M. and H. Liu, 1997. "Feature selection for classification," *Intelligent data analysis*, vol. 1*,* no. 1-4, pp. 131-156, 1997.

[14] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507-2517, 2007.

[15] J.C. Ang, A. Mirzal, H. Haron, and H.N.A. Hamed,"Supervised, unsupervised, and semi-supervised feature selection: a review on gene selection," *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 13, no. 5, pp. 971-989, 2016.

[16] Q. Shen, R. Diao, and P. Su, "Feature selection ensemble," in Proc. Turing-100, vol. 10, pp. 289–306, 2012.

[17] W. Awada, T.M. Khoshgoftaar, D. Dittman, R. Wald, and A. Napolitano, "A review of the stability of feature selection techniques for bioinformatics data," In *IEEE 13th International Conference on Information Reuse & Integration (IRI)*, pp. 356-363, 2012.

[18] Y. Peng, Z. Wu, and J. Jiang, "A novel feature selection approach for biomedical data classification," *Journal of Biomedical Informatics*, vol. *43,* no. 1, pp. 15-23, 2010.

[19] J.C. Ang, A. Mirzal, H. Haron, and H.N.A. Hamed, "Supervised, unsupervised, and semi-supervised feature selection: a review on gene selection," IEEE*/ACM transactions on computational biology and bioinformatics*, vol. 13, no. 5, pp. 971-989, 2016.

[20] D. Rodríguez, R. Ruiz, J. Cuadrado-Gallego, and J. Aguilar-Ruiz, "Detecting fault modules applying feature selection to classifiers," *IEEE International Conference on Information Reuse and Integration*, pp. 667-672, 2007.

[21] P.M. Shakeel, A.Tolba, Z. Al-Makhadmeh, and , M.M. Jaber, "Automatic detection of lung cancer from biomedical data set using discrete AdaBoost optimized ensemble learning generalized neural networks," *Neural Computing and Applications*, pp.1-14, 2019.

[22] N. Hoque, M. Singh, and D.K. Bhattacharyya, "EFS-MI: an ensemble feature selection method for classification," *Complex & Intelligent Systems*, vol. 4*,* no. 2, pp. 105-118, 2018.

[23] H. Wang, T.M. Khoshgoftaar, and A. Napolitano, "December. A comparative study of ensemble feature selection techniques for software defect prediction," In IEEE 2010 Ninth International Conference on Machine Learning and Applications, pp. 135-140, 2010.

[24] H. Liu, H. Motoda, R. Setiono, and Z. Zhao, "Feature selection: An ever evolving frontier in data mining," In *Feature Selection in Data Mining*, pp. 4-13, 2010.

[25] D. Pavithra, and B. Lakshmanan, 2017, "June. "Feature selection and classification in gene expression cancer data," In IEEE. *International Conference on Computational Intelligence in Data Science (ICCIDS)*, pp. 1-6, 2017.

[26] M. Dashtban, and M. Balafar,"Gene selection for microarray cancer classification using a new evolutionary method employing artificial intelligence concepts," *Genomics*, vol.*109*, no. 2, pp. 91-107, 2017.

[27] A. Rouhi, and H. Nezamabadi-pour, "A hybrid method for dimensionality reduction in microarray data based on advanced binary ant colony algorithm," In IEEE 2016 1st Conference on Swarm Intelligence and Evolutionary Computation (CSIEC), pp. 70-75, 2016.

[28] Y. Li, G. Wang, H. Chen, L. Shi, and L. Qin, "An ant colony optimization based dimension reduction method for high-dimensional datasets," *Journal of Bionic Engineering*, vol. 10*,* no. 2, pp. 231-241, 2013.

[29] Uzma, Z. Halim, "Optimizing the minimum spanning tree-based extracted clusters using evolution strategy," *Cluster Computing*, vol. 21, no. 1, pp. 377-391, 2018.

[30] B.A. Garro, K. Rodríguez, and R.A. Vázquez, "Classification of DNA microarrays using artificial neural networks and ABC algorithm," *Applied Soft Computing*, vol. 38, pp. 548-560, 2016.

[31] B. Duval, and J.K. Hao, "Advances in metaheuristics for gene selection and classification of microarray data," *Briefings in bioinformatics*, vol. 11, no. 1, pp. 127-141, 2009.

[32] S.M. Ayyad, A.I. Saleh, and L.M. Labib,"Gene expression cancer classification using modified K-Nearest Neighbors technique," *Biosystems*, vol. 176, pp.41-51, 2019.

[33] S.A. Ludwig, S. Picek, and D. Jakobovic, 2018. "Classification of cancer data: analyzing gene expression data using a fuzzy decision tree algorithm," In *Operations Research Applications in Health Care Management*, Springer, Cham, pp. 327-347, 2018.

[34] B. Pes, " Ensemble feature selection for high-dimensional data: a stability analysis across multiple domains," *Neural Computing and Applications*, pp.1-23, 2019.

[35] M. Ghosh, S. Adhikary, K.K. Ghosh, A. Sardar, S. Begum, and R. Sarkar, "Genetic algorithm based cancerous gene identification from microarray data using ensemble of filter methods," *Medical & Biological Engineering & Computing*, vol. 57, no.1, pp. 159-176, 2019.

[36] B. Xue, M. Zhang, W.N. Browne, X. Yao, "A survey on evolutionary computation approaches to feature selection," *IEEE Transactions on Evolutionary Computation*, vol. 20, no. 4, pp. 606-626, 2015.

[37] Z.Y. Chen, W.C. Lin, S.W. Ke, C.F. Tsai, "Evolutionary feature and instance selection for traffic sign recognition," *Computers in Industry*, vol. 74, pp. 201-211, 2015.

