Volume 2 Issue 1 | March 2026
e-ISSN : Applied for (Online)
Journal of Integrated Engineering Innovation and Applications

Comparative Analysis of Filter Feature Selection Methods on Microarray Datasets

Pushpa Makwane
Department of Applied Chemistry,
Government Holkar Science College,
Indore 452001, Madhya Pradesh, India
Email:

Madhuri Gokhale
Department of Computer Science and Engineering,
Jabalpur Engineering College
Jabalpur 482011, Madhya Pradesh, India
Email: mgokhale@jecjabalpur.ac.in

Saurabh Singh
Department of Computer Science and Engineering,
Jabalpur Engineering College
Jabalpur 482011, Madhya Pradesh, India
Email: ssingh@jecjabalpur.ac.in

Dinkar Dubey
Computer Science and Engineering,
Jabalpur Engineering College
Jabalpur 482011, Madhya Pradesh, India
Email:

ABSTRACT

Microarray technology is an emerging technology used to analyze large-scale gene expression data simultaneously. However, interpreting gene expression data remains a challenging task because of its high-dimensional and low-sample-size characteristics. Microarray datasets contain thousands of genes and only a limited number of samples, which complicates the classification process. Therefore, feature selection methods, also known as gene selection methods, are essential for identifying the most informative genes that provide maximum discriminative power between cancerous and normal tissues. Although several feature selection approaches have been proposed, there is still no universally accepted method that consistently produces optimal results across different datasets. In this study, a comparative analysis of four widely used filter-based feature selection methods, namely Chi-Square (χ^2), ReliefF, Mutual Information, and Symmetrical Uncertainty, is performed on five benchmark microarray cancer datasets: Colon, Central Nervous System (CNS), Leukemia, Lung, and Ovarian datasets. The selected features are evaluated using six machine learning classifiers, including Random Forest, Decision Tree, Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Naive Bayes, and Logistic Regression. Experimental results demonstrate that feature selection significantly improves classification performance by reducing irrelevant and redundant features. Among the evaluated classifiers, SVM combined with Mutual Information achieved the best overall performance on most datasets. The study provides a comprehensive evaluation of filter-based feature selection techniques and their impact on cancer classification accuracy using microarray data.

Index Terms:— *Microarray, classification, feature selection, gene selection.*

I. INTRODUCTION

Cancer is the second leading cause of death globally, with 9.6 million deaths per year. Further, there are about 18.1 million new cancer cases that emerge every year. There were over 100 types of cancer such as colon, liver, ovarian, and breast, and so on [1], [2]. Cancer research is one of the major research areas in the medical field. Accurate prediction of different tumor types has great value in providing better treatment and toxicity minimization on the patients. To gain a better insight into the problem of cancer classification, systematic approaches based on

gene expression analysis have been needed.

Machine learning (ML) is a set of tools utilized for the creation and evaluation of algorithms that facilitate prediction, pattern recognition, and classification. ML is based on four steps: Collecting data, picking the model, training the model, testing the model. Machine learning can be supervised or unsupervised. In unsupervised learning, the algorithm that will classify the data does not know which class the data belongs to. On the other hand, in supervised learning, the labeling information of the data is given to the algorithm which will classify the data. This information is used to gain experience in the machine, then it is expected to classify the data that the machine does not recognize.

Different classification methods from statistical and machine learning areas have been applied to cancer classification. With gene expression technology, microarray data is very different from any of the data these methods had previously dealt with. First, it has very high dimensionality, usually contains thousands to tens of thousands of genes. Second, publicly available data size is very small, all below 1000. Third, most genes are irrelevant to cancer distinction. Generally, in high-dimensional data, irrelevant and redundant genes do not only decrease the training strength but also negatively influence the performance of learning algorithms which is mainly caused by the curse of dimensionality. To address these problems, researchers have been investigated a large number of gene selection methods, many of them derived from the need to analyze microarray data to select the best discriminating gene called biomarker [3]. In a classification problem, gene selection is a central application of data reduction to avoid challenges, such as overfitting, high computational burden, and low interpretability of the final model.

