European Journal of Molecular & Clinical Medicine
ISSN 2515-8260 Volume 07, Issue 3, 2020

HYBRID ENSEMBLE FEATURE SELECTION (HEFS) MODEL FOR GENE EXPRESSION MICROARRAY DATA

[1]V. Kalaimani, Assistant Professor, Department of Computer Science (PG),
PSGR Krishnammal College for Women, Coimbatore.
mail id: kalaimani_95@yahoo.co.in
[2]*Dr.R.Umagandhi, Associate Professor and Head, Department of Computer Technology,*
Kongunadu Arts and Science and College, Coimbatore.
mail id: umakongunadu@gmail.com

ABSTRACT: The study of Gene Expression Profiling (GEP) of cells and tissue has become a major tool for discovery in medicine. In GEP, Gene Expression Microarray (GEM) data categorization is a difficult task because of its tremendous quantity of attributes (features) and also the limited size of sample. Feature Selection (FS) was explored to reduce the data dimensionality while maintaining the classifier accuracy. Recently, Swallow Swarm Optimization with Score-Based Criteria Fusion (Optimized SCF) wrapper FS approach has been suggested for predicting the tumors with better efficiency. However, relying upon simple statistical analyses or a unified FS might not increase prediction accuracy. Following the idea behind Ensemble FS (EFS), multiple algorithms are considered in this paper for increasing the robustness of classification. To give a contribution to the field, this Hybrid EFS (HEFS) system is introduced which encompasses different kinds of selection algorithms such as filter by Score-Based Criteria Fusion (SCF) and embedded method by Fuzzy Elephant Herding Optimization (FEHO) and Support Vector Machine- t (SVM-t). The outcomes from those approaches are aggregated via the Weighted Majority Voting (WMV). WMV is a popular and robust strategy to aggregate different algorithms, where each result according to their classification accuracy. The outcome of EFS is an attribute subgroup obtained by concatenating the results of different approaches on various data. It could improve the efficiency of the classifiers such as K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Recursive Neural Networks (RNN), and validate its superiority with four different datasets. Experimental results verify that the HEFS method shows improved results regarding precision, recall, accuracy and Area Under Curve (AUC) when compared to conventional FS methods.

INDEX TERMS: Tumor prediction, Feature selection, SCF, GEM data, FEHO, WMV.

1. INTRODUCTION

GEM data is an advanced technique employed during forecasting and diagnosing the tumors [1]. It offers the ability to simultaneously measure thousands of gene expression values. The GEM data collection is formulated according to 4 main stages: One of which is the extraction and filtering of actual information gathered from the devices in particular repositories. Consequently, such information is partly normalised to avoid distortion and the physiological data contained in relevant repositories are encoded. Finally, data mining techniques are applied to extract the physiological data from encoded information [2].

GEM information may be utilized for predicting the diseases, categorizing the tumor types and identifying the relevant genes [3]. Because GEM has million attributes compared to the sample size, the dimensionality issue has arisen. However, the classifier provides extremely less efficiency while using more features. So, FS technique is applied as a preprocessing task during classification [4]. This is the major approach applied to lessen the dimensionality through choosing the small subgroup of features [5] and time complexity of classifiers with the maximum prediction accuracy.

Many FS techniques are developed for extracting disease-mediated genes [6]. Normally, such techniques are split into filtering, wrapping and embedded-based types [7]. Filtering process relies on the building blocks of the dataset that is separate from the classification model and it uses some assessment rules focused on data analysis to pick a feature subgroup from the raw data. The wrapper approach relies on the classification output to assess the value of attribute subgroups whereas the embedded technique incorporates the benefit of filtering and wrapping approaches, by filtering specific genes by means of a pre-determined classifier [8]. The computation cost of such approaches is significantly lower as the filtering strategies are distinct of the classification model, which makes them ideal for large data analysis. However, wrapper approaches can be more reliable, but they're also overfit.

In earlier centuries, the optimization strategies are focused by many investigators for wrapper methods. Such algorithms are highly efficient to solve the complicated optimization challenges. The most well-known strategies are Genetic Algorithms (GA), Particle Swarm Optimization (PSO), Harmony Search (HS), Ant Colony Optimization (ACO), Simulated Annealing (SA) and Swallow Swarm Optimization (SSO) [9]. The competence of global hunt and convergence value is the foremost important criteria to assess their efficiency. Additionally, a major problem in many of the modern heuristic methods is prevention of local optima. Elephant

Herding Optimization process (EHO) seeks to prevent the capture of local optimums and population growth variations on the basis of higher responses [10].

But no particular algorithm is likely to attempt to guarantee optimum outcomes in statistical output and stable conditions (i.e. reliability of the input data modification), the efficiency of integration processes require the variety of various candidates was increasingly examined. Current and more comprehensive FS methods have been discussed with a view to develop a positive trade-off between forecasting efficiency and stability [11].

The ensemble model was explored [12-13] as an encouraging system to enhance the reliability of the classification algorithm, particularly in the deep and limited conditions where extracting balanced subsets of features is inevitably very challenging [14]. The strategies for choice from ensemble in relevant research can be generally divided into 2 distinct classes [15, 16]: heterogeneous methods involving the effects of advanced FS in similar dataset and homogeneous methods for various editions of the raw data, related to decision trees and wrapping. In order to take advantage of individual strategies, while also addressing their shortcomings, a combination of candidates leads to modern learning strategies.

In this article, HEFS is proposed that encompasses various categories of selection approaches: filtering by SCF and embedded method by FEHO and SVM-t. Then cancer prediction is performed via classifiers such as KNN, SVM and RNN. The major intent of this approach is to select the attributes which are well associated with the label and separate from every other. It gives improved results for cancer dataset. The remaining part of article is emphasized as follows: Section 2 surveys the previous EFS. Section 3 describes the HEFS methodology for choosing and categorizing the gene related attributes. Section 4 illustrates the efficiency of the FS approaches with benchmark datasets. Finally, Section 5 concludes the entire discussion and suggests an extension of this work.

2. LITERATURE SURVEY

Ke et al. [18] developed a SCF with SVM and KNN classifiers for cancer prediction using different public and low-dimensional datasets. The analysis confirmed that it has the ability for discovering number of discriminative attributes which were applied in prediction to classify the cancer.

Albashish et al. [19] suggested a supervised learning for Gleason grading of prostate histology. The primary tissue elements in the images were precisely categorizing into benign or malevolent. Also, the texture characteristics of the images were applied for creating a hybrid method. Further, an enhanced multi-scoring FS was introduced depending on SVM Recursive Feature Elimination (SVM-RFE) and Conditional Mutual Information (CMI) schemes.

Das et al. [20] suggested a sigFeature dependent on SVM for finding the most relevant attributes. The classifier efficiency was increased while using fewer amounts of attributes. Also, GEM enhancement evaluation using the elected attributes by sigFeature was performed. So, it was relatively enhanced compared to other approaches.

Morovvat and Osareh [21] introduced a hybrid FS which combines two approaches. The concept was that candidate attributes were chosen from the raw data using many filters. Then, the candidate attribute group was normalized by precise wrappers. The analysis was done using different GEM datasets to verify its efficiency while using fewer amounts of attributes.

Kavitha and Mahalekshmi [22] recommended enhanced EFS for diagnosing the breast cancer. Initially, Chi-square, Random Forest (RF) and Information Gain (IG) were combined to choose the optimized features. Then, SVM-RFE approach was used for selecting the subgroup of features. Further, classification was achieved by RF, SVM, Linear Discriminate Analysis (LDA), JRip, Recursive Partitioning and Regression Trees (RPART), J48 and Logistic Model Trees (LMT).

