Performance Analysis of Hybrid And Ensemble Techniques For Efficient Malicious Tumor Detection

S.SubashChandraBose, Research Scholar, PG and Research Department of Computer Science, Government Arts College Udumalpet, Tamilnadu, India, bose.milestone@gmail.com

Dr.T.Christopher, Assistant Professor PG & Research Department of Computer Science, Government Arts College, Coimbatore, India, chris.hodcs@gmail.com

Abstract - Biological Medical healthcare data are more complex and heterogeneous, single clustering algorithms are not efficient and not best in finding outliers and optimal solutions. While using bagging like ensemble methods, more techniques are combined to find more accurate and efficient tumor detection. To improve the efficiency, accuracy and robustness in malicious tumor detection we introduced Hybrid support vector machine (HSVM), Aggregate Linear Discriminate Analysis (ALDA-EC) and Deep Learning Spectral Cluster Gaussian Mixture (DL-SCGM) to improve the efficiency of classifying optimal results in minimum time with increase in accuracy. Here we compare these three methods and how one is best with another by finding the maximum malicious tumor detection.

***Keywords**: Ensemble, Cluster, Classification, Tumor, Hybrid support vector machine, Aggregate Linear Discriminate Analysis, Gaussian Mixture.*

I. INTRODUCTION

Cancer class discovery [1] is one of the most task in diagnosis of cancer class analysis and it supporting for biological medical applications to perform successful diagnosis and treatment of cancer classification, there are a number of previous research focus and work on cancer class discovery from microarray data .Most of the previous works based on [2] single clustering approaches such as non-negative matrix factorization (NMF) Self-organizing map (SOM), Hierarchical clustering.

(HC), though single clustering based tumor classification have been applied for various kinds of cancer gene expressions datasets for cancer class discovery. Quality of method is depends on the [3] similarity measure with high inter-cluster similarity and intra cluster similarity by its ability to discover all of the hidden patterns. Feature selection algorithm such as [4] Kernel F-score, and explicit margin-based feature selection is used to improve the classification accuracy for SVM classifiers on machine learning algorithms.

PCA [5] is mainly used for dimensional reduction and used in many applications such as pattern recognition, image processing, visualization and data compression, it reduces redundant values.PCA is evaluated from the Eigen vectors of the covariance matrix of the original variables .

GA contains a set of gene chromosomes contains solution in gene form it obtained by generating the population for every individuals and evaluate to find fitness of individuals and applies genetic operator mutation and crossover and find the [6] best fitness optimal solution.

Our work is to strongly focus on the hybrid intelligent system was developed by applying feature extraction for efficient tumor classification, ensemble classifiers [7],[8] Weighted vote-based Ensemble model developed and it uses different classifiers resulting in efficient tumor diagnosis.

II. RELATED WORKS

Most of the traditional clustering and classification approaches implemented to gene expression bimolecular data. To improve the performance and efficiency HSVM-Hybrid support vector machine framework is employed and integrate five phases, SVM is grouped by Kernel function for efficient classification, and finally SVM based classification is achieving by constructing the hyper plane for achieving maximum separation between class to recognize class belongs to benign or malignant [9], [10].Linear Discriminate Algorithm (LDA)[11] is supervised pattern recognition, is used to find a discriminative transformation matrix from the high dimensional data space to a reduced dimensionality and applying this prominent features for future classification.

K-Nearest Neighbour is algorithm is calculated by Euclidian distance and it is more efficient and accuracy, easy for calculation it finding the feature space and finding the maximum labels belongs to the same class. [12]K-Nearest is finding the similarity between the testing and the training instances, finally and finding most similar instance by majority instances by applying weighted vote rule.

 DOI : 10.35291/2454-9150.2019.0167

Spectral Cluster [13] is a technique uses for portioning the row of Matrix in their components of Vector matrix, and find the similarities of vertices of graph, rows points the d-dimensional space and column represents as coordinates and clustering fount to be optimal for given matrix.

Gaussian Mixture (GMM)[14] is a classification model for classification compare with other models, we used Gaussian mixture to predict earlier of malicious tumor through probability distance model by applying [15] Cluster matrix and finding the mean value, and finally measuring the maximum like hood function to obtain malicious and non-malicious tumor.

