How to Cite:
Ramya, P., & Reddy, T. B. (2022). A hybrid cluster based classification model for high dimensional disease prediction databases. *International Journal of Health Sciences*, *6*(S9), 4707–4719. https://doi.org/10.53730/ijhs.v6nS9.13989

A hybrid cluster based classification model for high dimensional disease prediction databases

P. Ramya
Research scholar, Dept. of Computer Science & Technology, Sri Krishnadevaraya University Ananthapuram, India

Dr. T. Bhaskar Reddy
Professor, Dept. of Computer Science & Technology, Sri Krishnadevaraya University Ananthapuram, India

***Abstract*---**As biomedical databases continue to expand, it becomes increasingly difficult to identify a crucial feature for a classification task due to big data size and sparsity issues. Traditional feature subset models rely on fixed-sized dimensions for the feature ranking and classification process, which is not suitable for addressing concerns with sparsity, missing values, and imbalance in the selection of crucial features for the data classification process. To enhance disease prediction effectiveness, this article proposes a hybrid ensemble feature selection method that employs an advanced cluster-based classification model. The model uses an ensemble of rated features to classify the disease with high accuracy and true positive rate. To improve the effectiveness of tree pruning and classification, we introduce a novel cluster-based classification model. We simulated experimental results using various training datasets to predict accuracy. Our proposed results demonstrate that the gene-chemical disease clustering-based classification framework outperforms traditional methods, statistical metrics, and classification models in terms of optimization.

***Keywords*---**Hybrid Cluster, high dimensional, databases.

1. Introduction

The size of microarray datasets has increased daily, making it more challenging to find a relevant feature in the vast feature space due to data size and sparsity issues[1]. Microarray feature ranking and classification is one of the biggest hurdles for scientific and biomedical researchers due to its high dimensional feature space and constrained sample size. A microarray can identify a specific gene-related disease, which is a collection of several identical DNA molecules[2].

Feature transformation, feature ranking, and data classification are the key methods for classifying high-dimensional data with a high true positive rate. Feature transformation refers to the technique of normalizing feature values within a restricted range. Feature transformation improves the feature ranking procedure in a high-dimensional feature space. Most traditional feature transformation methods, such as log transformation and min-max normalization, are not affected by outliers or data distribution. Medical data classification involves grouping a lot of medical information into relevant cluster sets, where each cluster represents a specific subject or situation[3]. The medical data in a group should be highly similar, while the similarity between different clusters of medical data should be minimal. Traditional classification methods were used to cluster medical data without considering the contextual information of the medical data set. For example, medical data are grouped into multiple clusters if two or more of them reflect the same issue using different semantically equivalent terms. This classification method can make information retrieval less effective. Therefore, medical data classification has become increasingly crucial to improve medical data exchange and communication in a dispersed context. Medical data classification has various uses in the fields of data mining and information retrieval. Other medical data classification strategies have arisen to better perform the classification and overcome the drawbacks of traditional methods[4].

In machine learning, a set of guidelines or decision criteria can be learned by a classifier from a set of labelled data that has been annotated by a subject matter expert[5]. This approach offers better scalability and lower costs for classifying medical data compared to systems that rely solely on manual input. The majority of research on machine learning-based classification of medical data has focused on binary classifiers, which use positive and negative examples to determine whether a medical data set belongs to a particular class. Previously, when dealing with a corpus containing medical data from numerous classes, a separate binary classifier was built for each class and the results of each binary classification were aggregated. Classification can take many forms, from fully automated systems that use no human intervention to semi-automatic systems that use a combination of human and machine approaches. One of the most important chronic diseases is the microarray dataset, which takes a long time to progress from moderate symptoms to severe illness and death. Medical data typically contain a set of cancer patches and their disease connections, making it difficult for doctors to identify patients at high risk of contracting the disease. In addition, patients' expertise and prior clinical history aid in detecting the disease more quickly. Structural changes to the airways can lead to a reduction in luminal diameter and airway wall thickening. Emphysema, a dangerous chronic lung disease that causes the breakdown of the walls of the alveoli without fibrosis[7].

