Feature Selection of Gene expression data for Cancer Classification using SCF with SVM

Kavitha K.R, Avani Prakasan, Dhrishya P.J
Department of Computer Science and Applications Amrita Vishwa Vidyapeetham Amrithapuri, India
kavithakr@am.amritha.edu,avaniprakasan111@gmail.com,dhrishya000001@gmail.com

Abstract—**Feature Selection is the technique used to select the features from microarray dataset. The selected feature must provide high accuracy to the intended classifier. An ordinary microarray dataset will have the characteristics such as high dimensionality limited sample and a large amount of noisy data. This basic features of microarray dataset will reduce the classification accuracy and elevate the run time of the proposed algorithm. To overcome this problem, dimensionality reduction techniques are deployed on the proposed dataset. Different dimensionality reduction techniques are available and they can be mainly categorised as feature selection and feature extraction. This research work focuses mainly on the filter based feature selection method. The proposed filter based combination method for performing dimensionality reduction is named as Feature selection of Gene expression data for Cancer Classification by using Score based Criteria Fusion (SCF) with SVM. The primary aim of the proposed research work is to minimize the classification time and make significant progression in the accuracy of algorithm.**

Index Terms—Gene expression data, Feature selection, Score-based method, Classification.

I. INTRODUCTION

Cancer gene selection and classification is one of the impotent research areas in the modern world. Many literature work has been published based on this topic. The microarray processing techniques have helped to process expressions of large number of genes concurrently. Feature selection or gene selection is the technique of selecting features that can elevate the accuracy of classifier. The microarray dataset has a huge number of features and has a limited sample this is the reason for which analysis of dataset is difficult. To avoid this problems we have to remove the unwanted features or noisy data by dimensionality reduction technique. Machine learning algorithm might be affected by this nosy data. So maximum reduction in the nosy data can improve the accuracy of the algorithm and it will also help in avoiding the unwanted interpretation during the classification. Common noise can be classified as attribute noise and class noise. Error in the attribute or feature value will generate attribute noise. Wrongly calculated variables or missing values will be resulted in such type of error. Class noise is caused by samples that are labelled reside in multiple class or miss classifications.

According to WHO the death cause due to cancer was increasing every year.Their are about more than 100 types of cancer that casing death of 8 million people across the world . The research in this field revels that the death cause due to cancer will increase 14 million in next two decades. The abnormal growth of of cell in the human body that affect other cells and organs and damaging them are called cancer. Cancer can be categorized as Carcinoma, Sarcoma, Myeloma, Leukemia, Lymphoma, and Mixed Types [1]. The number of cancer patients in the world is rapidly rising in every year. So research in the field of feature selection and classification of micro array dataset for cancer prediction is very relevant in modern world.

The most impotent factors that have to consider for choosing an algorithm for cancer prediction are accuracy and computational time. We need a classifier with maximum accuracy and minimum run time of the algorithm. To archive above factors, the process of feature selection should be performed on the proposed dataset.

There are mainly two categories of dimensionality reduction technique feature selection and feature extraction [2]. Feature selection is the method used for selecting desired feature from the existing dataset by eliminating noisy or repeated features. The attribute selection method does not make any modification on attributes. But in feature extraction, new attributes are derived from the current attributes. So feature modification is performed in this cause and selects the desired set of features. The proposed research focuses on filter based attribute identification method, which means the feature rank is assigned by some other method and depending on that rank, the desired set of attributes will be identified. Here, Symmetric Uncertainty (SU) and Relief(R) are used along with SVM to perform feature selection. The informative features are identified by all the above methods in an entirely different way. The combination of SVM, Relief, Su returns most informative attributes only, also help in rising the accuracy of classifier.