Feature selection is the process used for the reduction of dimension of the data, with the aim of improving the results which are recognized. The process involves following three steps: a search procedure for getting the desired results, an evaluation function, and a stopping criterion. We have considered five standard univariate measures, namely Chi-Square (χ_2), ReliefF, Mutual Information, and Symmetrical Uncertainty.

The Chi-Square approach implements a discretization algo- rithm based on statistics [4]. Each feature values present in given data represent an interval. The values are first sorted. Then the class relative frequencies of adjacent intervals are assessed using Chi-square statistics. For each couple of ad- jacent intervals the frequency similarities are checked and if they are above a given threshold, they are merged.

The ReliefF measure is an instance-based learning. It is an approach which is used to assign a weight to each feature.

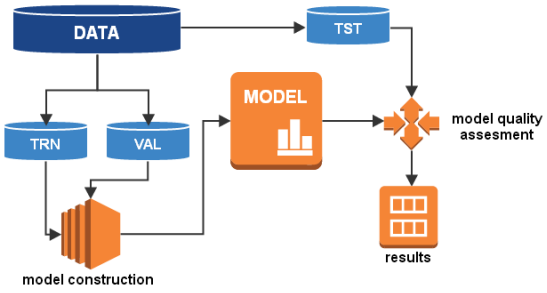

Figure 1 : Classification Model

These weights help in representing its relevance with respect to the target concept [5]. For each sample the nearest neighbor of the same class (nearest hit) and the nearest neighbor of a different class (nearest miss) are found. If the given feature receives a high weight, it takes different values for instances from different classes and similar values for instances belong- ing to the same class.

Mutual Information is an entropy-based measure. The entropy of a random variable is a function which identifies the "unpredictability" of a random variable [6], [7]. Mutual information is a quantity that is measured between two random variables that

are sampled simultaneously to define a relationship between them. It measures how much information is communicated, in one random variable about another.

The Symmetrical Uncertainty has been defined in such a way to compensate the bias towards the attributes taking more values. This is achieved by normalizing the values to the range [0, 1] [8].

Classification is the task of learning a target function "f "that maps each feature set "x "to one of the predefined class labels "y "The target function is also known as a classification model. The input data for a classification task is a collection of records/samples from a dataset [9]. Data classification is a two-step process. In the first step, a model is built describing a predetermined set of data classes or concepts. The model is constructed by analyzing database tuples described by features. Each tuple is assumed to belong to a predefined class, as determined by one of the features, called the class label feature. In the context of classification, data tuples are also referred to as samples. The data tuples analyzed to build the model collectively form the training dataset. The individual tuples making up the training set are referred to as training samples and are randomly selected from the sample population. In the second step, the model is used for classification. The predictive accuracy of the model is estimated, on a given test set is the percentage of test set samples that are correctly classified by the model. Figure 1 represents a classification model.

There are different types of classifiers available for classifying the dataset. Following are some of the popularly used classifiers:

- Decision Tree Induction Classifier
- K-Nearest Neighbors Classifier
- Naive Bayesian Classifier
- Support Vector Machine Classifier

Decision tree classification is the learning of decision trees from class labelled training tuples [10]. A decision tree is a flowchart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node holds a class label. They are robust in nature, therefore, they perform well even if its assumptions are somewhat violated by the true model from which the data were generated. Decision trees perform well with large data in a short time. Consequently, practical decision-tree learning algorithms are based on heuristic algorithms such as the greedy algorithm where locally optimal decisions are made at each node. Such algorithms cannot guarantee to return the globally optimal decision tree.

The k-Nearest Neighbor's algorithm (k-NN) is a method for classifying objects based on closest training examples in the feature space. K-NN is a type of instance-based learning, or lazy learning [11]. It can also be used for regression. The k-nearest neighbor algorithm is amongst the simplest of all machine-learning algorithms. The space is partitioned into regions by locations and labels of the training samples. A point in the space is assigned to the class c if it is the most frequent class label among the k nearest training samples. Usually Euclidean distance is used as the distance metric, however this will only work with numerical values. In cases such as text classification another metric, such as the overlap metric (or Hamming distance) can be used.