2. Related works

Traditionally, as the size of the training dataset is small, medical disease prediction rate could be dramatically reduced due to class imbalance and high dimension space. In this filtering method, each attribute is tested for missing values. Traditional probability estimation techniques such as Naïve Bayes, markov model, and Bayesian model are used to find the highest probability estimation variance among the gene and its related disease sets in biomedical

data sets. Medical data summaries represent instances or phrases extracted from different sources without any subjective human intervention or editorial touch and thus making the end product completely unbiased. [8] presented a new correlation based feature selection technique along with clustering of high dimensional data. Feature selection approach is considered as an important approach which can significantly decrease the issues of dimensionality. All conventional feature selection approaches are incapable in order to scale huge space[9]. This research paper emphasizes on an advanced technique in order to overcome dimensionality issues. Here, the clustering approach is merged with correlation technique to generate proper feature subset. Initially, each and every irrelevant feature is discarded with the help of k-mean clustering scheme. After that, all non-redundant features are chosen with the help of correlation measure from every individual cluster. [10] introduced an advanced two-phase grading technique for feature selection and classification of microarray data . They considered a new Pareto based feature ranking approach. High dimensional search space in case of microarray data along with huge numbers of genes increases the overall complexity of the problem exponentially. All the genes are not necessary; therefore only important genes are needed to be retrieved[11]. Hence, the process of dimension reduction play vital role during the process of classification. Almost all of the ranking schemes are implemented during the process of feature selection. Various kinds of ranking schemes are responsible for allocating different ranks to a particular gene[12]. Selection process depends upon a particular ranking scheme is not appropriate for all kinds of problems. Therefore, the classification performance can be decreased remarkably. In order to resolve the above mentioned issue, a bi- objective rank based approach is presented in this research paper. The presented approach is basically a two rank based approach which has the responsibility to produce numbers of features. Kumar, et.al, proposed a recursive memetic approach in order to carry out the process of gene selection in microarray data [12]. Feature selection approach has significant role the field of biomedical diagnosis process. It has the objective to choose a subset of genes those can be used to implement an appropriate classification algorithm in order to predict cancer disease. Apart from this, this approach is efficient enough to detect the type of cancer which is very complicated task. The researchers tried to obtain better classification accuracy along with extracting few numbers of features. They presented a recursive memetic algorithm which is efficient and effective in order to choose genes. This approach is a modified version of memetic algorithm that results better performance as compared to the traditional algorithm. Again, this algorithm is much better than that of genetic algorithm[13]. Dimension reduction is considered as an important step during the process of machine learning and data mining. It has wide range of applications in the field of medicines, bioinformatics genetics. In this piece of research work, they have presented a new two-step local dimension reduction technique which has the responsibility to classify microarray data. By considering the microarray data, differentially expressed genes are detected with the help of meta-analysis process. Genetic algorithm optimised artificial neural network algorithm is proposed in order to develop the most efficient prediction model. This model has the responsibility to filter candidate genes. In this research paper, they proposed a diagnostic and prognostic prediction scheme. This technique can be implemented in case of different high throughput data in order to identify various biomarkers.

Ong emphasized on feature selection and classification of microarray data with the help of latest machine learning approaches [14]. In order to resolve the above mentioned issues, different machine learning approaches are introduced. This machine learning approaches are usually implemented in the feature selection process. They have thoroughly studied and analysed different previously existing methods in order to choose most significant and important features from the microarray data set. They have integrated the most popular classification techniques along with different feature selection approaches at the time of classification. Most of the traditional approaches detect inappropriate and computationally infeasible patterns on high dimensional datasets. Hence, it is difficult to process all of the cancer patterns that are not required during the process of classification. Hence, the overall computational overhead also increases significantly. Unwanted noise is resulted during the process of classification. Hence, it is very much required to select essential cancer patches during the classification process. All of the traditional cancer selection techniques involve a perfect combination of filter and wrapper schemes. Filtering approaches have the responsibility to rank every individual feature according to their goodness. During the process of ranking, the relationship among every individual cancer with respective class label is considered. Univariate scoring metric play a significant role in the above ranking process. As the size of the feature space increases, traditional ensemble classifiers select a predefined number of features for classification. Learning classification models with all the high dimensional features may result serious issues such as performance and scalability[15].

3. Proposed Model

Improved EM Gene-disease clustering approach

In the expectation maximization model, two phases are implemented on the training data to predict the best clustered features for the gene-disease prediction. EM is an iterative soft cluster that estimates cluster densities. Basically, cluster membership is a hidden latent variable that the maximum likelihood EM method estimates. To start with, a random assignment of instances to clusters is used - actually, in Weka's EM library implementation, the starting point is provided by the best of 10 runs of the k-means algorithm. Initial distributions for each cluster are learned from this starting point and then the E and M step of the algorithm are executed in subsequent iterations. The E step estimates the cluster membership of each instance given the current model - this is a soft, probabilistic membership where the predicted density/probability distribution is used to weight each instance. Then the M step re-estimates the parameters of the normal and discrete distributions for each cluster using the weights computed by the E step. Iteration stops when the likelihood of the training data with respect to the model does not increase enough from one iteration to the next, or the maximum number of iterations have been performed.