Classification is the approach that is used to foretell the class label of unknown data. A good classifier can be developed mainly by two steps training and testing. During training Supervised learning method is used that means a set of input data and corresponding out are provided and their by develop a model. This developed model will be used for testing. During testing an unlabelled data is provided and

978-1-7281-6387-1/20/$31.00 ©2020 IEEE

based on the trained information it will predict the class label of the feature. Different classification methods are used. The accuracy of classifier in completely depend up on type of the data that we provide as input. If we provide input data with unwanted attributes the accuracy of classifier get diminished. So maximum reduction in unwanted data can maximize the accuracy of classifier.

II. LITERATURE REVIEW

Zifa Li, Weibo Xie, Tao Liu studied about modified version of the SVM-RFE called SVM- RFE with a variable step size. Usually, SVM will plot the data in the n-dimensional space and rank the features. Based on the rank RFE reject the features with minimal rank. RFE discard features one by one. It took more computational time. But in SVM -RFE with variable step size method initially set a step size value and it will be reduced to half then the number of features that to be eliminated also reduced to half this process will continue until step size reaches one. This makes it faster than the traditional method [3]

In the paper by Jorge R. Vergara and Pablo Estevez studies the mutual information based on attribute selection. Mutual information indicates how much information one arbitrarily selected variable shares to other one. It indicates how much one feature dependent on others, related to this we can grade the features and can reject the features with minimum rank. Mutual information closely dependent on the entropy of features if two feature which has 0 entropy indicates that they are independent and 1 indicate maximum dependency. MI can be also used to identify the redundancy between the features. MI is one of the efficient redundancy estimation methods.[4].

Baris Senliol, Gokhan Gulgezen, Lei Yu and Zehra Cataltepe studied about the feature selection method FCBF (Fast correlation-based filter). This is a multivariate model of attribute identification method. This approach usually uses symmetric uncertainty to measure the dependency between the feature and feature selection is done by the backward elimination method. The basic idea behind this method is correlation. A good feature which is highly correlated to its class and less related to other features means they are irrelevant. This method eliminates one feature at a time and begins the next iteration with the rest of the features. Relief is another example of a correlation-based feature selection.[5].

Univ Waikato and Hamilton, studied about a type of filter method named as Correlation based attribute selection. Correlation based attribute selection is basically a multivariate model. It identifies the feature that are highly related to class and reduce the dependency between the features. This method rank the features based on identifying the correlation between the features and eliminate the features with lowest rank. According to this method efficient set of features must have more dependency with class and have least mutual dependency. This help to remove redundant set of features [6].

Kamitomioka and Nagaoka studied about another multivariate model of feature selection named as minimum redundancy maximum relevance. This is a attribute selection method pick the most informative attribute,at the same time this method reduces the mutual dependency or reject the repeated features . MRMR type of attribute selection method mostly used by collaboration with other algorithms, helps them to choose top quality attribute and lessen the repeated attributes .[7]

In this paper Changki Lee, Gary Geunbae Lee told about Information Gain (IG). The attribute that has the highest information gain can be chosen because that attribute might have towering information and that feature will be selected. Those features which have the least information gain would have less amount information and such type of features are eliminated and the remaining set of features will be selected and this can be used for classification. The information gain can be measured in range 0 to 1. 1 indicate maximum information gain and 0 indicate minimum information gain [8][9][10].

In most of the feature selection method the algorithms focuses only on single feature selection method which may mostly affect the time and accuracy of classifier. Combination of one or two feature selection method can reduce the time and increase the accuracy. Combination of many weak algorithms perform far better than using a single strong algorithmic method. SCF is a attribute selection method that is the combination of two algorithm such as SU and Relief That gives far better result than using a single feature selection method, or using Su alone or relief alone. But in our analysis we found that we can again make improvisation in this algorithmic approach resulted in the development of new algorithm named as Feature selection of Gene expression data for Cancer Classification using SCF with SVM.