Normalized Spectral Clustering Technique is used to analysis about the tumor and non-tumor.it is used to find the optimal closest clusters centers, and finally the detection of the cluster is obtained through the Gaussian mixture to detect the malicious tumor among obtained clusters.

III. PERFORMANCE OF HSVM, ALDA-EC AND DL-SCGM WITH EXISTING METHODS

In recent years different clustering and classification techniques are used by several researchers. Hybrid fuzzy cluster ensemble framework [16] (HFCEF) employed for tumor clustering from cancer gene expression data .HFCEF framework was the combination of soft clustering and hard clustering into clustering ensemble framework.HFCEF framework involve a process namely, Creation of set of new datasets through affinity propagation algorithm, Consensus function employed to generate fuzzy matrices and obtain result as tumor data or non-tumor data.

Ensemble selection algorithm [17] was employed to select threshold values for reducing the false positive rate below a specified maximum value. Ensemble selection approach algorithm applied to clinical dataset for reducing the performance of classification.

Random Forest using Class Decomposition (RF-CD) [18] method was investigated in for medical diagnosis. RF-CD method is applied in any classification method including single classifier system.

Eensemble system [19] for cancer classification employed to provide solutions for enhancing the accuracy by using ensemble technique to more cancer types and avoiding the problems related to over fitting cancer classification. Ensemble method used to improve tumour classification accuracy and different classifiers used as base members

Feature Selection-based Semi Supervised Cluster Ensemble[20] (FS-SSCE) framework employed for clustering tumor cells from bio molecular data, to significantly improved the accuracy of Random Forests-means clustering was applied to instances that belong to each class by varying number of clusters

FS-SSCE framework adopted feature selection techniques to eliminate the effect of noisy genes. Double selection based semi-supervised cluster ensemble framework (DS-SSCE) applied feature Selection technique to perform gene selection on gene dimension.

Single clustering algorithms not better in finding malicious for high dimension data, thus to improve the efficiency we proposed and used Hybrid Support Vector Machine (HSVM),Aggregate Linear Discriminate Analyzed Ensemble classifiers (ALDA-EC),Deep Learning Spectral Cluster Gaussian Mixture (DL-SCGM).

HSVM is a hybrid technique [21], it achieved by integrating five phases, PCA-Principal compound Analysis is used to obtain reduced dataset contains positive uncorrelated variables and it used for future extraction.

Extracted features given as input for the GA-Genetic algorithm for feature selection, and it identifies feature subset evaluated by fitness function, relevant feature generated by Genetic algorithm, and HSVM framework applies Markov Chain Clustering for GA features by alternating two operations: expansion and inflation, finally nearest clusters are grouped.SVM is uses to for the prediction of benign or malignant tumor.

Aggregate Linear Discriminate Analysis-based Ensemble Classification [22], ALDA-EC is a framework technique used to increase the efficiency and accuracy than HSVM by ensemble more techniques it reduces dimensionality using Iterative Scattering matrix algorithm to find the distance within the class and between the classes, so that we can improve the similar class and find the consistent features for ensemble classifiers.

To overcome the lack of robustness and reduce the classification time in ALDA-EC method we introduced a [23] Deep learning Feature extraction-DLFE, used to improve the robustness and accuracy in less classification time by measuring the element present in the tumor image.

DLFE reduces the redundant features and obtain most prominent features by using mean activation function, obtain relevant features are performed to ensemble clusters for better classification, thus increases the accuracy, efficiency, robustness in minimum classification.