Alejandro et al. [23] applied an enhanced FS method for detecting highly significant microRNA and categorizing the cancer according to the consensus on attribute similarity. The dataset was gathered and a meta-analysis was conducted for discovering the most relevant attributes. Then, machine learning algorithms were used for classification task. Xu et al. [24] proposed a Correlation-based FS (CFS) by Neighborhood Mutual Information (NMI) and PSO. An efficient FS approach namely NMICFS-PSO was proposed. Moreover, SVM with leave-one-out cross-validation was used for classification.

Wang et al. [25] proposed a Sort Aggregation-EFS (SA-EFS) method via aggregating Chi-Square, the highest data coefficient and XGBoost to obtain the attribute subset. These were classified by using KNN, RF and XGBoost. Bilen et al. [26] developed an improved method for categorizing GEM of leukemia cancer via choosing the most relevant genes and minimizing the data dimensionality. First, a gene filtration was done through designing an ensemble method using Fisher correlation rank, Wilcoxon rank sum and IG. Then, a modified GA was applied to elect the optimized genes.

III. PROPOSED METHODOLOGY

In the cancer classification, a single FS approach appears likely to ensure maximum efficiency, for both prediction and reliability, so the combinations of various approaches: filtering, wrapper and embedded have been concentrated by analysts. Attribute subset produced by HEFS is employed for categorization according to the selected features. If those two functions are achieved higher than those features or genes are selected for classification. Then classifiers are validated by different GEM datasets. The overall framework of HEFS approach is shown in the Figure 1.

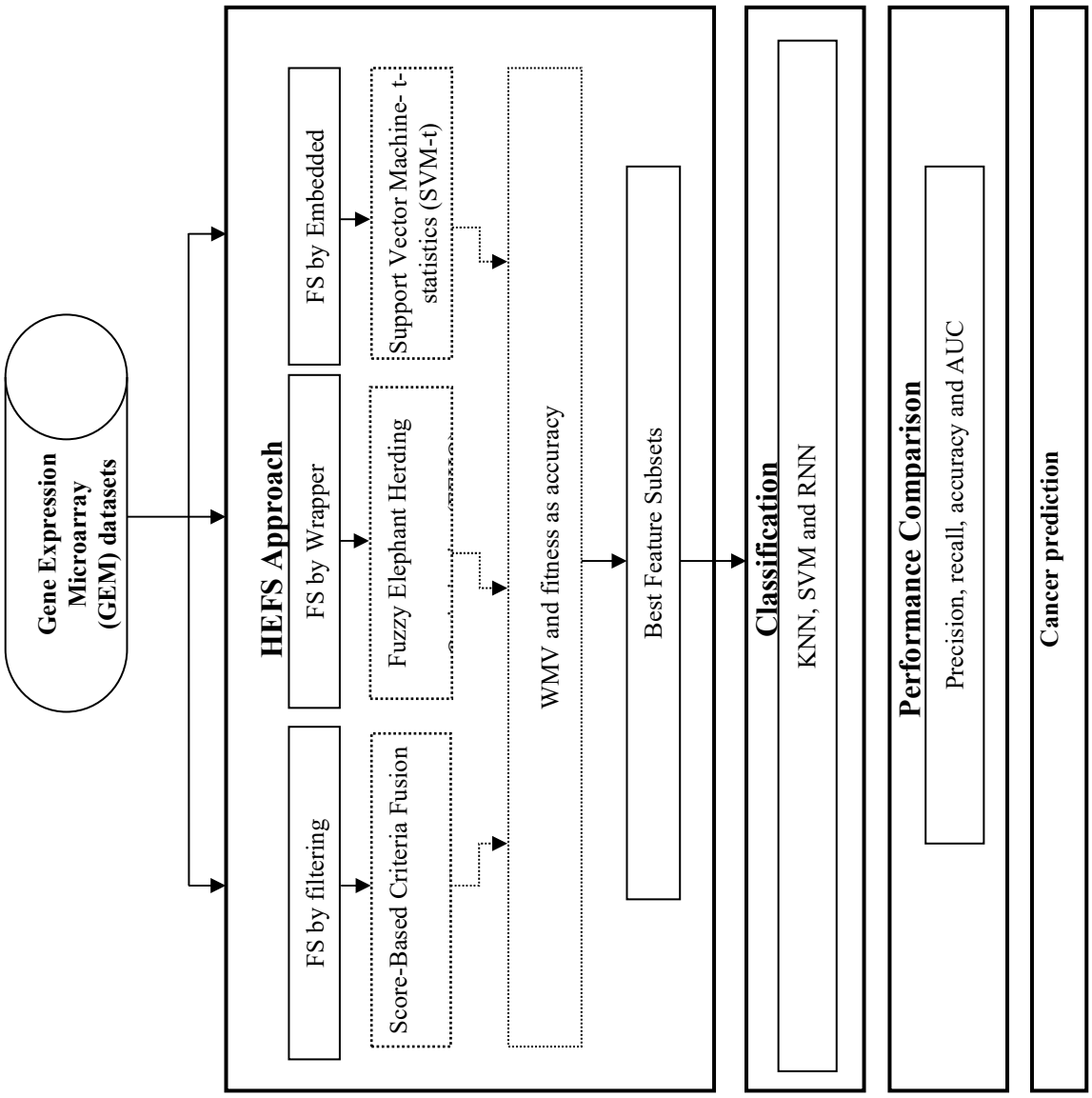

Figure 1. Overall Research Framework of Proposed OSCF-based FS

3.1. HEFS APPROACH

It combines SCF, FEHO and SVM-t algorithms. The outcomes of these algorithms are aggregated via WMV.

3.1.1. FILTERING-BASED FS

It uses statistical techniques for analyzing the correlation between each input and output factors as well as these scores are employed as the basis to choose (filter) those input factors which can be applied in the model. In SCF, the similarity measure involves Symmetrical Uncertainty (SU) and ReliefF, merged via a score-based multi-criteria fusion scheme [27]. By using this scheme, every fundamental criterion creates a score vector consisting values of each attribute and many score vectors are combined into single vector through a fusion strategy [28].

At last, attributes are ranked in accordance with their ranges in the resultant score vector. Besides, the fusion strategy merges 2 score vectors via multiplying the weight value as [18]**:**

$$R_{i,cl} = \mu SU_{i,cl} + (1-\mu)W_i \quad (1)$$

In Eq. (1), the weight value $\mu \in [0,1]$ provides a trade-off between SU and ReliefF, and its range computes the significances of 2 fundamental criteria to the determined similarity. SU is applied for compensating the bias for mutual information as [18]:

$$SU(X,Y) = \frac{2I(X,Y)}{H(X)+H(Y)} \quad (2)$$

In Eq. (2), the SU $\in [0,1]$. If SU=1, the data of single random parameter X is fully classified by the data of other parameter Y, as well as if SU=0 means, X and Y are autonomous. H(X) and H(Y) are edge entropy, $I(X,Y)$ is mutual information [18].