III. PROPOSED METHOD

A. Support Vector Machine

SVM is one of the most widely used and powerful feature selection method, most probably applied on microarray data analysis. SVM is a cracking feature identification method based on kernel technique. Mostly it will be linear kernel [11]. SVM is one of the impotent machine learning algorithms. This algorithm is chiefly used for classification and regression analysis. SVM can be used for removing noisy data from our dataset and thereby we can improve our feature selection method. SVM plot the data in an n-dimensional space. The decision boundary that separate the dataset into two classes the points that are close to the decision boundaries are called support vector which means small changes in the decision boundary can affect the classification. By the analysis of support vector and thereby remove unwanted features. The features which are far away from the decision boundary will have a larger margin and they are the best features that can be used for the classification because small changes cannot alter the classification result.

The human body comprise of thousands of genes among this we pick the features that are accountable for cancer or help in the prediction of cancer in a patient. Extensively used feature selection method is SVM-RFE (support vector machine with Recursive Feature Elimination) [12] as the name implies SVM will rank the features based on the support vector. The data points that are neighbour of hyper plane are called support, slight variation in the data point might affect the classification. After ranking the feature we will remove features with the lowest rank by recursive feature elimination method. But SVM-RFE will takes more time to calculate the weight and remove the features. They only remove one feature at a time. This makes the entire process more time consuming.

B. Symmetric Uncertainty (SU)

Symmetric uncertainty is one of the efficient feature selection methods. Mutual information is the root of SU. Mutual information measure the amount of information one arbitrary variable shared with others. For the calculation of mutual information, we need a good understanding of entropy. entropy is the amount of impurity within the dataset. SU based feature selection work as follows. Su will rank the features based on the mutual information and we will sort them in increasing or decreasing order according to their rank and we will select the feature that all have elevated values. The selected features will be the best among them. The definition of mutual information is as:

$$I(R1,S1)=\sum_{r1,s1}p(r1,s1)log\frac{p(r1,s1)}{p(r1),p(s1)}$$

In the above equation let us take R and S as two random variables. P(r1,s1) denote the joint probability distribution function of R and S. p(r1) is the edge probability distribution of R1,as the same way p(s1) indicate the edge probability distribution of S1 respectively. By using the above formula we calculates the mutual information or dependency between the random variables R and S

$$SU(R1,S1)=\frac{2I(R1,S1)}{H(R1)+H(S1)}$$

Su value of R1 and S1 are calculated using above equation. In this formula edge entropy of arbitrary variable R1 and S1 are represented using H(R1) and H(S1)[13][14][15].

C. Relief

Relief is a uncomplicated, quick and efficient feature selection method that was first introduced by Kera and Rendell. This feature selection method is entirely different from symmetric uncertainty in the sense that symmetric uncertainty will select the features based on their mutual information. But relief will select the attributes regarding the distance between them. The main two terms we have to consider for relief based feature selection is the nearest hit and the nearest miss value. Consider an instant from the dataset and we have to find the nearest neighbors to calculate the weight of the feature. The nearest neighboring value that belongs to the same class is called nearest hit (Htj) and the neighboring value that belong to the different class is called nearest miss value(Msj) after calculating the hit and miss value we will subtract miss value from hit value this process is repeated for several numbers of time and thereby the weight of the feature get elevated. This technique is repeated for the entire sample or some sample. Finally, the weight of the feature is divided by no of features in the dataset and finally, we will get a normalized value within a range of 1 and -1 [16][17]. The wight calculation of instance B as follows, i indicate a instant of the dataset Dt.