Step 1: E-step: In this step, based on the model parameters, proposed model computes the probability of each data point as a segment.

In the M step, filling missing labels and find the model that improves similarity likelihood of the data.

Find the features using EM clusters as S_{IF}.

For each feature pair in F[]

Do

$$\eta_1 = \sum_{i=1}^{|S_r|} F_i[i].S(F_p, F_q)$$

$$\eta_2 = \sum_{i=1}^{|S_{IF}|} F_j[i].S(F_p, F_q)$$

$$MI = Mutual - Information(S_i, S_j)$$

$$Chebyshev = d(i,j) = \max_r |x(i,r) - x(j,r)|$$

$$S(F_p, F_q) = Max\{MI, Min\{Chebyshev(F_r(x,y)), Chebyshev(F(x,y))\}\}$$

Similar Segmented measure = $Max\{\eta_1, \eta_2\}$

Step 3: The probability estimator is used to improve the occurrence of disorder patterns in the given dataset.

Proposed Classification Algorithm

1: Read pre-processing training datasets.

All these input patterns are partitioned 'm' clusters based on EM approach.

2: To each clustered |

3: do

4: Apply proposed ensemble decision tree model on each cluster data.

5: In the proposed classification model, a novel gene-disease feature selection measure is implemented on each cluster.

Classifier-1

Probabilistic kernel density based KNN classification model

Input: Filtered data D, Training data classes C, KNN data points K.

1. Input filtered data D with different class labels C
2. Perform KNN approach on the filtered data D.
3. Compute local density estimation on the KNN data points in the E-step of EM model

$$d_c = \mu^K + \Pr\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(\psi_i^K - \mu^K)^2}$$

where N is the number of k nearest neighbor points in the KNN approach,

ψ_i^K is the is the mean distance between Kth nearest neighbor to point i.

$\psi_t^K = max_j \in KNN_i(d_{ij})$, and

μ^K is the average of ψ_i^K , computed as $\mu^K = \frac{1}{N}\sum_{i=1}^{N}\psi_i^K$

$\Pr=\Pr(i,C_m)=\text{Max}\{\text{Prob}(i/C_m);i \in \text{KNN}\}$

$$LDE(\rho_i) = \frac{1}{\eta}\exp\left(-\frac{\|d_{ij} - d_c\|^2}{2\sigma^2}\right)\sum_{j \in KNN_i}\Pr(d_{ij}, C_m)\exp\left(-\frac{d_{ij}^2}{d_c^2}\right)$$

Attribute ranking measure:

a) Enhanced entropy:

$$\Pr = -\text{Prob}(D_i).\log(\text{Prob}(D_i))$$

$$\text{Ent}(D) = \sum_i \Pr$$

Math.cbrt(entropy(data)*total*GHDSplitCriterion.computeHellinger(data))*Pr/ (chiVal(data));

In the proposed boosting algorithm, a set of weak classifiers are used to improve the classification rate using the boosting mechanism.

Modified Score for BN:

$\theta = \text{Conditional PriorProb}(s_i)$;

$\phi = \text{Joint Prob}(D / s_i)$;

$\text{PropBayesScore} = \log(\theta) + \log(\phi)$

where

$$\text{Joint Prob}(D / s_i) = \prod_{i=0}^{n}\prod_{j=0}^{q_i}\frac{\Gamma(\sum_{k=1}^{r}\alpha_{ijk}\log(\alpha_{ijk}))}{\Gamma(\sum_{k=1}^{r}\alpha_{ijk}\log(\alpha_{ijk}) + \sum\log(N_{ij}))}\prod_{k=1}^{r}\frac{\Gamma(\sum_{k=1}^{r}\exp(\alpha_{ijk})\log(\alpha_{ijk}) + \sum N_{ijk})}{\Gamma(\sum_{k=1}^{r}\alpha_{ijk}\log(\alpha_{ijk}))}$$

4. Experimental Results

Experimental results are performed on real-time medline biomedical documents and micro-array datasets. Proposed feature selection-based ensemble methods increase the efficiency of the F-measure, recall and accuracy on high dimensional datasets. Proposed model uses the entire training data set for construction of decision patterns; therefore, the prediction accuracy of each cross validation tends to be more accurate than the traditional ensemble classification models. Simulation results represent the proposed ensemble classification improves the overall true positive and false negative rate. Also, the main advantage of using proposed model is to reduce the error rate on high dimensional features.