ReliefF is dependent on attribute's capability of differentiating adjacent data for determining the attribute's weight via several iterations. During every iteration, a data R is elected at random and ReliefF explores KNN (closest hit, H) from data label and explores KNN (closest miss, M) from every other label. For an attribute, if the space between R & H is lower than the space between R & M, then the attribute can differentiate data from various labels and its weight have to be high; or else, it is not used and its weight have to be less. For a dataset consisting n attributes, this task is continued n iterations for acquiring the weight W of every attribute [18].

Here, SCF created by SU $\in [0,1$, when the scores created by ReliefF are uncertain. So, score normalization is applied for ReliefF as [18]:

$$u_i' = \frac{u_i - u_{min}}{u_{max} - u_{min}} \quad (3)$$

A group of high-quality attributes must be appropriate to labels & non-appropriate with every other. Here, redundancy [29] is applied for removing the unwanted attributes in the chosen attribute subgroup. It regularizes the shared data for eliminating its bias towards attributes having maximum ranges. This is defined as:

$$NI(f_i, f_s) = \frac{I(f_i, f_s)}{min[H(f_i), H(f_s)]} \quad (4)$$

$$D_i = \frac{1}{|S|}\sum_{f_i, f_s \in S} NI(f_i, f_s) \quad (5)$$

Where |S| denotes the amount of chosen attributes & its reciprocal is applied for balancing similarity with redundancy term. Based on this, a novel SCF approach is built and defined as:

$$G = R_{i,cl} - D_i \quad (6)$$

After all, an appropriate and non-appropriate attribute subgroup is chosen via increasing cost factor (6). Observe that if $\mu = 0$ & $\mu = 1$, Eq. (6) can rewritten as Eq. (7) and Eq. (8) which are the SCF's modification. They are considered by including redundancy into SU and ReliefF.

$$G_1 = W_i - D_i \quad (7)$$

$$G_2 = SU_{i,cl} - D_i \quad (8)$$

Additionally, the SCF's explore policy pursues the incremental forward choice.

3.1.2. WRAPPER FS

Wrapper FS via analyzing the attribute subgroups by machine learning approach which uses a search policy for viewing through the attribute space, analyzing every subgroup according to the efficiency of a

considered approach. It follows a FEHO by evaluating all the possible combinations of features against the evaluation. FEHO is stimulated via the herding characteristic of elephant set [30]. The basic FEHO is explained by the below refined principles [30]:

1) Elephants from various groups, guided by a matriarch, stay together to make the best choice of attributes. For the optimum range of attributes, every group has a specific amount of elephants. For classification purposes, imagine that every group has the same quantity of elephants in order to ensure the optimum range of attributes.

2) The elephant roles in a group are modernized according to their interaction with the matriarch. FEHO process, actions conducted via the optimum range of attributes of the updated function.

3) Male adult elephants abandon their relatives in the ideal range of attributes to stay independently. Expect that a certain percentage of male elephants abandon their groups in the optimum choice of attributes at every iteration. FEHO devises the upgrade method in the best possible range of attributes by a filtering function.

4) The elder female elephant in any group is usually the matriarch. In order to ensure optimum range of attributes, a matriarch is regarded as the most appropriate elephant in the group.

3.1.2.1. Clan Updating Function

Consider that an elephant group (clan) is indicated as c_i. The successive location of any elephant j in the group is modified by,

$$x_{new,ci,j} = x_{ci,j} + \alpha * \left(x_{best,ci} - x_{ci,j}\right) * r \quad (9)$$

In Eq. (9), $x_{new,ci,j}$ is the changed location and $x_{ci,j}$ is the previous location of j in c_i. $x_{best,ci}$ is the matriarch of c_i; it is the best independent elephant in c_i. The scale variable $\alpha \in [0,1]$ computes the matriarch impact of c_i on $x_{ci,j}$. $r \in [0,1]$. It must be observed that $x_{ci,j} = x_{best,ci}$ refers that the matriarch (best elephant) in c_i can't be modified via Eq. (9). This is resolved by modifying the best elephant as:

$$x_{new,ci,j} = x_{center,ci} \times \beta \quad (10)$$

The data from each elephant in c_i is considered for creating the fresh elephants $x_{new,ci,j}$. The mid of c_i, $\left(x_{center,ci}\right)$ is determined for d^{th} size via D estimations where D refers to the overall size as:

$$x_{center,ci,d} = \frac{1}{n_{ci}} \times \sum_{j=1}^{n_{ci}} x_{ci,j,d} \quad (11)$$

Here, $1 \leq d \leq D$ represents the d^{th} size, n_{ci} is the amount of elephants in c_i and $x_{ci,j,d}$ is the d^{th} size of the elephant $x_{ci,j}$.

Algorithm 1. Clan Updating Function

Start

for c_i = 1 to n (each group in elephant populace)

for j = 1 to n (each elephant in c_i)

Renew $x_{ci,j}$ & create $x_{new,ci,j}$ using Eq. (9).

If $x_{ci,j} = x_{best,ci}$ then

Renew $x_{ci,j}$ and create $x_{new,ci,j}$ using Eq. (10).

end if

end for

end for

Terminate

3.1.2.2. Separating Function

In elephant clans, male elephants abandon their families and stay alone with chosen attributes when they hit puberty. Now consider that the specific elephants with worse condition enforce the separating function for every group in an optimal range of attributes to further enhance the searching capacities of the FEHO as:

$$x_{worst,ci} = x_{min} + (x_{max} - x_{min} + 1) * Frand \quad (12)$$

In Eq. (12), x_{max} and x_{min} denote the maximum and the minimum limit, accordingly, for the location of an elephant. $x_{worst,ci}$ indicates the worst elephant in c_i. $Frand \in [0,1]$ defines a fuzzy distribution i.e., the likelihood of a fuzzy value between 0 and 1.

Algorithm 2. Separating Function

Start

for c_i = 1 to n (every group in elephant populace)

Swap the worst elephant in c_i using Eq. (12).

end for

Terminate

For the FEHO, a type of intellectualism technique is applied to preserve the strongest elephants from the misery of group adjustments and separators for an optimum range of attributes. At first the strongest elephant is preserved with the maximum variety of attributes, and the worst is exchanged with the most preserved elephant at the completion of the training cycle with the optimum choice of attributes. This means that the volume of subsequent elephants isn't necessarily weaker than that of the previous populace.

Algorithm 3. FEHO

Initialization

Assign the iteration t = 1

Initialize the populace P of elephants at a random with fuzzy

distribution in the search space

Assign the amount of preserved elephants nKEL, the highest iteration MaxIter, r α, β, n and the amount of elephants

for the c_i^{th} group n_{ci}.

Analyze every independent elephant based on its location

While t <MaxIter

Rank every independent elephant based on their fitness.

Keep the nKEL elephants.

Execute the clan updating function using Algorithm 1.

Execute the separating function using Algorithm 2.

Analyze the populace based on freshly upgraded locations.

Swap the worst elephant with the nKEL ones.

t = t + 1

End while

Obtain the finest result

3.1.3. EMBEDDED FS

In the design of the machine learning model, the embedded methods execute the FS procedure. It is done on the basis of SVM-t facts to pick attributes from the data. In order to construct the maximum separation, hyperplane SVM exploit the data of support vectors and define the labels for each dataset. In order to create the attribute, the SVM-t filtering system involves the most relevant attribute subgroup of the dataset. The basic 2-sample statistics are used to assess the substantial difference between 2 groups. Therefore, the key discrepancy for individual genes can be classified among the closest data with the variance of the tests:

$$|t_j| = \left| \frac{(u_j^+ - u_j^-)}{\sqrt{((s_j^+)^2/n^+) + ((s_j^-)^2/n^-)}} \right| \quad (13)$$

In Eq. (13), n+ (accor., n−) denotes the no. of support vectors designed for label +1 (accor., −1). Compute average u_j^+ (resp.,u_j^-) and standard variance s_j^+ (accor., s_j^-) via the attribute's support vectors of

label +1 (accor., −1) to calculate the range of every attribute. These attributes along with improved the ranges are the attributes having important variation among the two categories. It is perceptive to choose features by way of the improved efficiency.