$$B[i]=B[i]-\sum_{j=1}^{k}\frac{diff(i,Dt,Htj)}{m.k}+$$

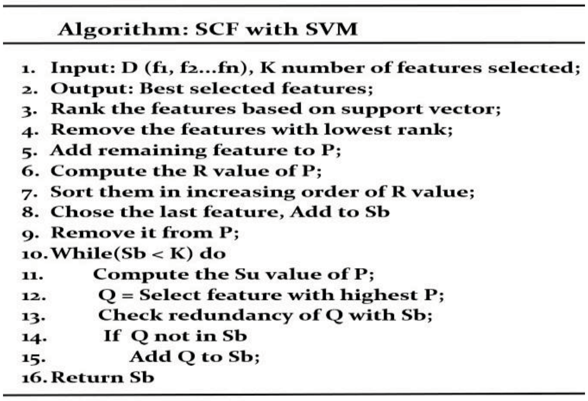
$$\sum_{c!=class(Dt)}\frac{Pr(Y)}{1-Pr(Class(Dt))}.$$

$$.\sum_{j=1}^{k}\frac{diff(i,Dt,Msj(Y)}{m.k}$$

$$diff(i,Dt,Htj)=\frac{|value(i,I0)-value(i,I1)|}{max(i)-min(i)}$$

The prior probability of Y is indicated as Pr(Y) in the above formula. 1-Pr(class(Dt)) denote the cost of miss class probability.

D. Feature selection of Gene expression data for Cancer Classification using SCF with SVM



```
Algorithm: SCF with SVM

Input: D (f1, f2...fn), K number of features selected;
Output: Best selected features;
Rank the features based on support vector;
Remove the features with lowest rank;
Add remaining feature to P;
Compute the R value of P;
Sort them in increasing order of R value;
Chose the last feature, Add to Sb
Remove it from P;
While(Sb < K) do
    Compute the Su value of P;
    Q = Select feature with highest P;
    Check redundancy of Q with Sb;
    If Q not in Sb
        Add Q to Sb;
Return Sb
```

SCF is a feature selection method which focuses on combination of two different algorithmic method such as Su and Relief. In our research we found that adding an initial filer to this algorithm can again improve the accuracy and reduce the computational time. So we used SVM as a initial filter. SCF with SVM is a feature selection method developed by combining three different algorithms named SU, Relief, and SVM all of the above algorithms use different criteria for feature selection. For feature selection or attribute selection, SU uses mutual information as the basic criteria. Relief is a powerful feature selection method that is completely dependent on the distance between the feature, relief positions them in increasing or decreasing order based on the score assigned by calculating the distance, depending upon that we choose the features.

key point behind our concept is, if we use su as a feature selection method or Relief alone as a feature selection method both this algorithmic approach select entirely different set of feature or attributes. During our study about these algorithms, it was clear that each of the algorithm follows absolutely different rules for feature selection and the result from each algorithm or set of selected features from each method was also unique. The separate usage of these algorithm does not help as to reach in our destination. So we found that fusing these algorithms can help us to increase the accuracy of classifier and also we can decline the computational time [18]. But here to improve accuracy we used SVM alone with SCF. SVM will remove easily available noisy data and the output is used for applying combination of SU and Relief

Score based method work as follows Initially we will plot the dataset in N dimensional space and remove all easily identifiable noisy features by the analysis of support vector and remaining are selected. This will be given as the input to the combinations algorithm of SU and Relief then we will set the number of features to be selected and score the features using the relief algorithm (computing R value of feature). Sort the features in the increasing order according to their rank. Select the last feature and add that to sbest (The best set of features to be outputted) then the remaining no of features are again filtered using the symmetric uncertainty and rank them again and chose the feature with the highest rank and check the redundancy between the selected features and if it is non redundant then it will be added to the sbest otherwise it will be removed thereby we will select all the no of desired features. The features selected by this method will be used for classification. The best thing behind this method is, actually it acts like a triple filter, so the combination method of SCF with SVM removes all unwanted features and select best feature set from available features or from the dataset.