The Bayesian Classification [12] represents a supervised learning method as well as a statistical method for classification. It assumes an underlying probabilistic model and allows capturing uncertainty about the model in a principled way by determining probabilities of the outcomes. It can solve diagnostic and predictive problems. It is based on the Bayesian theorem. It is particularly suited when the dimensionality of the inputs is high. Parameter estimation for naive Bayes models

uses the method of maximum likelihood. In spite of over-simplified assumptions, it often performs better in many complex real world situations. One of the advantages of Naive Bayes classifier is that it requires a small amount of training data to estimate Support Vector Machines.

SVM belong to a family of generalized linear classifiers and can be interpreted as an extension of the perceptron [9]. SVMs are a group of supervised learning methods that can be applied to classification or regression. It is primarily a two class classifier. SVMs can efficiently perform non-linear classification using what is called the kernel function; indirectly map their inputs into high-dimensional feature spaces. It can also solve the multiclass problem with the help of kernel methods and kernel function. It aims to maximize the width of the margin between classes, that is, the vacant area between the decision boundary and the nearest training pattern. The basic idea of SVM classifier is to choose the hyperplane that has maximum margin.

Logistic regression (LR) is a supervised learning which was introduced in 1958 by David Cox [13]. LR is used while the output variable is categorical and input variable is either discrete or continuous . LR estimates the parameters and predicts the probability of output variable (cancer vs. control) based on the input variables and choosing the cutoff point values. If the probability value is greater than the cutoff point, it belongs to one class and vice-versa [14].

Random forest (RF) is an ensemble learning method for regression and classification in machine learning which involves the construction of multiple DTs via bootstrap aggregation [15]. RF classifies the trees based on the prediction of the tree structure. The main advantage of RF is that it is a better fit for the categorical data after obtaining the final solution in the majority voting system, where result of each tree is judged.

II. RELATED WORK

Numerous studies have revealed that most genes present in DNA (Deoxyribonucleic Acid) Microarray datasets are not pertinent in the accurate diagnosis of different diseases [16]. To avoid the problem of the curse of dimensionality, feature selection (which is sometimes called gene selection) is the technique employed to seek the most informative genes that can increase the diagnosis and predictive accuracy of diseases. Therefore, feature selection is one of the exciting research areas in the domain of data mining, pattern recognition, machine learning, statistics, and bioinformatics. The main advantage of feature selection is to increase the classification accuracy performance of the classification algorithm through the removal of redundancy and irrelevant features in the microarray dataset. Feature selection (FS) methods are categorized into four (4) approaches, namely filter, wrapper, embedded, and hybrid approaches [17].

Filter Method: The Filter method accomplish the FS task before the classification of the data. Filter methods estimate the goodness of feature by observing only the basic data characteristics, in which (this) typically a single feature or a subset of features is evaluated against the class label. Classical filter methods are usually applied to microarray data, such as Correlation Feature Selection (CFS), Fast Correlation-Based Filter (FCBF), ReliefF, or The mRMR (minimum Redundancy Maximum Relevance).

Wrapper Method: Wrapper methods make use of a certain machine-learning algorithm to select a subset of features. The filter method is very fast in computation and low accuracy while wrapper approaches have better accuracy performance with less computation rate. In a domain with large microarray datasets, the wrapper approach has not received the same amount of attention as the filter methods, due to its high computational cost. As the number of features grows, the space of feature subsets grows exponentially.

This is something that becomes a critical aspect when ten thousand features are considered. Furthermore, they have the risk of overfitting due to the small sample size of microarray data. As a result, the wrapper approach has been largely avoided in the literature.

Hybrid Method: A Hybrid method is a novel approach that tries to seek the benefits of both filter and wrapper techniques. The idea behind the hybrid approach is that the filter approach is used to remove irrelevant features (dimensionality reduction) from the original dataset. Then, the wrapper technique is employed to find the best feature subset from the selected feature pool. This approach speeds up feature selection as the filter technique swiftly reduces the irrelevant features from the dataset.