Training datasets

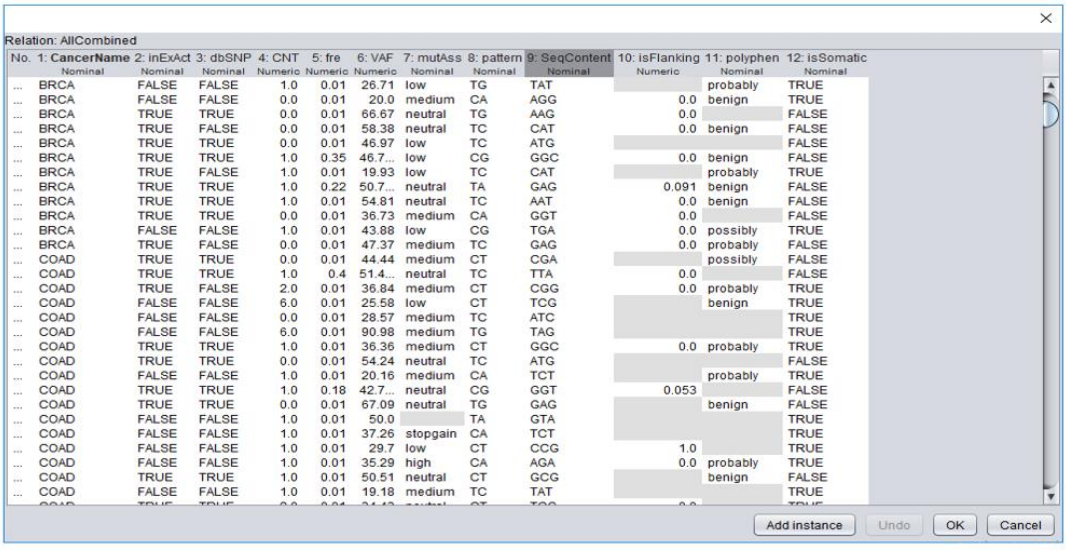
1. Autistic Spectrum Disorder Screening Data for Adult

Autistic Spectrum Disorder (ASD) is a neurodevelopment condition associated with significant healthcare costs, and early diagnosis can significantly reduce these. Unfortunately, waiting times for an ASD diagnosis are lengthy and procedures are not cost effective. The economic impact of autism and the increase in the number of ASD cases across the world reveals an urgent need for the development of easily implemented and effective screening methods. Therefore, a time-efficient and accessible ASD screening is imminent to help health professionals and inform individuals whether they should pursue formal clinical diagnosis. The rapid growth in the number of ASD cases worldwide necessitates datasets related to behaviour traits. However, such datasets are rare making it difficult to perform thorough analyses to improve the efficiency, sensitivity, specificity and predictive accuracy of the ASD screening process. Presently, very limited autism datasets associated with clinical or screening are available and most of them are genetic in nature. Hence, we propose a new dataset related to autism screening of adults that contained 20 features to be utilised for further analysis especially in determining influential autistic traits and improving the classification of ASD cases. In this dataset, we record ten behavioural features (AQ-10-Adult) plus ten individuals characteristics that have proved to be effective in detecting the ASD cases from controls in behaviour science.

Table 1: Features and their descriptions

Attribute	**Type**	**Description**
Age	Number	Age in years
Gender	String	Male or Female
Ethnicity	String	List of common ethnicities in text format
Born with jaundice	Boolean (yes or no)	Whether the case was born with jaundice
Family member with PDD	Boolean (yes or no)	Whether any immediate family member has a PDD
Who is completing the test	String	Parent, self, caregiver, medical staff, clinician ,etc.
Country of residence	String	List of countries in text format
Used the screening app before	Boolean (yes or no)	Whether the user has used a screening app
Screening Method Type	Integer (0,1,2,3)	The type of screening methods chosen based on age category (0=toddler, 1=child, 2= adolescent, 3= adult)
Question 1 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 2 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 3 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 4 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 5 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 6 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 7 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 8 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 9 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 10 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Screening Score	Integer	The final score obtained based on the scoring algorithm of the screening method used. This was computed in an automated manner

2. Somatic Cancer: Evaluation of the classifiers was performed on two different cancer datasets: COAD (colon adenocarcinoma), BRCA (breast invasive carcinoma).