The finest feature selection method should have some characteristics such as the selected set of feature does not allow repetition of features or redundancy in features [19]. Another important quality the selected feature should be related to class label. In order to ac-cure these quality here we should apply one more formula. This is to avoid repetition in selected set of features. Let us consider we have L number of features and two sets named as U and V respectively. U is a set which contains L number of features. In beginning V will be null. Here we check the redundancy between the features by considering the mutual information between them. Consider we need h number of features from L number of available feature set. At first we add the top most ranked feature from U to V. Next time we add a feature from U to V we check the mutual information between already added features in V and currently selected feature. If that feature is not available in selected list V then it should be appended to V otherwise get eliminated from U. This procedure perform like an iteration until the size of V reaches h or we selects required number of features from available feature set U[20]..

$$NI(f0,fv) = \frac{I(f0,fv)}{minH(f0),H(fv)}$$

The average normalised value is calculated by the following formula.

$$\frac{1}{|V|}\sum NI(fi,fv)$$

E. System Architecture

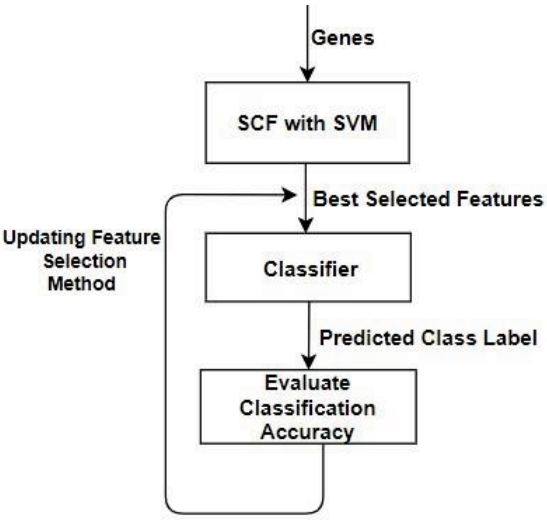


Figure 1. Architectural Diagram

IV. PERFORMANCE ANALYSIS

A. Dataset

For examining our algorithm we are using the Leukaemia dataset. This dataset carries gene expression of large no of genes. This data set holds two classes marked as AML and ALL. The count of AML and ALL classes are 25 and 47 respectively. The sample count of this dataset is 72 and the feature count is 7128. We perform dimensionality reduction in the Leukaemia dataset to eliminate unwanted or noisy features.

B. Experimental Analysis

We started experimenting SCF with SVM algorithm in the microarray dataset and analysed the run time and accuracy of our algorithm by varying the variables no of feature selected (NFS) and neighbour to be considered (NC). Our algorithm consists of mainly two variables. First one is No of features selected this is a variable specify the number of best features to be selected from 7521 features. Next one is neighbouring value to be considered which specifies how many neighbours to be considered while feature selection Our experimental result as follows. We conducted our experiment on microarray dataset and analysed the run time and accuracy of our algorithm by varying the variables no of feature selected and neighbour to be considered. Our experimental result as follows.

At the beginning we set NC value as 5 and assigned 5 different value for NFS our experimental result mentioned in table Table 1.

Table I
ANALYSIS OF RUN TIME AND ACCURACY OF ALGORITHM WITH NC=5

NFS	NC	Run time	Accuracy
10	5	0.050s	0.88
30	5	0.036s	0.96
100	5	0.090s	1.0
200	5	0.081s	1.0
500	5	0.081s	1.0

In the above experiment we got average run time as 0.24s and maximum accuracy 1.

Next we conducted our experiment by setting the NC value as 35 and 70 and assigned 5 dissimilar value for Variable NFS. Our observation is marked on the tables Table 2 and Table 3.