Embedded Method: The main disadvantage of the filter approach is the fact that it does not interact with the classifier, as a result, worse performance than those obtained with wrappers. However, the wrapper model comes with a high computational cost, which is particularly aggravated by the high dimensionality of microarray data. An intermediate solution for researchers is the use of embedded methods, which use the core of the classifier to establish criteria to rank features.

As mentioned before FS can be grouped into three approaches, that is filter techniques, wrapper techniques, and hybrid techniques, and embedded techniques. Table 1 gives a brief review of existing work on, filter techniques, wrapper techniques, hybrid techniques for feature selection methods on microarray dataset for cancer classification and discusses some major issues or limitations of their work.

Chandra et al. [18] proposed a novel and efficient feature selection approach termed ERGS (Effective Range based Gene Selection). The governing principle for the ERGS algorithm is based on the fact that a feature should be given higher weightage if it discriminates the classes. ERGS algorithm is based on statistically defined effective ranges of each class for a given feature. The main drawback of the algorithm was only Binary class dataset was used and the approach may not likely cope with the multi-class dataset.

Mao et al. [19] proposed a statistic derived from the regression coefficients in a series - partial least squares discriminant analysis (PLS-DA) model is employed to evaluate the significance of the genes. The method used multiple linear regression (MLR) classification model is used for the classification of selected genes. The only limitation observed was all datasets used were binary class data, the approach may not cope with multi-class dataset.

Santos et al. [20] used the ensemble FS algorithms that consist of the following: Information Gain, Gain Ratio, Sym- metrical Uncertainty, and Chi-square. They evaluated selected features with SVM, Bagging using the RPART function (BAG), Random Forest (RF) and Naive Bayes (NB) learning algorithms. The drawback of the paper was the algorithm was tested only on breast dataset.

A new fast feature selection approach based on multiple SVDD on the multi-class microarray data was developed by Cao et al. [21]. The method used recursive feature elimination (RFE) algorithm was repeatedly used to eliminate irrelevant features. Their proposed approach was tagged with multiple SVDDRFE (MSVDDRFE). They employed KNN and SVM classifiers to evaluate the performance of their approach. The limitation of the algorithm was the result of the lung cancer dataset was satisfying not and SVDD is time-consuming for gene selection.

Mohammadi et al. [22] proposed a filter-based feature selection method via Maximum–Minimum Correntropy Criterion and SVM with the linear kernel as classifier. It was used with 25 different datasets. The time, accuracy and stability of feature selected were the metrics

used for performance evaluation of the approach. The only drawback of the algorithm was all datasets used were binary class data, the approach may not cope with multi-class data

He et al. [23] in his work, proposed a class imbalance-aware Relief algorithm, called imRelief. This algorithm corrects the bias towards the majority classes. It avoids the influence of the majority classes for estimating the weights of features. The drawback of the algorithm was it used KNN classifier as the only classifier to measure the performance of the algorithm.

III. EXPERIMENTAL EVALUATION

The two main objectives of performing feature selection on the microarray is to find informative genes and class prediction. For the class prediction, there is a requirement for machine learning techniques such as supervised classification. However, if the objective is to find informative genes, the classification performance is ignored and the selected genes have to be dually evaluated. The experimental studies are focused on class prediction, which is an important reason to use feature selection methods in microarray analysis. As an experimental review of existing feature selection methods, four classical feature selection methods widely used by the researchers in this field were chosen to be applied in this study (Mutual Information, ReliefF, χ_2, symmetric uncertainty). We have opted for these four methods because they are used extensively in the literature. To apply these feature selection methods we have considered five widely-used cancer microarray datasets (Colon, Central Nervous System, Leukemia, Lung, and ovarian) table 2 shows a brief description of all 5 datasets. To evaluate the behavior of the feature selection methods after applying a classifier, accuracy measures are used. In order to obtain the classification accuracy, 6 well-known classifiers (Random Forest, Decision tree, support vector machine, KNN, Naive Bayes, and Logistic Regression) were used.