[38] S.E. Hosseini, M.H. Moattar, "Evolutionary feature subsets selection based on interaction information for high dimensional imbalanced data classification," *Applied Soft Computing*, vol. 82, pp. 105581, 2019.

[39] C.S. Yusta, "Different metaheuristic strategies to solve the feature selection problem," Pattern Recognition Letters, vol. 30, no. 5, pp. 525-534, 2009.

[40] S.R. Ahmad, A.A Bakar, M.R. Yaakub, "Metaheuristic algorithms for feature selection in sentiment analysis," In 2015 Science and Information Conference, pp. 222-226, 2015.
[41] Sayed, G. I., Tharwat, A., & Hassanien, A. E. (2019). Chaotic dragonfly algorithm: an improved metaheuristic algorithm for feature selection. *Applied Intelligence*, *49*(1), 188-205.
[42] M. Mafarja, A. Qasem, A.A. Heidari, I. Aljarah, H. Faris, S. Mirjalili, "Efficient hybrid nature-inspired binary optimizers for feature selection," *Cognitive Computation*, vol. 12, no. 1, pp. 150-175, 2020.
[43] O. Gokalp, E. Tasci, A. Ugur, "A novel wrapper feature selection algorithm based on iterated greedy metaheuristic for sentiment classification," *Expert Systems with Applications*, vol. 146, pp. 113176, 2020.
[44] A.K. Shukla, P. Singh, M. Vardhan, "Gene selection for cancer types classification using novel hybrid metaheuristics approach," Swarm and Evolutionary Computation, vol. 54, pp. 100661, 2020.
[45] Y.Xue, Y.Tang, X.Xu, J.Liang, and F.Neri, "Multi-objective Feature Selection with Missing Data in Classification", In *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2021.
[46] Y.Xue, B.Xue, and M.Zhang, "Self-adaptive particle swarm optimization for large-scale feature selection in classification", *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol.*13,no.*5, pp.1-27,2019.
[47] Y.Zhang, D.W.Gong, and J.Cheng, "Multi-objective particle swarm optimization approach for cost-based feature selection in classification" In *IEEE/ACM transactions on computational biology and bioinformatics*, vol.*14,no.*1, pp.64-75,2015.
[48] M. S. Uzer, N.Yilmaz, and o.Inan "Feature Selection Method Based on Artificial Bee Colony Algorithm and Support Vector Machines for Medical Datasets Classification", the scientific world journal, 2013.
[49]M. Günay, and Z. Orman, "A modified firefly algorithm-based feature selection method and artificial immune system for intrusion detection", *Uludağ University Journal of The Faculty of Engineering*, vol, *25, no.*1, pp.269-288,2020.
[50] B.Guan, Y.Zhao, Y.Yin, and Y. Li, "A differential evolution based feature combination selection algorithm for high-dimensional data", *Information Sciences*, vol.*547*, pp.870-886,2021.
[51] Y.Zhang, D.W.Gong , X.Z.Gao, T.Tian, and X.Y.Sun, "Binary differential evolution with self-learning for multi-objective feature selection", *Information Sciences*, vol.*507*, pp.67-85,2020.
[52] X.F.Song, Y.Zhang, Y.N.Guo, X.Y.Sun, and Y.L.Wang, "Variable-size cooperative coevolutionary particle swarm optimization for feature selection on high-dimensional data",In *IEEE Transactions on Evolutionary Computation*, vol.*24,no.*5, pp.882-895, 2020.
[53] X.F.Song, Y.Zhang, D.W.Gong, and X.Y.Sun, "Feature selection using bare-bones particle swarm optimization with mutual information", *Pattern Recognition*, vol.112, p.107804, 2021.
[54] X.F.Song, Y.Zhang, D.W.Gong, and X.Z.Gao, "A fast hybrid feature selection based on correlation-guided clustering and particle swarm optimization for high-dimensional data", In *IEEE Transactions on Cybernetics*,2021.
[55] S. Tiwari, B. Singh, and M. Kaur, "An approach for feature selection using local searching and global optimization techniques," *Neural Computing and Applications*, vol. 28, no. 10, pp. 2915-2930, 2017.
[56] Uzma, F. Al-Obeidat, A. Tubaishat, B. Shah, and Z. Halim, "Gene encoder: a feature selection technique through unsupervised deep learning-based clustering for large gene expression data", *Neural Computing and Applications*, pp.1-23, 2020.
[57] M. J. Rani, D. Devaraj, "Two-stage hybrid gene selection using mutual information and genetic algorithm for cancer data classification", *Journal of medical systems*, vol. 43, no. 8, pp. 235, 2019
[58] D. Guan, W. Yuan, Y.K. Lee, K. Najeebullah, and M.K. Rasel, "A review of ensemble learning based feature selection," *IETE Technical Review*, vol. 31, no. 3, pp. 190-198, 2014.
[59] T. Abeel, T. Helleputte, Y. Peer, P. Dupont, and Y. Saeys,"Robust biomarker identification for cancer diagnosis with ensemble feature selection methods," *Bioinformatics*, vol. 26, no. 3, pp. 392 98, 2010
[60] Y. Saeys, T. Abeel, and Y. Peer, "Robust feature selection using ensemble feature selection techniques," in Proceedings of the ECML PKDD, vol. 5212, pp. 313 25, 2008.
[61] Z. Halim, M. Atif, A. Rashid, C. A. Edwin, "Profiling Players Using Real-World Datasets: Clustering the Data and Correlating the Results with the Big-Five Personality Traits," *IEEE Transactions on Affective Computing*, vol. 10, no. 4, pp. 586-584, 2019.
[62] Q. Zhang, H. Wang, S.W. Yoon, "A 1-norm regularized linear programming nonparallel hyperplane support vector machine for binary classification problems," *Neurocomputing*, vol. 376, pp. 141-152, 2020.