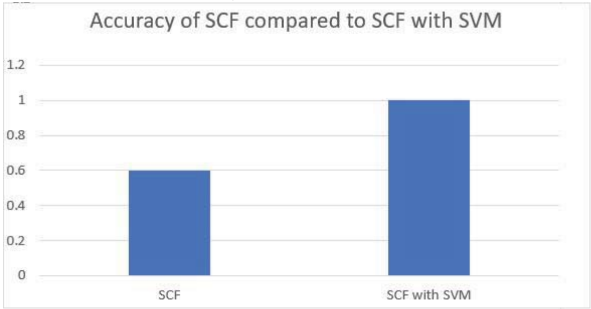
Table II
ANALYSIS OF RUN TIME AND ACCURACY OF ALGORITHM WITH NC= 35

NFS	NC	Run time	Accuracy
10	35	0.050s	0.86
30	35	0.05s	0.98
100	35	0.098s	1.0
200	35	0.100s	1.0
500	35	0.081s	1.0

Table III
ANALYSIS OF RUN TIME AND ACCURACY OF ALGORITHM WITH NC= 70

NFS	NC	Run time	Accuracy
10	70	0.050s	0.88
30	70	0.14s	0.86
100	70	0.048s	1.0
200	70	0.086s	1.0
500	70	0.05s	1.0

From the above experimental result we can conclude that we gets average run time of the algorithm SCF with SVM as 0.6s and maximum accuracy 1. If NFS value is assigned above 100, a maximum accuracy can be obtained.



SCF is a feature selection method which combine two feature selection algorithm Such as Su and Relief. Using SCF method we got an accuracy 0.65 as the experimental result. But by the combination of SCF with SVM gives us the accuracy of 1 and minimum run time 0.6 seconds. Here we used SVM as an initial filter to the combinational algorithm of SCF that helped us to improve the accuracy and reduce the computational time. That is clearly represented in the above graph.

V. CONCLUSION

Feature selection of gene expression data for cancer classification with SVM (SCF with SVM) is an efficient feature selection method. This method is the fusion of three different feature selection method that follows entirely different rules for feature selection. By the fusion of SVM , SU and Relief help us to reduce the feature size or dimensionality of the dataset, at the same time we get the maximum accuracy for our classifier and it also took lest time for computing the result. By examining each of the algorithms individually, we realized that the combination of these can add extra efficiency to the algorithm and can reduce the dimensionality of the dataset more easily and effectively thus we developed SCF with SVM. This algorithm gives us maximum accuracy one and average computational time 0.6 seconds. The combination of SU and Relief alone is not much efficient for the feature selection. But before applying this fusion algorithm, SVM is used as initial filter to make the proposed algorithm very effective for feature selection. Thus, the proposed model has increased the classification accuracy. Using SCF with SVM, the minimum best features are selected, thus time wastage is avoided by training the classifier with large number of feature. From the proposed research and experiment, it is clear that the combination of algorithms will perform better than the individual algorithms.

REFERENCES

[1] Analysis of Cancer Classification of Gene Expression Data A Scientometric Review 1Joseph M. De Guia, 2Madhavi Devaraj, PhD

[2] A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data Zena M. Hira

[3] Zifa Li, Weibo Xie*, Tao . Efficient feature selection and classificationfor microarray data

[4] Jorge R. Vergara · Pablo A. Estevez. A Review of Feature Selection Methods Based on Mutual Information

[5] Baris Senliol1, Gokhan Gulgezen1, Lei Yu2, and Zehra Cataltepe1 1. Fast Correlation Based Filter (FCBF) with a Different Search Strategy. Istanbul Technical University, Computer Engineering Department, Istanbul, Turkey 2Binghamton University, Computer Engineering Department, Binghamton, NY, USA senior, glazen, cataltepe@itu.edu.tr, lyu@cs.binghamton.edu

[6] "Correlation-based feature selection for machine learning," Univ. Waikato, Hamilton, New Zealand, Tech. Rep. 19, Apr. 1999

[7] 1603-1, Kamitomioka, 940-2137, Nagaoka, Japan Yoshiki Mikami Department of Management and Information Systems Science, Nagaoka University of Technology, 1603-1, Kamitomioka, 9402137, Nagaoka, Japan A Minimum Redundancy Maximum Relevance-Based Approach for Multivariate Causality Analysis Yawai Tint Information science and Control Engineering, Graduate School of Engineering, Nagaoka University of Technology,

[8] q Changki Lee *, Gary Geunbae Lee * .Information gain and divergencebased feature selection for machine learning-based text categorization

[9] Feature Selection based on Information Gain B.Azhagusundari, Antony Selvadoss Thanamani

[10] Xiaohui Lin *, Chao Li, Yanhui Zhang, Benzhe Su, Meng Fan, and Hai Wei
Selecting Feature Subsets Based on SVM-RFE and the Overlapping Ratio with Applications in Bioinformatics.