Table 1 : The Dataset Considered for the Experiments

Datasets	Features	Samples	Classes
Colon	2000	62	2
Central Nervous System (CNS)	7129	60	2
Leukemia	7129	72	2
Lung	12600	203	5
Ovarian	15154	253	2

A. About Dataset

Colon is disease in which cancer causing tissues are found in colon tissues. The dataset contains 62 samples out of which 40 tumor biopsies are from colon adenocarcinoma specimens and 22 are from normal biopsies which form healthy parts of the colons of the same patients. The total number of genes to be tested are 2000 [2].

Leukemia is a primary bone marrow disorder. It is malignant neoplasms of hematopoietic stem cells. The dataset contains 72 samples, each represented by 7129 genes [24].

Lung cancer is the most common and cancer. It is characterized by uncontrolled cell growth in the lung tissues. Techniques based on the expression levels of a small number of genes are useful in the early and accurate diagnosis of lung cancer. The lung dataset contains 181 tissue samples, each described by 12,533 genes [25].

Ovarian cancer is caused due to uncontrollable growth of the cells in the ovaries. The lump of tissues are produced due to abnormal growth. It is most common types of cancer in women. The symptoms of the cancer are similar to those of some more common conditions, and this make it not always easy to diagnose. The Ovarian dataset contains 216 samples (100 patients and 116 controls) [26].

Table 2 : Performance of Various Machine Learning Classifiers Colon Dataset

Classifiers	CS		MI		SU		RE	
RF	87.09	20	90	10	88.7	60	85.48	40
DT	79.03	10	80.64	10	86.64	10	83.87	10
SVM	87.09	80	87.09	60	87.09	20	87.99	60
KNN	83.87	40	83.87	20	85.48	10	87.09	100
NB	88.7	20	83.87	10	87.9	10	87.09	100
LR	88.7	100	87.09	60	85.9	10	82.25	20

Table 3 : Performance of Various Machine Learning Classifiers CNS Dataset

Classifiers	CS		MI		SU		RE	
RF	88.33	40	98.61	60	98.6	100	90	40
DT	68.37	10	87.5	20	88.88	10	68.33	10
SVM	91.66	80	99.44	20	97.22	10	91.6	60
KNN	81.66	10	97.22	100	98	100	90	40
NB	80	20	99.22	100	97.22	60	75	60
LR	83	80	97.22	20	97	100	83.33	40

The CNS datasets has 7129 features. The number of samples this dataset are 61. The tumorous causing tissues are generated in the central nervous system of the body.

B. Analysis of Results

As mention above for the experiment, 5 different cancer datasets are used with four(4) feature selection methods and six different classifiers to evaluate the accuracy of the dataset without feature selection and with feature selection. The re- quired number of genes selected cannot be determined using a common standard, but several hundred of genes are verified to be sufficient to achieve high accuracy [27]. Therefore, different numbers of genes are selected for different filters with the number of genes ranging from 10 to 100. The table shows the result of 6 classifiers with 4 feature selection methods.

For the colon dataset, Na¨ıve baise classifier gives an accuracy of 88.7% with only 20 features using Chi-Square (CS) feature selection method, Random Forest classifier gives an accuracy of 90% with only 10 representative features using mutual Information (MI) as feature selection method, Random Forest classifier again give the accuracy of 88.7% using Symmetric Uncertainty (SU) as a feature selection method, to compare with ReliefF feature selection method SVM provides 87.99 % accuracy with only 60 topmost features.

For the CNS dataset SVM provide 91.66% accuracy with the cs feature selection method with only 80 features, for MI as feature selection method again SVM give99.44% accuracy with only 20 features. SU feature selection method random forest gives 98.6% accuracy with 100 top-ranked features, 91.6% accuracy with 60 features for SVM classifier using ReliefF feature selection method.

For Leukemia dataset, random forest using CS as feature selection method gives 98.61% accuracy with 80 features, MI as feature selection method 96.65% accuracy with only 10 feature provided by logistic regression classifier, for SU and ReliefF feature selection method 98.61% and 85% accuracy is provided by random forest and SVM classifier respectively for 80 topmost features.