 DOI : 10.35291/2454-9150.2019.0167

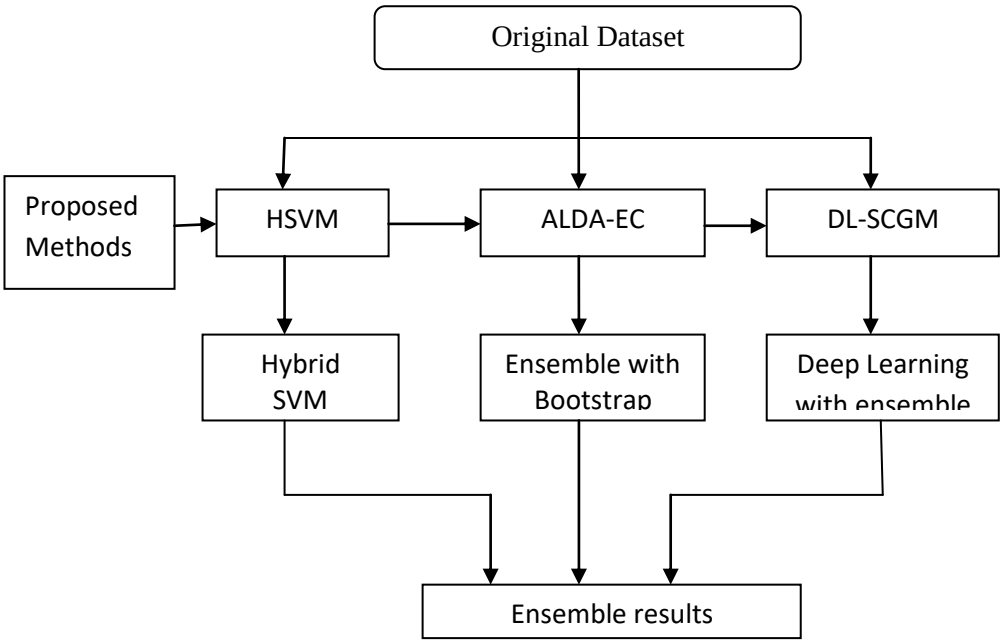

Fig 1.Block of performance analysis of proposed methods

Method	HFC	Eensemble selection-based algorithm	RF-CD	HSVM	ALDA-EC	DL-SCGM
Accuracy (%)	72.67	75.15	81.19	74.90	78.52	85.28
Sensitivity (%)	92.02	90.12	94.32	94.08	96.23	97.19
Sensitivity%	30.62	35.45	42.58	33.25	37.54	44.94
Recall (%)	91.23	87.35	89.17	94.08	94.86	95.50
Error rate (%)	27.32	24.85	18.81	25.09	21.48	14.72

Table 1.Performance Evolution of HFC, ESBA, RF-CD HSVM, ALDA-EC and DL-SCGM using Lung Cancer Dataset