[11] Zifa Li, Weibo Xie*, Tao . Efficient feature selection and classification for microarray data.

[12] Multiple SVM-RFE for Gene Selection in Cancer Classification With Expression Data Kai-Bo Duan, Jagath C. Rajapakse*, Senior Member, IEEE, Haiying Wang, Member, IEEE, and Francisco Azuaje, Senior Member, IEEE

[13] Application of Symmetric Uncertainty and Mutual Information to Dimensionality Reduction of and Classification Hyperspectral Images ELkebir Sarhrouni*, Ahmed Hammouch** and Driss Aboutajdine* [14] Mutual Information between Discrete and Continuous datasets Brian C. Ross

[15] A Feature Subset Selection Technique for High Dimensional Data using Symmetric Uncertainty Bharat Singh, Nidhi Kushwaha, Om Prakash Vyas

[16] RELIEF: Feature Selection Approach Francisca Rosario Head of the Department, Department of Computer Applications Arignar Anna (Arts Science) College, Krishnagiri, Tamil Nadu, India Dr. K. Thangadurai Head of the Department, PG Research Department of Computer Science Government Arts College (Autonomous), Karur, Tamil Nadu, India

[17] Relief-Based Feature Selection: Introduction and Review Ryan J. Urbanowicza, , Melissa Meekerb , William LaCavaa , Randal S. Olsona , Jason H. Moore

[18] A New Filter Feature Selection Based on Criteria Fusion for Gene Microarray Data WENJUN KE1, CHUNXUE WU 1, YAN WU2, AND NEAL N. XIONG3

[19] Redundancy Based Feature Selection for Microarray Data Lei Yu, Huan Liu, Department of Computer Science Engineering

[20] P. A. Estevez, M. Tesmer, C. A. Perez, and J. M. Zurada, “Normalized′ mutual information feature selection,” IEEE Trans. Neural Netw., vol. 20, no. 2, pp. 189–201, Feb. 2009.

[21] Ani, R., Jose, J., Wilson, M., Deepa, O.S. Modified rotation forest ensemble classifier for medical diagnosis in decision support systems (2018) Advances in Intelligent Systems and Computing, 564, pp. 137146.

[22] Kavitha, K.R., Harishankar, U.N., Akhil, M.C. PSO based feature selection of gene for cancer classification using SVM-RFE (2018) 2018 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2018, art. no. 8554429, pp. 1012-1016.[6]

[23] Kavitha, K.R., Ram, A.V., Anandu, S., Karthik, S., Kailas, S., Arjun, N.M. PCA-based gene selection for cancer classification(2018) 2018 IEEE International Conference on Computational Intelligence and Computing Research, ICCIC 2018, art. no. 8782337, . [5]

[24] A comparative study of hybrid feature selection methods using correlation coefficient for microarray data, Arunkumar Chinnaswamy; Ramakrishnan, S

[25] Decision support system for diagnosis and prediction of chronic renal failure using random subspace classification Ani, R., Sasi, G., Sankar, U.R., Deepa, O.S.

[26] Pandian, A. Pasumpon. ”Identification and classification of cancer cells using capsule network with pathological images.” Journal of Artificial Intelligence 1, no. 01 (2019): 37-44.

[27] Manoharan, Samuel. ”Study On Hermitian Graph Wavelets in Feature Detection.” Journal of Soft Computing Paradigm (JSCP) 1, no. 01 (2019): 24-32.