Table 4: Performance of Various Machine Learning Classifiers Leukemia Dataset

Classifiers	CS		MI		SU		RE	
RF	94.08	80	99.6	100	92.11	20	94.08	60
DT	84.72	60	97.23	10	88.74	20	89.16	60
SVM	94.58	100	100	40	94.58	80	95.56	40
KNN	93.1	40	99.6	40	92.61	80	93.1	80
NB	93.1	100	98.41	60	90.14	20	86.78	40
LR	93.1	100	99.6	60	93.1	20	96.62	80

Table 5 : Performance of Various Machine Learning Classifiers Lung Dataset

Classifiers	CS		MI		SU		RE	
RF	94.08	80	99.6	100	92.11	20	94.08	60
DT	84.72	60	97.23	10	88.74	20	89.16	60
SVM	94.58	100	100	40	94.58	80	95.56	40
KNN	93.1	40	99.6	40	92.61	80	93.1	80
NB	93.1	100	98.41	60	90.14	20	86.78	40
LR	93.1	100	99.6	60	93.1	20	96.62	80

For Lung dataset SVM classifier provides accuracy of 94.58%, 100%,94.58%, 95.56% for four feature selection methods CS,MI, SU and ReliefF with 100,40,80,40 to most features respectively.

For Ovarian dataset CS feature selection method give an accuracy of 99.6% for random forest classifier with 80 features, among all classifier SVM give 87.09%, 100%, and 100% accuracy for an ovarian dataset using MI, SU and ReliefF feature selection method respectively for 60, 80,80 topmost features.

Finally, Table (VII) summarizes the best result obtained for each microarray dataset by applying the feature-selection techniques considered in this study. For comparison purposes, the table also shows the best results obtained by using the whole feature set.

IV. BIOLOGICAL SIGNIFICANCE

Feature selection methods help identify important genes associated with cancer by removing irrelevant and redundant

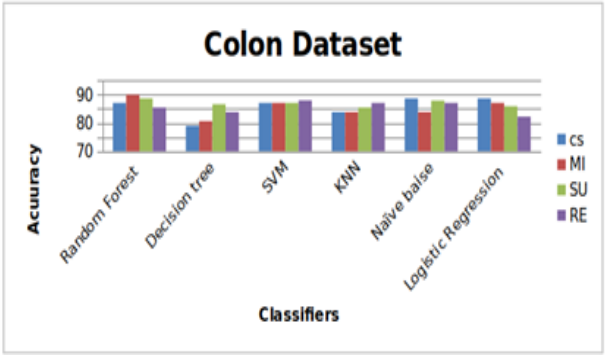

Figure 2: Performance of various Machine learning classifiers on Colon dataset

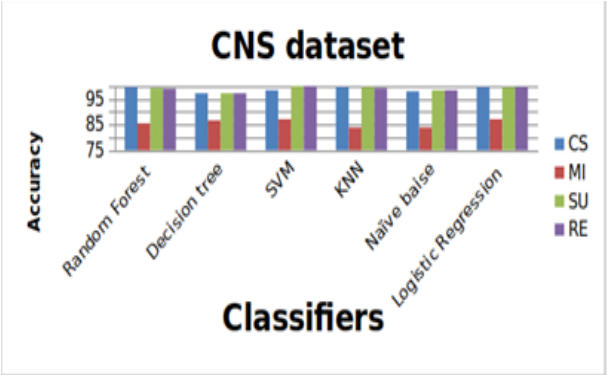

Figure 3: Performance of various Machine learning classifiers on CNS dataset

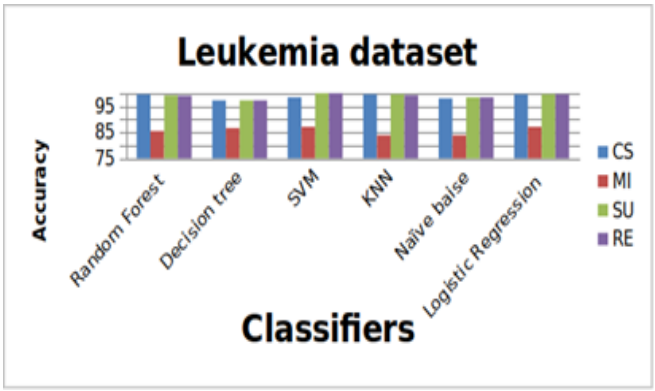

Figure 4: Performance of various Machine learning classifiers on Leukemia dataset