Relation: AllCombined

No.	1: CancerName	2: inExAct	3: dbSNP	4: CNT	5: fre	6: VAF	7: mutAss	8: pattern	9: SeqContent	10: isFlanking	11: polyphen	12: isSomatic
	Nominal	Nominal	Nominal	Numeric	Numeric	Numeric	Nominal	Nominal	Nominal	Numeric	Nominal	Nominal
...	BRCA	FALSE	FALSE	1.0	0.01	26.71	low	TG	TAT		probably	TRUE
...	BRCA	FALSE	FALSE	0.0	0.01	20.0	medium	CA	AGG	0.0	benign	TRUE
...	BRCA	TRUE	TRUE	0.0	0.01	66.67	neutral	TG	AAG	0.0		FALSE
...	BRCA	TRUE	FALSE	0.0	0.01	58.38	neutral	TC	CAT	0.0	benign	FALSE
...	BRCA	TRUE	TRUE	0.0	0.01	46.97	low	TC	ATG			FALSE
...	BRCA	TRUE	TRUE	1.0	0.35	46.7...	low	CG	GGC	0.0	benign	FALSE
...	BRCA	TRUE	FALSE	1.0	0.01	19.93	low	TC	CAT		probably	TRUE
...	BRCA	TRUE	TRUE	1.0	0.22	50.7...	neutral	TA	GAG	0.091	benign	FALSE
...	BRCA	TRUE	TRUE	1.0	0.01	54.81	neutral	TC	AAT	0.0	benign	FALSE
...	BRCA	TRUE	TRUE	0.0	0.01	36.73	medium	CA	GGT	0.0		FALSE
...	BRCA	FALSE	FALSE	1.0	0.01	43.88	low	CG	TGA	0.0	possibly	TRUE
...	BRCA	TRUE	FALSE	0.0	0.01	47.37	medium	TC	GAG	0.0	probably	FALSE
...	COAD	TRUE	TRUE	0.0	0.01	44.44	medium	CT	CGA		possibly	FALSE
...	COAD	TRUE	TRUE	1.0	0.4	51.4...	neutral	TC	TTA	0.0		FALSE
...	COAD	TRUE	FALSE	2.0	0.01	36.84	medium	CT	CGG	0.0	probably	TRUE
...	COAD	FALSE	FALSE	6.0	0.01	25.58	low	CT	TCG		benign	TRUE
...	COAD	FALSE	FALSE	0.0	0.01	28.57	medium	TC	ATC			TRUE
...	COAD	FALSE	FALSE	6.0	0.01	90.98	medium	TG	TAG			TRUE
...	COAD	TRUE	TRUE	1.0	0.01	36.36	medium	CT	GGC	0.0	probably	TRUE
...	COAD	TRUE	TRUE	0.0	0.01	54.24	neutral	TC	ATG			FALSE
...	COAD	FALSE	FALSE	1.0	0.01	20.16	medium	CA	TCT		probably	TRUE
...	COAD	TRUE	TRUE	1.0	0.18	42.7...	neutral	CG	GGT	0.053		FALSE
...	COAD	TRUE	TRUE	0.0	0.01	67.09	neutral	TG	GAG		benign	FALSE
...	COAD	FALSE	FALSE	1.0	0.01	50.0		TA	GTA			TRUE
...	COAD	FALSE	FALSE	1.0	0.01	37.26	stopgain	CA	TCT			TRUE
...	COAD	FALSE	FALSE	1.0	0.01	29.7	low	CT	CCG	1.0		TRUE
...	COAD	FALSE	FALSE	1.0	0.01	35.29	high	CA	AGA	0.0	probably	TRUE
...	COAD	TRUE	TRUE	1.0	0.01	50.51	neutral	CT	GCG		benign	FALSE
...	COAD	FALSE	FALSE	1.0	0.01	19.18	medium	TC	TAT			TRUE