3.1.4. WMV

Predictability can boost decision-making on these eligible attributes which can offer higher significance to the choice in the voting and can ultimately greatly increase efficiency than SVM. In WMV, each vote is weighted by the prediction accuracy value of the features via classifier that is denoted here Acc. The number of overall voting for a class c_k is recomputed as:

$$T_k = \sum_{i=1}^{M} Acc(A_i) \times F_k(c_i) \quad (14)$$

3.2. FINTESS COMPUTATION

Fitness is computed via merging the classification accuracy and SCF. If both are greater than the features are selected from the GEM dataset. Fitness is determined from the KNN, SVM and RNN. Chosen attributes are given to classifier, and then accuracy will also be considered with fitness along with SCF criteria.

Classification accuracy is defined as a fraction between the numbers of correctly assigned labels and the overall amount of objects to be classified.

IV. RESULTS AND DISCUSSION

To evaluate SA-EFS and HEFS approaches, 4 different GEM datasets are considered.

4.1. DATASET DESCRIPTIONS

Prostate cancer

It is acquired at http://www.gems-system.org/. It comprises 102 data in which 50 are prostate tumors and 52 are typical. Every data has 10509 genes.

SRBCT data

It is obtained at http://www.biolab.si/supp/bi-cancer/projections/info/ SRBCT.html. It comprises 83 data and each has 2308 genes. The tumors are Burkett's Lymphoma (BL), the Ewing family of tumors (EWS), Neuro Blastoma (NB) and Rhabdo Myo Sarcoma (RMS). Among 83 data, 63 are used for training and 20 are used for testing. The training set encompasses 8, 23, 12 and 20 data of BL, EWS, NB and RMS, accordingly. The test set encompasses 3, 6, 6 and 5 data of BL, EWS, NB and RMS, accordingly.

Leukemia

It holds 7129 genes in use over 72 models. It comprises 72 data, 25 data of Acute Myeloid Leukemia (AML) and 47 data of Acute Lymphoblastic Leukemia (ALL). A basis of the GEM values is obtained from 63 bone marrow data and 9 peripheral blood data. It is obtained at http://cilab.ujn.edu.cn/datasets.html.

Lymphoma

It has 2 different tumor subcategories: germinal center B cell-like DLCL and activated B cell-like DLCL. It comprises 24 data of germinal center B-like and 23 data of activated B-like DLCLs. It holds 42 data acquired from Diffuse Large B-cell lymphoma (DLBCL), 9 data from Follicular Lymphoma (FL) and 11 data from Chronic Lymphocytic Leukemia (CLL). It is obtained at http://csse.szu.edu.cn/staff/zhuzx/Datasets.html. An entire dataset holds 4026 genes. Also, it contains some missing values which are imputed using KNN. Table 1 briefly summarizes these datasets.

Table 1. Gene Datasets Characteristics

DATASETS	#GENE	#INSTANCE	#CLASS
Leukemia	7129	72	2
Lymphoma	4026	62	3
Prostate cancer	10509	102	2
SRBCT	2308	83	4

4.2. PERFORMANCE METRICS

To analyze the efficiency of proposed FS approach, KNN, SVM and RNN classifiers are used. The effectiveness of HEFS is compared with standard EFS with respect to Prostate cancer, SRBCT, Leukemia and Lymphoma datasets via MATLAB environment. Such approaches are assessed using the classification metrics like precision, recall, accuracy and Area Under Curve (AUC). Four effective measures such as True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN) are calculated from confusion matrix result in Table 2.

Precision or positive predictive value is computed as:

$$\text{Precision}=TP/(TP+FP) \quad (15)$$

Recall or sensitivity is a ratio of TP to the overall of TP and FN as:

$$\text{Recall}=TP/(TP+FN) \quad (16)$$

Accuracy works by considering the amount of exactly classified samples to the fraction of the overall amount of test samples as:

$$\text{Accuracy}=(TP+TN)/(TP+TN+FP+FN) \quad (17)$$

The AUC is the graphical plot of the TP Rate (TPR) versus the FP Rate (FPR).

Table 2. Confusion Matrix

Total population		Predicted class	
		Positive	Negative
Actual class	Positive	TP	FN
	Negative	FP	TN

4.3. *COMPARISON OF PERFORMANCE METRICSVS.METHODS*

To analyze the efficiency of SA-EFS and HEFS, KNN, SVM & RNN classifiers are used.