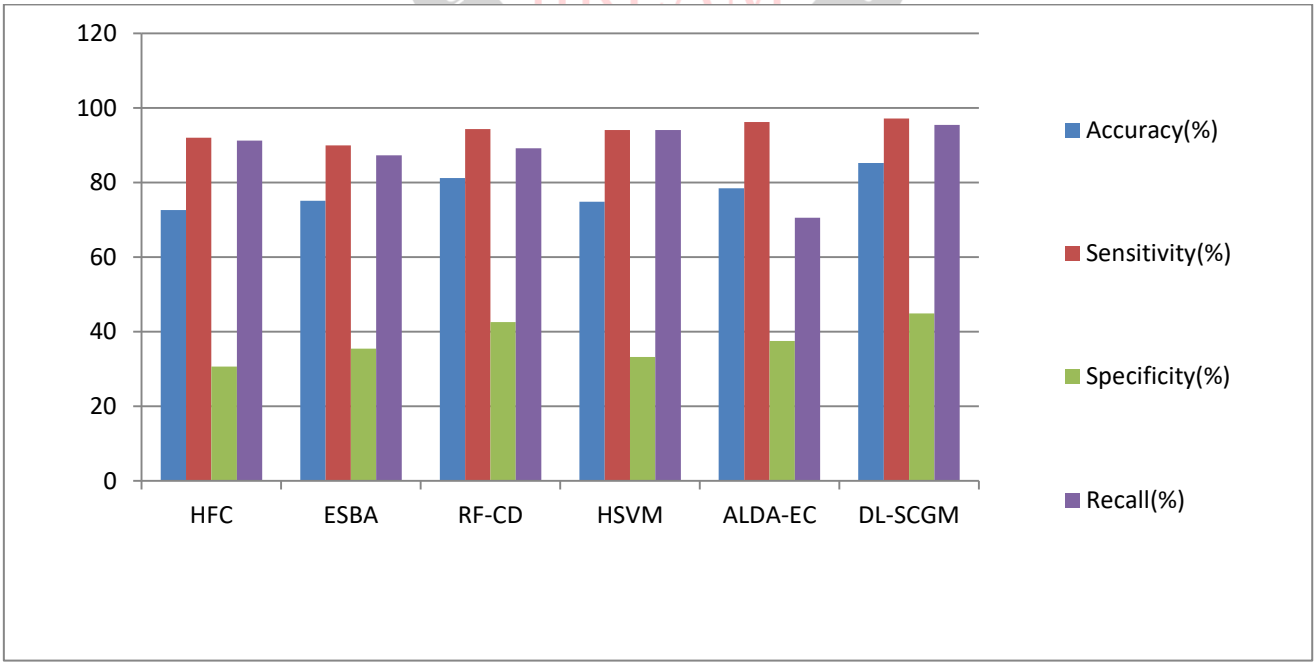

Fig 2.Comparative results of HFC, ESBA, RF-CD, DL-SCGM, RF-CD and DS-SSCE using lung cancer datas

 DOI : 10.35291/2454-9150.2019.0167

Combined results for proposed DL-SCGM using lung cancer datasets explanation

Using the lung cancer dataset accuracy is increased by 85.28%, sensitivity is increased by 97.19%, specificity is increased by 44.94%, recall is increased by 95.50% and time is reduced by 70.62ms in DL-SCGM technique than the other methods.

Using the leukemia dataset accuracy is increased by 76.32%, sensitivity is increased by 72.14%, specificity is increased by 68.13%, recall is increased by 72.15% and time is reduced by 72.65ms in DL-SCGM techniques than the other methods. Proposed DL-SCGM improves the accuracy, sensitivity, specificity, recall better than HFC. Ensemble selection-based algorithm and RF-CD.

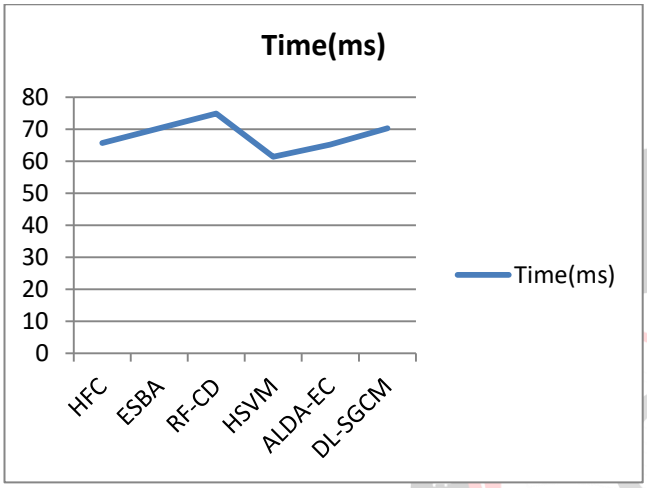

Fig 3.Lung Cancer Dataset using Time

Classification time is defined as time taken to classify the correct samples to the total number of sample cases. Classification time is measured in terms of milliseconds (ms).Classification time is lower, method is more efficient.

Classification time is expressed astime taken to cluster the tumor with respect to the total number of sample cases. Tumor clustering time is measured in terms of milliseconds (ms).Tumor clustering time is lower, method is more efficient.

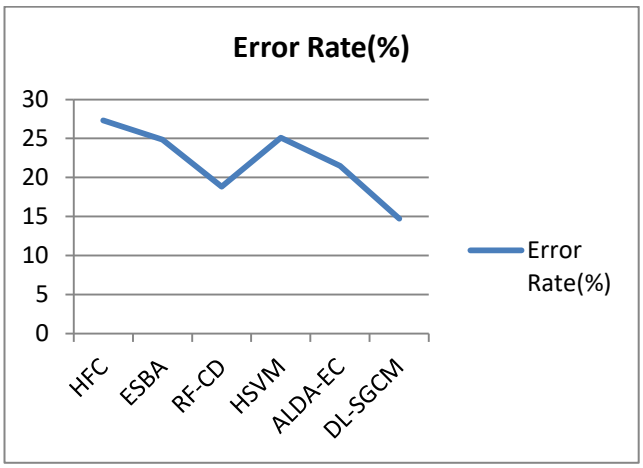

Fig 4. Lung Cancer Dataset using Error rate

Error rate is defined as the number of sample cases that are incorrectly classified to the total number of sample cases. Error rate is measured in terms of (%).Error rate is lesser method is more efficient.