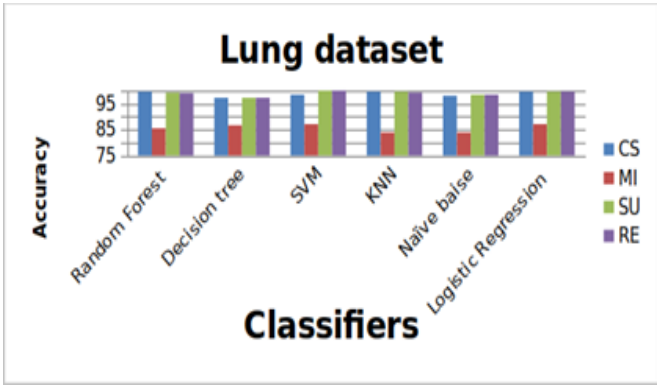

Figure 5: Performance of various Machine learning classifiers on Lung dataset

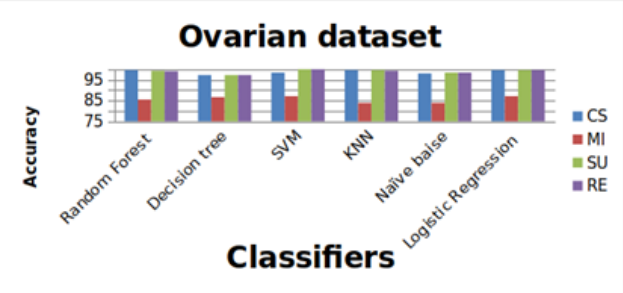

Figure 6. Performance of various Machine learning classifiers on Ovarian dataset

features from high-dimensional microarray datasets. By selecting informative genes and eliminating irrelevant features, the proposed methods improve classification accuracy and support the discovery of potential biomarkers for cancer diagnosis and prognosis. The selected genes obtained through feature selection may serve as important biomarkers for cancer diagnosis and prognosis. Genes such as ANXA1, TRPS1, TBX3, FOXA1, and CAV1 are known to be associated with tumor progression, cell proliferation, and cancer development [28]. The identification of such biologically relevant genes demonstrates that the applied feature selection methods can effectively capture meaningful genetic information from microarray datasets, thereby improving cancer classification and supporting bioinformatics research.

V. CONCLUSION

It is interesting to note that in all the experiments the random forest and logistic Regression classifier provided the best results without feature selection. This outcome is in good accordance with the theory since it is known that such a classifier performs very well when the feature number is very high. The data in Table (VII) also show that the best performance with feature selection was provided by SVM classifiers. As SVM classifier performance is better when the number of features are less. It should be noted, however, that the best results provided by the classification schemes, namely DT, RF, LR, KNN, and NB are slightly lower, but obtained by selecting on average smaller number of features.

REFERENCES :

[1] S. M. Alladi, P. Shinde Santosh, V. Ravi, U. S. Murthy, Colon cancer prediction with genetic profiles using intelligent techniques, Bioinformation 3 (3) (2008) 130.

[2] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, A. J. Levine, Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, Proceedings of the National Academy of Sciences 96 (12) (1999) 6745–6750.

[3] H. M. Zawbaa, E. Emary, C. Grosan, V. Snasel, Large-dimensionality small-instance set feature selection: a hybrid

bio-inspired heuristic approach, Swarm and Evolutionary Computation 42 (2018) 29–42.

[4] H. Liu, R. Setiono, Chi2: Feature selection and discretization of numeric attributes, in: Proceedings of 7th IEEE International Conference on Tools with Artificial Intelligence, IEEE, 1995, pp. 388–391.

[5] I. Kononenko, Estimating attributes: Analysis and extensions of relief, in: European conference on machine learning, Springer, 1994, pp. 171 182.