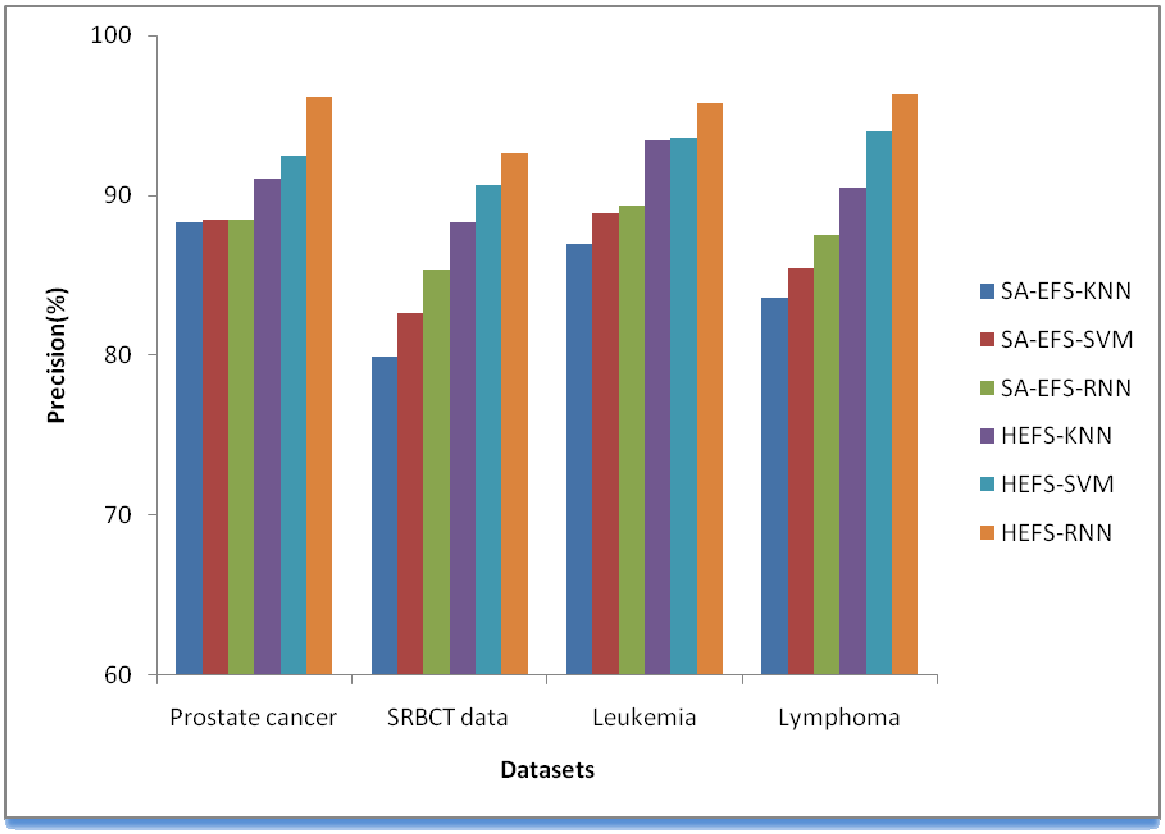

Figure.2 Comparison of Precision vs. Different Datasets

Figure 2 shows the precision for HEFS and SA-EFS using KNN, SVM and RNN. From this analysis, it is analyzed that the HEFS can achieve better precision than other existing EFS. From three classifiers, HEFS–RNN gives the highest precision of 96.2% for prostate cancer dataset which is 7.74% higher than SA-EFS-RNN (See Table 3). Similarly, the precision value of HEFS–SVM for prostate cancer dataset is 2.72% higher than SA-EFS-SVM and HEFS–KNN is 4.04% higher than SA-EFS–KNN (See Table 3).

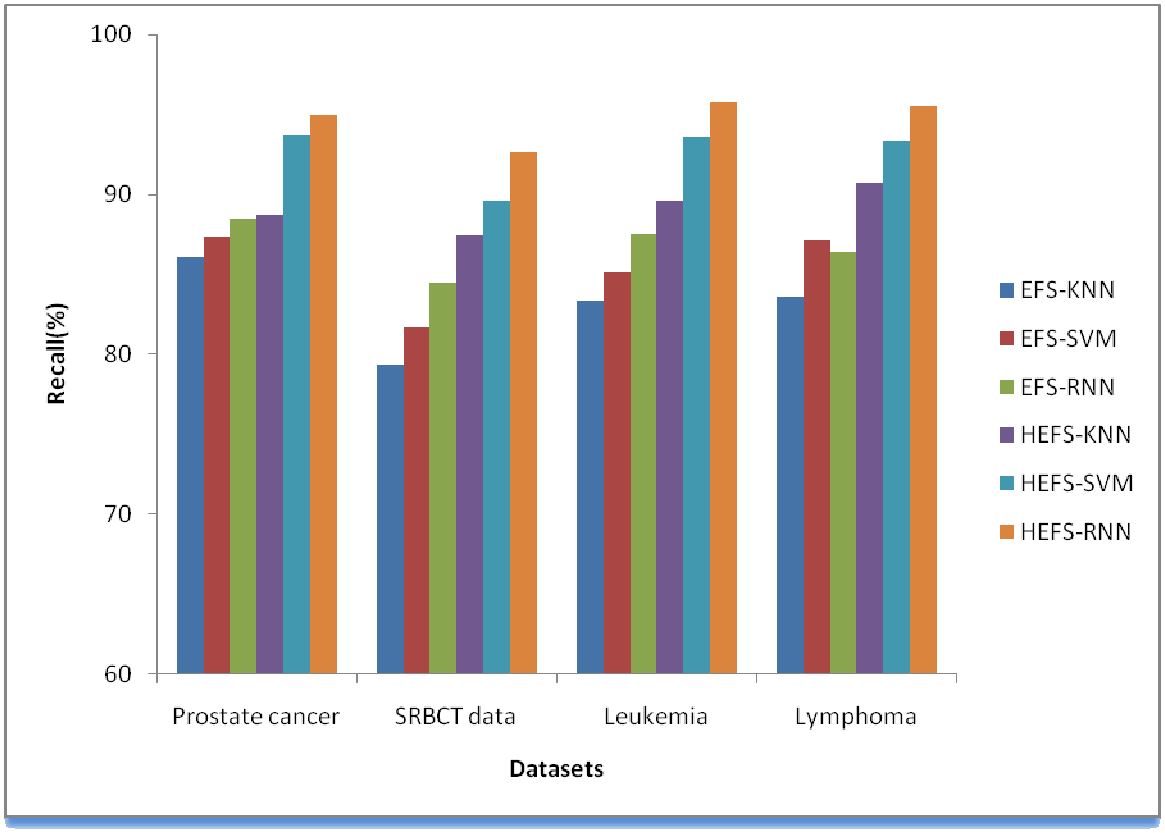

Figure.3 Comparison of Recall vs. Different Datasets

Figure 3 displays the recall for HEFS and SA-EFS using KNN, SVM and RNN. From this analysis, it is analyzed that the HEFS can achieve better recall than other SA-EFS. From three classifiers, HEFS–RNN gives the highest recall of 95% for prostate cancer dataset which is 6.538% higher than SA-EFS-RNN (See Table 3). Similarly, the recall value of HEFS–SVM for prostate cancer dataset is 6.329% higher than SA-EFS-SVM and HEFS-KNN algorithm is 2.674% higher than SA-EFS-KNN (See Table 3).