IV. PERFORMANCE EVALUATION

To validate the performance we provide an experimental evaluation of HSVM, ALDA-EC and DL-SCGM techniques, and implemented with java using various experiments conduct on WDBC (Wisconsin Diagnosis Breast Cancer), Breast tissues (BT), Leukemia and Lung cancer datasets.

Confusion Matrix is a performance analysis tool used in supervised learning. It is used to evaluate the test result in the proposed techniques, each column matrix represents predicted class and row represents instance in actual class, few of the equations for performance analysis given below. *Accuracy* (AC) is proportion of the total number of predictions that were positive, it is determined using the equation. *True positive* (TP) *rate* or recall is proportion of correctly identified positive cases, as calculated using equation.

False positive (FP) *rate* is incorrectly classified as positive, as calculated using equation. The *true negative* (TN) *rate* is defined as the proposition of negative cases classified correctly .as calculated by the equation.

The *false negative* (FN) *rate* is the proportion of positive cases that were incorrectly classified as negative, as calculated the equation. Finally *Precision* (P) is the proportion of the predictive positive cases, that were correct, as calculated using the equation.

V. CONCLUSION

In this paper we compare three ensemble techniques for tumor cluster and classification from gene expression datasets to increase the classification efficiency and accuracy we combining several techniques to achieve malicious tumor detection. Ensemble techniques are acquiring most relevant features for tumor classification; HSVM is a hybrid technique extracts more relevant features by constructing hyper plane that uses to separate the maximum possibilities of cancer tumors. Aggregate Linear Discriminate Analysis-based Ensemble Classification ALDA-EC is a framework technique used to increase the efficiency and accuracy than HSVM by ensemble more techniques it reduces dimensionality using Iterative Scattering matrix algorithm to find the distance within the class and between the classes, so that we can improve the similar class and find the consistent features for ensemble classifiers. To overcome the lack of robustness and reduce the classification time in ALDA-EC method we introduced a Deep learning Feature extraction-DLFE, used to improve the robustness and accuracy in less

 DOI : 10.35291/2454-9150.2019.0167

classification time by measuring the element present in tumor image. DLFE reduces the redundant features and obtain most prominent features by using mean activation function, obtain relevant features are performed to ensemble clusters for better classification, thus increases the accuracy, efficiency, robustness in minimum classification.

VI. FUTURE WORK

In future we will explore how to improve the performance of existing technique. We examine the performance of existing technique by implementing LogitBoost algorithm to build a strong ensemble classifier to obtain minimum likehood for maximum accuracy by aggregating the sequence of weak hypothesis.

REFERENCES

[1]T.R.Golub,D.k.Slonim,P.Tamayo,C.Huard,M.Gaasenbeek,J.P. Mesirov,H.Coller,M.Loh,J.Downing,M.Caligiuri,C.Bloomfield,a n

d E.Lander,"Molecular classification of cancer: Class discovery and class prediction by gene expressions", Science,Vol.286, no.5439, pp.531-537,1999.

[2] K.S. Leung, K.H.Lee, J.-F.Wang, et al., "Data Mining on DNS Sequences of Hepatitis B Virus", IEEE/ACM Transactions on Computational Biology and Bioinformatics,vol.8,no.2,pp.428-440,2011.

[3] Xu R. and Wunsch D, "Survey of clustering algorithms", IEEE Transactions Neural Networks,Vol.16, No.3, pp.645-678.

[4] C.Deisy,S. Baskar, N.Ramraj, J.S.Koori, and P.Jeevanandam, "A novel information theoretic-interact algorithm(IT-IN)for feature selection using three algorithms", Experts systems with Application,Vol.37, no.12, pp.7589-7597,2010.

[5] Gumus, E.,Kilic, N., Sertbas, A., & Ucan, O. N.," Evaluation of Face Recognition technique using PCA,Wavelets and SVM", Expert Systems with Applications.Vol.37, pp.6404-6408.

[6] Farser A.S.,"Simulation of genetic systems by automatic digital computers and Effects of linkage on rates under selection", Austral.J.Biol.Sci.Vol.10, 1957, pp.492-499.