[6] B. Gierlichs, L. Batina, P. Tuyls, B. Preneel, Mutual information analysis, in: International Workshop on Cryptographic Hardware and Embedded Systems, Springer, 2008, pp. 426–442.

[7] A. Kraskov, H. St¨ogbauer, P. Grassberger, Estimating mutual information, Physical review E 69 (6) (2004) 066138.

[8] M. A. Hall, Correlation-based feature selection for machine learning (1999).

[9] K. S. Durgesh, B. Lekha, Data classification using support vector machine, Journal of theoretical and applied information technology 12 (1) (2010) 1–7.

[10] S. Bhandari, N. Agrawal, N. S. Parande, Design a binary neural network classifier algorithm with parallel training in hidden layer.

[11] P. Mewada, J. Patil, Performance analysis of k-nn on high dimensional datasets, International Journal of Computer Applications 975 (2011) 8887.

[12] K. P. Murphy, et al., Naive bayes classifiers, University of British Columbia 18 (60) (2006).

[13] T. Sapatinas, Discriminant analysis and statistical pattern recognition (2005).

[14] Q. Shen, W.-M. Shi, W. Kong, Hybrid particle swarm optimization and tabu search approach for selecting genes for tumor classification using gene expression data, Computational Biology and Chemistry 32 (1) (2008) 53–60.

[15] D. K. Slonim, From patterns to pathways: gene expression data analysis comes of age, Nature genetics 32 (4) (2002) 502–508.

[16] O. A. Alomari, A. T. Khader, M. A. Al-Betar, L. M. Abualigah, Mrmrba: a hybrid gene selection algorithm for cancer classification, J Theor Appl Inf Technol 95 (12) (2017) 2610–2618.

[17] L. Rangarajan, et al., Bi-level dimensionality reduction methods using feature selection and feature extraction, International Journal of Computer Applications 4 (2) (2010) 33 –38.

[18] B. Chandra, M. Gupta, An efficient statistical feature selection approach for classification of gene expression data, Journal of biomedical informatics 44 (4) (2011) 529–535.

[19] Z. Mao, W. Cai, X. Shao, Selecting significant genes by randomization test for cancer classification using gene expression data, Journal of biomedical

informatics 46 (4) (2013) 594–601.

[20] V. Santos, N. Datia, M. Pato, Ensemble feature ranking applied to medical data, Procedia Technology 17 (2014) 223–230.

[21] J. Cao, L. Zhang, B. Wang, F. Li, J. Yang, A fast gene selection method for multi-cancer classification using multiple support vector data description, Journal of biomedical informatics 53 (2015) 381–389.

[22] M. Mohammadi, H. S. Noghabi, G. A. Hodtani, H. R. Mashhadi, Robust and stable gene selection via maximum–minimum correntropy criterion, Genomics 107 (2-3) (2016) 83–87.

[23] Y. He, J. Zhou, Y. Lin, T. Zhu, A class imbalance-aware relief algorithm for the classification of tumors using microarray gene expression data, Computational biology and chemistry 80 (2019) 121–127.

[24] 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, et al., Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, science 286 (5439) (1999) 531–537.

[25] G. J. Gordon, R. V. Jensen, L.-L. Hsiao, S. R. Gullans, J. E. Blumen stock, S. Ramaswamy, W. G. Richards, D. J. Sugarbaker, R. Bueno, Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma, Cancer research 62 (17) (2002) 4963–4967.

[26] E. F. Petricoin III, A. M. Ardekani, B. A. Hitt, P. J. Levine, V. A. Fusaro, S. M. Steinberg, G. B. Mills, C. Simone, D. A. Fishman, E. C. Kohn, et al., Use of proteomic patterns in serum to identify ovarian cance The lancer 359 (9306) (2002) 572-577.

[27] H. Vural, A. Subas¸ı, Data-mining techniques to classify microarray gene expression data using gene selection by svd and information gain, Modeling of Artificial Intelligence (2) (2015) 171–182.

[28] Genecards: The human gene database, https://www.genecards.org/, accessed: 2024-10-25 (2016)

* * * * *