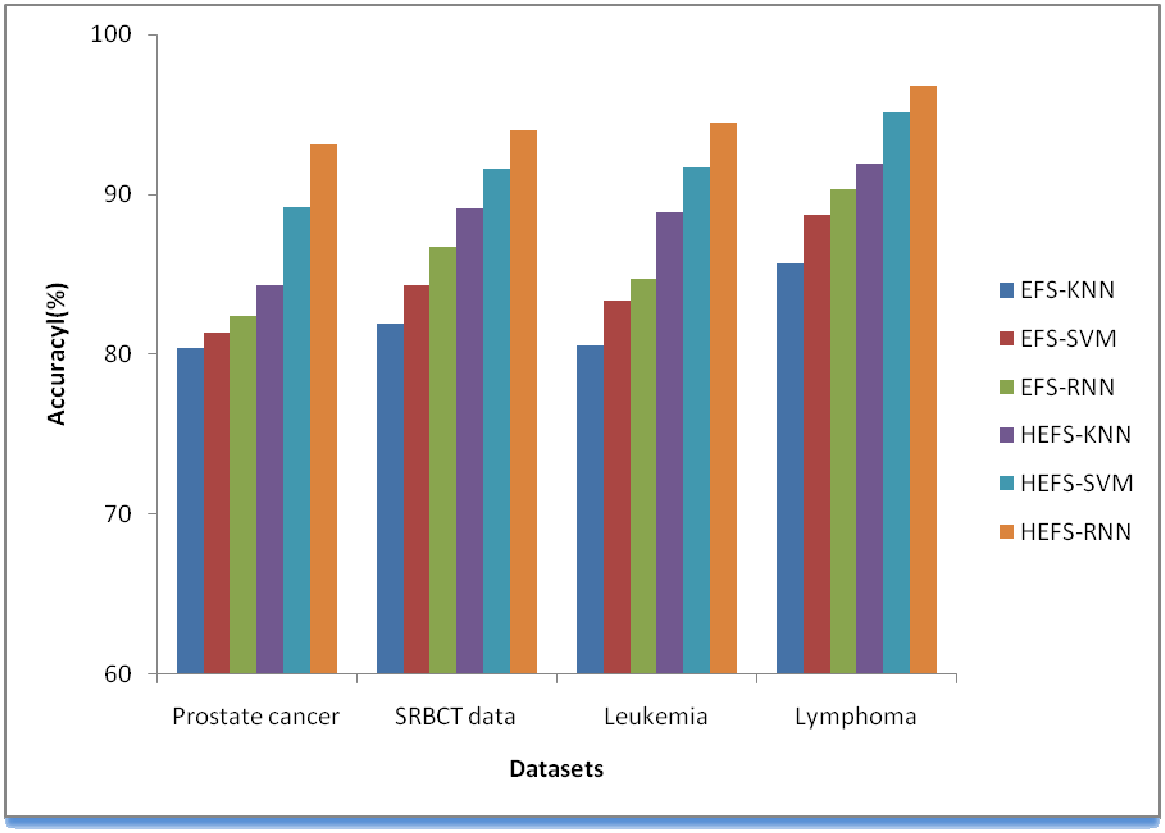

Figure.4 Comparison of Accuracy vs. Different Datasets

Figure 4 shows the accuracy for HEFS and SA-EFS using KNN, SVM and RNN. From this analysis, it is analyzed that the HEFS can achieve better recall than other SA-EFS. From three classifiers, HEFS–RNN gives the highest accuracy of 96.774% for Lymphoma dataset which is 6.451% higher than SA-EFS-RNN (See Table 3). Similarly, the accuracy value of HEFS–SVM for Lymphoma dataset is 6.451% higher than SA-EFS-SVM and HEFS–KNN algorithm is 6.451% higher than SA-EFS-KNN (See Table 3).

Table 3. Results Comparison of Different Datasets vs. Classifiers

Methods/Metrics	Leukemia Dataset			Prostate Cancer Dataset		
	Precision (%)	Recall (%)	Accuracy (%)	Precision (%)	Recall (%)	Accuracy (%)
SA-EFS-KNN	86.95652	83.33333	80.5555556	88.31	86.076	80.3921569
SA-EFS-SVM	88.88889	85.10638	83.3333333	88.46	87.342	81.372549
SA-EFS-RNN	89.3617	87.5	84.7222222	88.46	88.462	82.3529412
HEFS-KNN	93.47826	89.58333	88.8888889	91.03	88.75	84.3137255
HEFS-SVM	93.61702	93.61702	91.6666667	92.5	93.671	89.2156863
HEFS-RNN	95.74468	95.74468	94.4444444	96.2	95	93.1372549
Methods/Metrics	Lymphoma Dataset			SRBCT Dataset		
	Precision (%)	Recall (%)	Accuracy (%)	Precision (%)	Recall (%)	Accuracy (%)
SA-EFS-KNN	83.5767	83.5767	85.714	79.9285	79.3147	81.928
SA-EFS-SVM	85.4494	87.124	88.71	82.654	81.7085	84.337
SA-EFS-RNN	87.5144	86.3746	90.323	85.3472	84.4407	86.747
HEFS-KNN	90.4557	90.7144	91.935	88.3565	87.434	89.157
HEFS-SVM	93.9707	93.3334	95.161	90.617	89.623	91.566
HEFS-RNN	96.3214	95.5557	96.774	92.6297	92.6345	93.976