[7] Subrata Kumar Mandal, "Performance Analysis of Data Mining Algorithms for Breast Cancer Cell Detection Using Naïve Bayes,Logistic Regression and Decision Tree" ,International Journal of Engineering and Computer Science, Vol 6,Issue 2, Feburary 2017, pp 20388-20391.

[8] Milan Joshi,Anuraj Joshi, "On Comparitive Study of Breast Cancer Classification Using Ensembles in Stastical Modelling", International Journal of Computer Science and Technology, Volume 8,Issue 1,March 2017,pages 18-21.

[9] Issam EI-Naqa ,Yongyi Yang,Miles N.Wernick, Nikalos P.Galatsanos and Robert M.Nishikawa,"A Support Vector Machine Approach for Detection of Microcalcifications",IEEE Transactions on Medical Imaging,Vol.21,Issue 12,Dec 2002,Pages 1552-1563.

[10] Yukai Yao,Hongmei Cui,Yang Liu,Longjie Li,Long Zhang,and Xiaoyun Chen,"PSVM:An Optimized Support Vector Machine Classification Algorithm Based on PCA and Multilevel Grid Search Methods"Hindawi Publishing Corporation Mathematical Problems in Engineering,Vol 2015,Article ID 320186,Pages 1-15

[11] George Saon and Mukund Padmanabhan, "Minimum Bayes Error Feature Selection for Continuous Speech Recognition ", Advance in neural Information Processing System, Vol.13,2001, Pages 800-806.

[12] Aman Kataria,M.D.Singh,"A Review of Data Classification Using K-Nearest Neighbour Algorithm"International Journal of Emerging Technology and Advance Enginering",Vol.3,Issue 6,June 2013,Pages 354-360.

[13] Wei Zheng,HaiDong Wang,Lin Ma ,Ruo Yi Wang,"An Improved K-Nearest Classification Algorithm Using Shared Nearest Similarity"Journal of Computational Physics,Vol.10,2015,Pages 133-137 .

[14] Ravi Kannan and Santhosh Vempala and Adrian Vetta, " On Clustering:Good, Bad and Spectral" Journal of the ACM, Vol.51, Issue 3, May 2004, Pages 497-515.

[15] Haitain Ling, Kunping Zhu, "Predcting Precipitation Events Using Gaussian Mixture Model" Journal of Data Analysis and Information Processing, Vol.5, Issue 10, Oct 2017, Pages 131-139.

[16] ZhiwenY u, JaneYou, HantaoChen, Le Li, Xiao Wei Wang,"Tumor Clustering Based on Hybrid Cluster Ensemble Framework in Computerized Healthcare (ICCH), International Conference on 2012.pages 95-101

[17] Zhiwen yu, Hongsheng Chen,Jane You, Hau-San Wong,Jiming liu,Le Li and Guoqiang Han, "Double Selection Based Semi-Supervised Clustering Ensemble for Tumor Clustering from Gene Expression profiles,IEEE/ACM Transactions on Computational Biology and Bioinformatics April 2011.pages1108-1122.

[18] Eyad Eylan, Mohamed Medat Gaber,"A Fine Gradient Random Forests using Class Decomposition: an application to medical Diagnosis", Neural computing and Application, Issue Sep2015, Page674-682.

[19] YunpengLi, Emily Porter, Adam Santorelli, Milica Popovic, Mark Coates, " Microwave breast cancer detection via cost-sensitive ensemble classifiers: photonand patient investigation", Elsevier, Biomedical Signal Processing and Control, volume 31,January 2017, Pages366-376.

[20] Eyad Elyan and Mohamed Medhat Garber, "A Fine-Grained Random Forests using Class Decomposition: An Applications to Medical Diagnosis", Neural Computing and Applications, vol.27, No.8, 2015, Pages 2279-2288.

[21] S.Subashchandrabose, T.Christopher, Hybrid Support Vector Machine based Markov Clustering for Tumor Detection from Bio-Molecular Data",Asian Research Publish Network(ARPN),Volume.13,Issue 9,May 2018,Pages 3270-3279

[22] S.Subashchandrabose, T.Christopher, "Deep Learning Extraction with Ensemble Spectral Cluster and Guassian Mixture for Malicious Tumor Detection", ICACT Journal on Sotf Computing, Volume.08, Issue 4, July 2018, Pages 1750-1757.

 DOI : 10.35291/2454-9150.2019.0167