Add instance | Undo | OK | Cancel

Figure 1: Sample Somatic cancer dataset

Somatic cancer features visualization

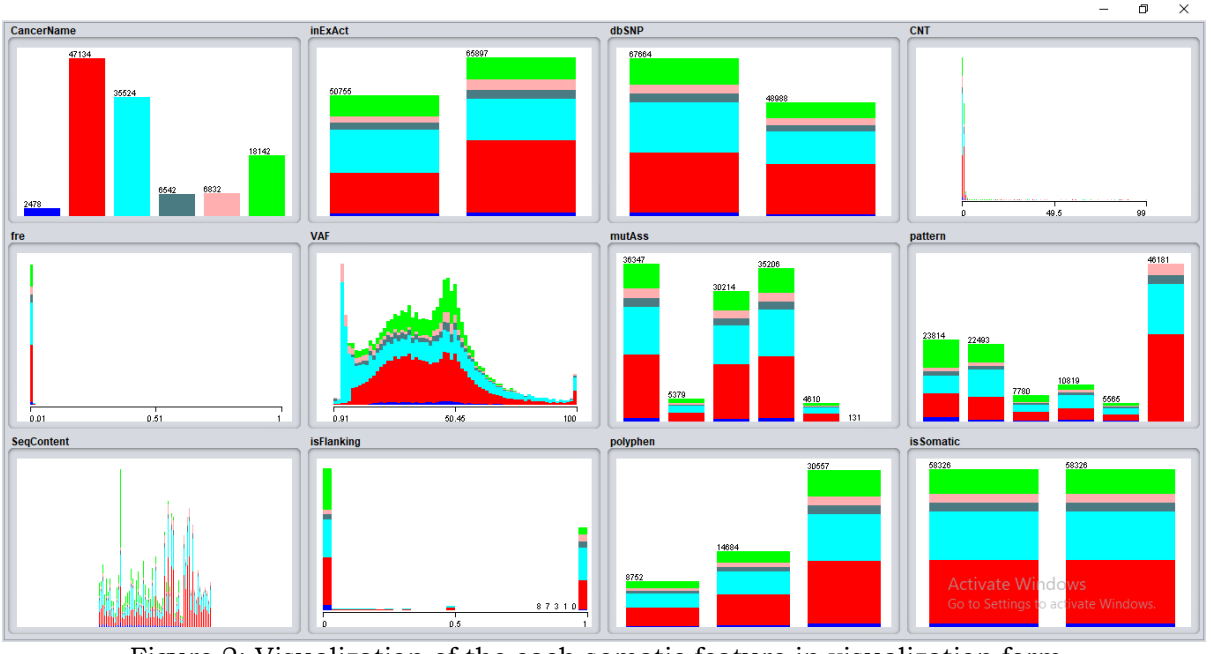


Figure 2: Visualization of the each somatic feature in visualization form

Each plot in the above figure represents the frequency for different values in each attribute.

No.	Label	Count
1	BRCA	2478
2	COAD	47134
3	ESCA	35524
4	KIRC	6542
5	PAAD	6832
6	UCEC	18142

Figure 3: Each somatic cancer type and its instance count

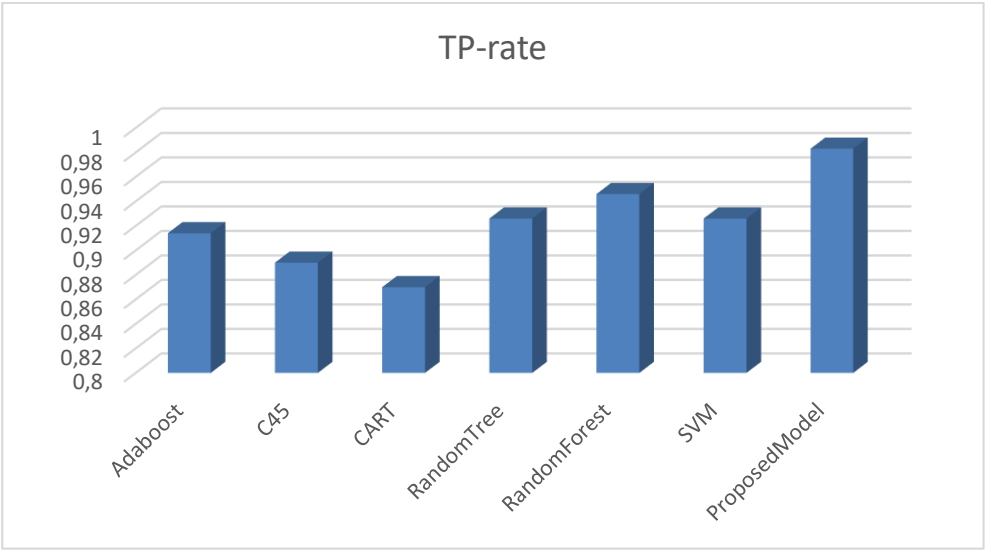


Figure 4: Comparative analysis of proposed model to the conventional models on somatic cancer datasets

Figure 4, describes the performance of the proposed model on somatic dataset. Here, all the cancer datasets are evaluated using the proposed model to find the average true positive rate on the high dimensional datasets. From the figure, it is visualized that the present approach has better true positive rate over the existing models.