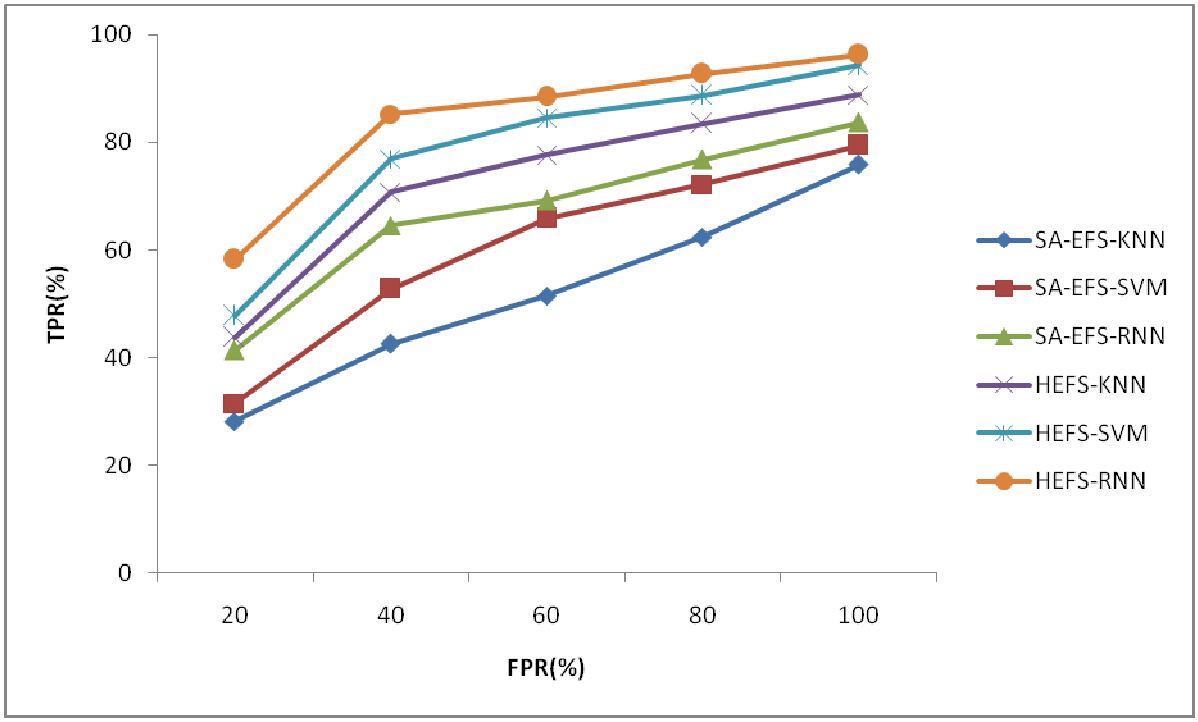

Figure.5 (a) ROC Curve for Prostate Cancer Dataset

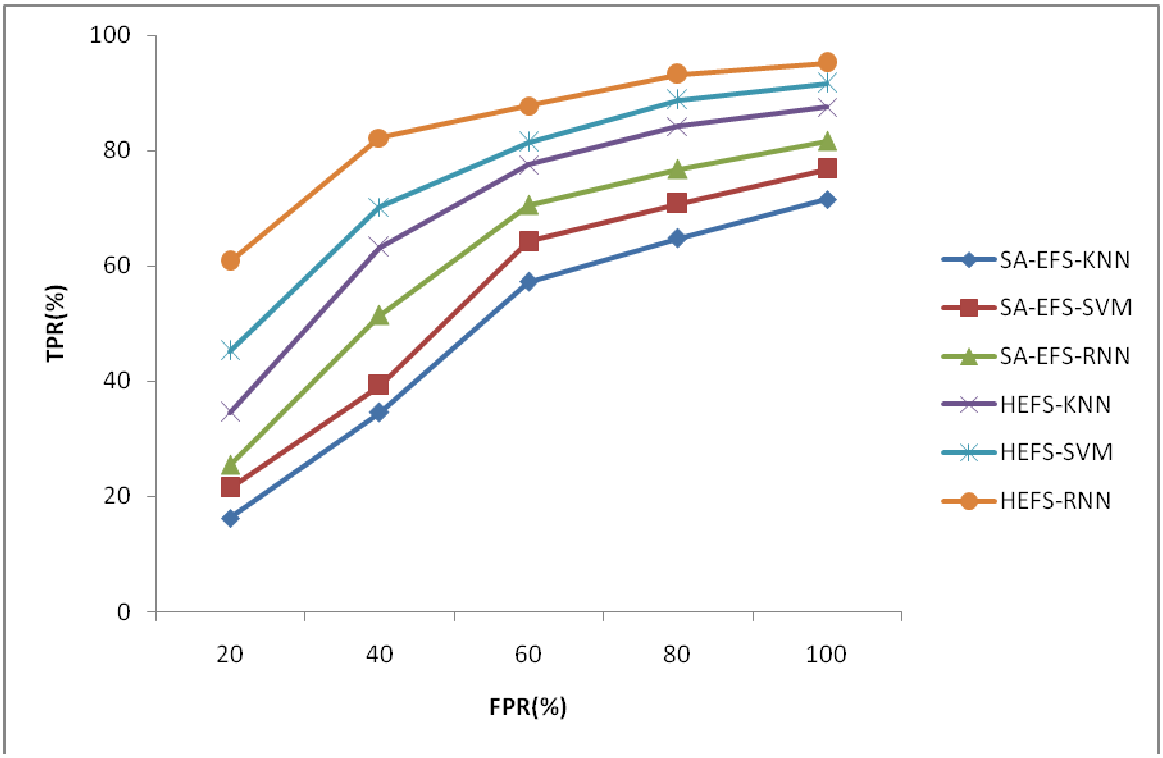

Figure.5 (b) ROC Curve for SRBCT Dataset

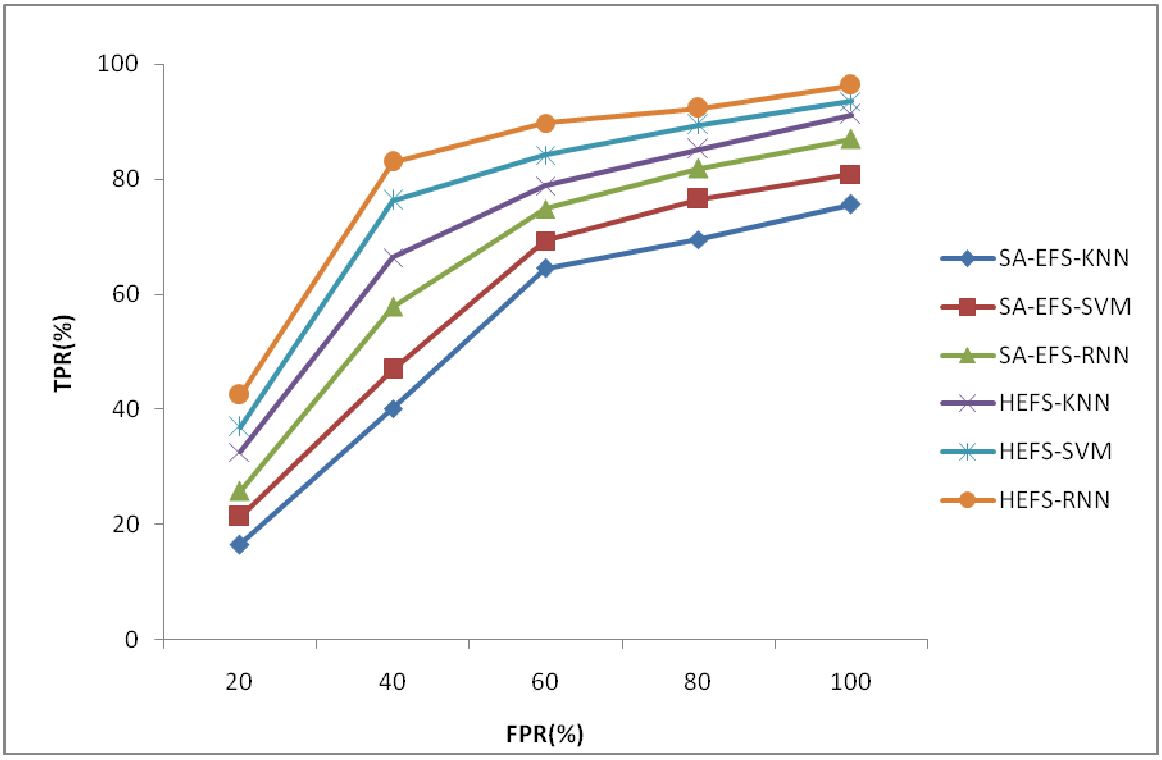

Figure.5 (c) ROC Curve for Leukemia Dataset

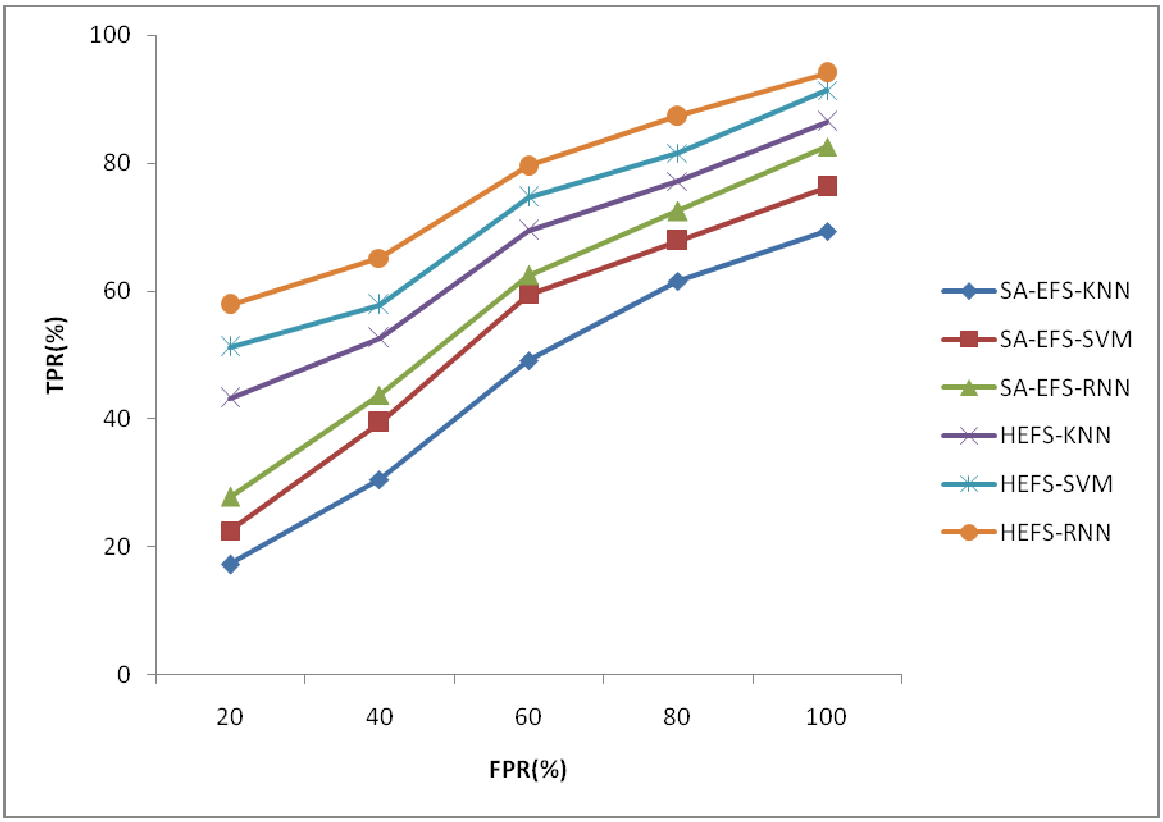

Figure.5 (d) ROC Curve for Lymphoma Dataset

Figure 5(a)-(d) portrayed that the ROC curves for HEFS-RNN is the largest than other approaches. AUC for HEFS-RNN is greater for prostate cancer dataset than the other datasets. In Figure 5(a)-(d), viewed that the impact of HEFS with 3 classifiers is higher than the impact of SCF with 3 classifiers.

V. CONCLUSION AND FUTURE WORK

This article presents an HEFS approach for cancer prediction via discovering the group of features. It encompasses SCF, FEHO and SVM-t. It is worked by aggregating the FS results offered via the single feature selectors into a final via WMV. It needs to be combined with a classifier for measuring the efficiency of FS. Fitness of HEFS approach is calculated from KNN, SVM and RNN. Finally, experimental outcomes verified that HEFS-RNN has an enhanced efficiency than the EFS with classifiers based on precision, recall, accuracy and ROC for different GEM datasets. Also, it is noticed that the efficiency is increased in most cases and RNN classifier gives higher efficiency compared to the other classifiers. The upcoming work will develop classifiers to together optimize the accuracy and robustness.

REFERENCES

1. Bosio M, Bellot P, Salembier P, Verges AO. Microarray classification with hierarchical data representation and novel feature selection criteria 2012; In: IEEE 12thInternational Conference on Bioinformatics & Bioengineering (BIBE), pp. 344-349.
2. Guzzi PH, Cannataro M. Challenges in Microarray Data Management and Analysis. Computer-Based Medical Systems 2011; 24(3): 1-6.
3. Liang S, Ma A, Yang S, Wang Y, Ma Q. A Review of Matched-pairs Feature Selection Methods for Gene Expression Data Analysis. ComputStructBiotechnol J 2018; 16: 88-97.
4. Li, Z., Xie, W. and Liu, T., 2018. Efficient feature selection and classification for microarray data. PloS one, 13(8), p.e0202167.
5. He J., D. Wu, N. Xiong, and C. Wu, "Orthogonal margin discriminant projection for dimensionality reduction," J. Supercomput., vol. 72, no. 6, pp. 2095-2110, 2016.
6. Peng H, Ding C, Long F. Minimum redundancy- maximum relevance feature selection. BioinformaComput Biol. 2005;3(2):185–205.
7. Saeys Y, et al. A review of feature selection techniques in bioinformatics. Bioinformatics. 2007;23(19):2507–17.
8. Hira, Z.M. and Gillies, D.F., 2015. A review of feature selection and feature extraction methods applied on microarray data. Advances in bioinformatics, Volume 2015, Article ID 198363, pp1-13.
9. Duval B, Hao JK. Advances in metaheuristics for gene selection and classification of microarray data. Brief Bioinform. 2010;11(1):127–41.
10. Wang, G.G., Deb, S. and Coelho, L.D.S., 2015, Elephant herding optimization. In 2015 3rd International Symposium on Computational and Business Intelligence (ISCBI) ,pp. 1-5.
11. Zengyou H, Weichuan Y (2010) Stable feature selection forbiomarker discovery. ComputBiolChem 34:215–225
12. Saeys Y, Abeel T, Van de Peer Y (2008) Robust feature selectionusing ensemble feature selection techniques. In: Machine learningand knowledge discovery in databases. Lecture notes incomputer science, vol 5212. Springer, Berlin, pp 313–325.
13. Yang F, Mao KZ (2011) Robust feature selection for microarraydata based on multicriterion fusion. IEEE/ACM Trans ComputBiolBioinform 8(4):1080–1092.
14. Awada W, Khoshgoftaar TM, Dittman D, Wald R, Napolitano A(2012) A review of the stability of feature selection techniquesfor bioinformatics data. IEEE 13th international conference oninformation reuse and integration, pp 356–363.
15. Guan D, Yuan W, Lee YK, Najeebullah K, Rasel MK (2014) A review of ensemble learning based feature selection. IETE Tech Rev 31(3):190–198
16. Seijo-Pardo B, Porto-Dı´az I, Bolo´n-Canedo V, Alonso-Betanzos A (2017) Ensemble feature selection: homogeneous and heterogeneous approaches. Knowl-Based Syst 118:124–139.
17. Bu¨hlmann P (2012) Bagging, boosting and ensemble methods. In: Gentle J, Ha¨rdle W, Mori Y (eds) Handbook of computational statistics. Springer handbooks of computational statistics. Springer, Berlin.
18. Ke, W., Wu, C., Wu, Y. and Xiong, N.N., 2018. A new filter feature selection based on criteria fusion for gene microarray data. IEEE Access, 6, pp.61065-61076.
19. Albashish, D., Sahran, S., Abdullah, A., Adam, A., AbdShukor, N. and Pauzi, S.H.M., 2015, Multi-scoring Feature selection method based on SVM-RFE for prostate cancer diagnosis. In 2015 International Conference on Electrical Engineering and Informatics (ICEEI) , pp. 682-686.
20. Das, P., Roychowdhury, A., Das, S., Roychoudhury, S. and Tripathy, S., 2020. sigFeature: Novel Significant Feature Selection Method for Classification of Gene Expression Data Using Support Vector Machine and t Statistic. Frontiers in Genetics, 11, pp.1-12.
21. Morovvat, M. and Osareh, A., 2016. An Ensemble of Filters and Wrappers for Microarray Data Classification. Mach. Learn. Appl. An Int. J, 3(2), pp.1-17.

22. Kavitha C.R, and Mahalekshmi T , "SVMFilefs- A novel ensemble feature selection technique for effective breast cancer diagnosis", International Journal of Civil Engineering and Technology (IJCIET), Vol.9, no. 11, 2018, pp. 1526–1533.
23. Alejandro, L.R., Marlet, M.A., Ulises, M.R.G. and Alberto, T., 2018. Ensemble Feature Selection and Meta-Analysis of Cancer miRNA Biomarkers. bioRxiv, pp.1-17.
24. Xu, J., Sun, L., Gao, Y. and Xu, T., 2014. An ensemble feature selection technique for cancer recognition. Bio-medical materials and engineering, 24(1), pp.1001-1008.
25. Wang, J., Xu, J., Zhao, C., Peng, Y. and Wang, H., 2019. An ensemble feature selection method for high-dimensional data based on sort aggregation. Systems Science & Control Engineering, 7(2), pp.32-39.
26. Bilen, M., Işik, A.H. and Yiğit, T., 2020. A New Hybrid and Ensemble Gene Selection Approach with an Enhanced Genetic Algorithm for Classification of Microarray Gene Expression Values on Leukemia Cancer. International Journal of Computational Intelligence Systems.
27. Yang F. and K. Z. Mao, "Improving robustness of gene ranking by multicriterion combination with novel gene importance transformation," Int. J. Data Mining Bioinf., vol. 7, no. 1, pp. 22-37, 2013
28. Yang F.and K. Z. Mao, "Robust feature selection for microarray data based on multicriterion fusion," IEEE/ACM Trans. Comput. Biol. Bioinf., vol. 8, no. 4, pp. 1080-1092, 2011.
29. Estévez P. A., 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, 2009.
30. Wang, G.-G.; Deb, S.; Coelho, L.d.S. Elephant herding optimization. In Proceedings of 2015 3rd International Symposium on Computational and Business Intelligence (ISCBI 2015), Bali, Indonesia, 7–9 December 2015; pp. 1–5.