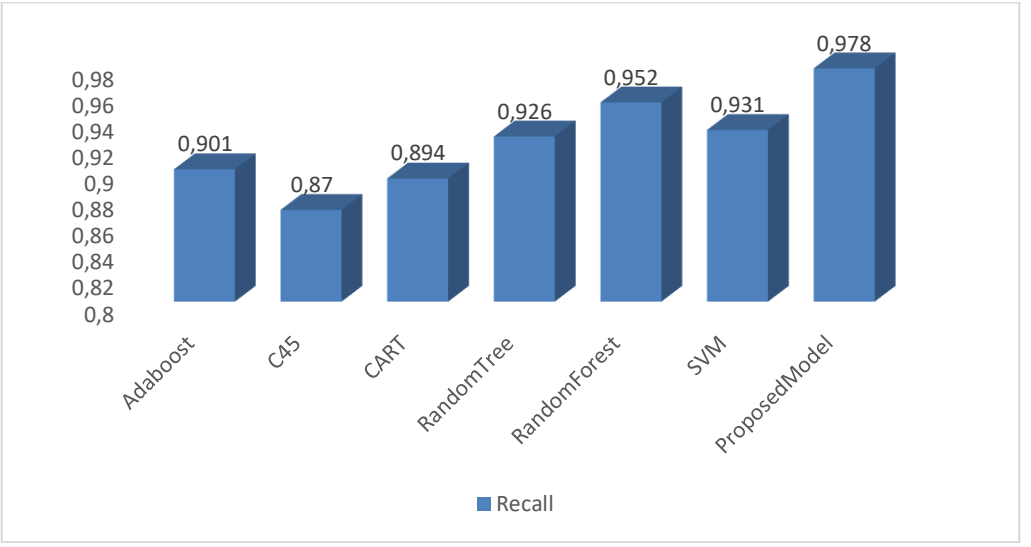


Figure 5: Comparative analysis of proposed model to the conventional models on Autistic datasets using Recall measure

Figure 5, describes the performance of the proposed model on Autistic dataset. Here, all the cancer datasets are evaluated using the proposed model to find the average recall on the high dimensional datasets. From the figure, it is visualized that the present approach has better recall over the existing models.

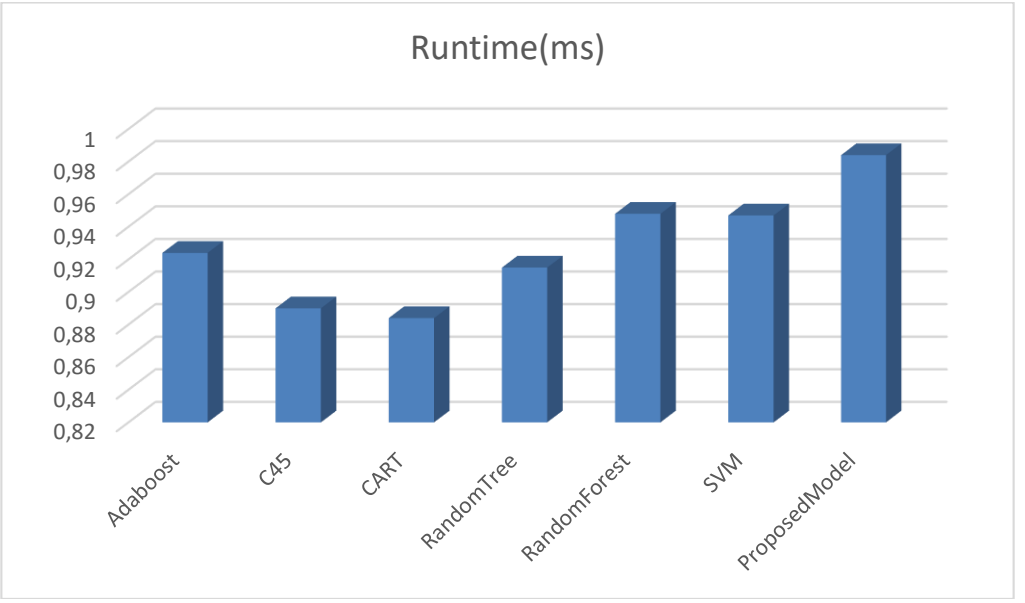


Table 6: Comparative runtime analysis of present technique to the conventional techniques by using accuracy on different real-time average accuracy of all the training datasets.

Conclusion

Ensemble classification algorithm is one of the best classification learning models for high dimensional datasets. In this work, a hybrid feature selection measure are implemented to test the accuracy of the test data. In this model, different feature ranking sets are extracted to predict the best majority voting of the classification model. Experimental results proved that the present feature subset selection-based classification model is best applicable to high dimensional datasets with high accuracy and runtime than the traditional algorithms. In the future work, an advanced cluster based classification model is used to cluster the unknown cancer test data for classification problem.

References

[1] N. K. Berry et al., “Enrichment of atypical hyperdiploidy and IKZF1 deletions detected by SNP-microarray in high-risk Australian AIEOP-BFM B-cell acute lymphoblastic leukaemia cohort,” Cancer Genetics, vol. 242, pp. 8–14, Apr. 2020, doi: 10.1016/j.cancergen.2020.01.051.

[2] N. K. Berry, R. J. Scott, P. Rowlings, and A. K. Enjeti, “Clinical use of SNP-microarrays for the detection of genome-wide changes in haematological malignancies,” Critical Reviews in Oncology/Hematology, vol. 142, pp. 58–67, Oct. 2019, doi: 10.1016/j.critrevonc.2019.07.016.

[3] B. Cao et al., “Multiobjective feature selection for microarray data via distributed parallel algorithms,” Future Generation Computer Systems, vol. 100, pp. 952–981, Nov. 2019, doi: 10.1016/j.future.2019.02.030.

[4] R. Dash, “A two stage grading approach for feature selection and classification of microarray data using Pareto based feature ranking techniques: A case study,” Journal of King Saud University - Computer and Information Sciences, vol. 32, no. 2, pp. 232–247, Feb. 2020, doi: 10.1016/j.jksuci.2017.08.005.

[5] R. Dash, R. Dash, and R. Rautray, “An evolutionary framework based microarray gene selection and classification approach using binary shuffled frog leaping algorithm,” Journal of King Saud University - Computer and Information Sciences, Apr. 2019, doi: 10.1016/j.jksuci.2019.04.002.

[6] N. A. Firdausanti and Irhamah, “On the Comparison of Crazy Particle Swarm Optimization and Advanced Binary Ant Colony Optimization for Feature Selection on High-Dimensional Data,” Procedia Computer Science, vol. 161, pp. 638–646, Jan. 2019, doi: 10.1016/j.procs.2019.11.167.

[7] M. Ghosh, S. Begum, R. Sarkar, D. Chakraborty, and U. Maulik, “Recursive Memetic Algorithm for gene selection in microarray data,” Expert Systems with Applications, vol. 116, pp. 172–185, Feb. 2019, doi: 10.1016/j.eswa.2018.06.057.

[8] B. I. Grisci, B. C. Feltes, and M. Dorn, “Neuroevolution as a tool for microarray gene expression pattern identification in cancer research,” Journal of Biomedical Informatics, vol. 89, pp. 122–133, Jan. 2019, doi: 10.1016/j.jbi.2018.11.013.

[9] Y. He, J. Zhou, Y. Lin, and T. Zhu, “A class imbalance-aware Relief algorithm for the classification of tumors using microarray gene expression data,” Computational Biology and Chemistry, vol. 80, pp. 121–127, Jun. 2019, doi: 10.1016/j.compbiolchem.2019.03.017.

[10] C. Kang, Y. Huo, L. Xin, B. Tian, and B. Yu, “Feature selection and tumor classification for microarray data using relaxed Lasso and generalized multi-class support vector machine,” Journal of Theoretical Biology, vol. 463, pp. 77–91, Feb. 2019, doi: 10.1016/j.jtbi.2018.12.010.
[11] K. Kappel, E. Eschbach, M. Fischer, and J. Fritsche, “Design of a user-friendly and rapid DNA microarray assay for the authentication of ten important food fish species,” Food Chemistry, vol. 311, p. 125884, May 2020, doi: 10.1016/j.foodchem.2019.125884.
[12] A. Kumar, S. C. Pandey, and M. Samant, “DNA-based microarray studies in visceral leishmaniasis: identification of biomarkers for diagnostic, prognostic and drug target for treatment,” Acta Tropica, vol. 208, p. 105512, Aug. 2020, doi: 10.1016/j.actatropica.2020.105512.
[13] M. Momenzadeh, M. Sehhati, and H. Rabbani, “A novel feature selection method for microarray data classification based on hidden Markov model,” Journal of Biomedical Informatics, vol. 95, p. 103213, Jul. 2019, doi: 10.1016/j.jbi.2019.103213.
[14] H. F. Ong, N. Mustapha, H. Hamdan, R. Rosli, and A. Mustapha, “Informative top-k class associative rule for cancer biomarker discovery on microarray data,” Expert Systems with Applications, vol. 146, p. 113169, May 2020, doi: 10.1016/j.eswa.2019.113169.
[15] S. P. Potharaju and M. Sreedevi, “Distributed feature selection (DFS) strategy for microarray gene expression data to improve the classification performance,” Clinical Epidemiology and Global Health, vol. 7, no. 2, pp. 171–176, Jun. 2019, doi: 10.1016/j.cegh.2018.04.001.