International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-2, July 2019

Classification of Gene Expression Data with Optimized Feature Selection

T.Ragunthar, S.Selvakumar

Abstract—There are different types of fatal diseases that could possibly outspread to various parts of the body. It thus becomes obligatory to predict the existence of such anomalies, in order to prune the extent of their spread. Examining the characteristics of genes provides a deep intuition about the disease classification, as they play a vital role in influencing how an organism appears, behaves and survives in an environment. The detection of the abnormal genes could be efficiently modelled using statistical methods and machine learning approaches. Gene expression data derived from a microarray could act as an aid for this statistical computation. Microarray being a recent leap in molecular biology, provides a scope for hybridization of DNA samples that can be interpreted as values based on the gene expression level that the genome possesses. We propose an idea to select a subset of features from the huge number of samples retrieved from the gene expression profiles using Boruta feature selection algorithm. A comparative study with various supervised classification algorithms is made to categorize this subset to a normal and deviant gene. This serves to discover the most appropriate algorithm to classify the gene expression data. Hence assorting the abnormal genes in future could be accelerated with ease.

KEYWORDS- Boruta algorithm, DNA samples, Feature selection, Gene expression data, Kernel, Machine learning, Microarray, Random forest, SVM.

I. INTRODUCTION

A. Gene Expression Data And Microarray Technology

Diseases are caused because of division of cells or uncontrolled growth due to cellular changes. In order to form new cells usually cells receive information to die. On the other hand, the cancer cells would lack the component which would instruct them to stop dividing and instead die. As a result, they can form tumors, impairing the immune system. Genetic factors can contribute to various deadly diseases, as it is a person's genetic code that instructs if a cell has to divide or expire. Every cell in our body consists of same no. of genes as well as similar type of genes.

It is the gene expression of each cell that distinguishes between normal and affected cells. Based on the environment, the gene expression varies. To check the gene expression the two-phenomenon involved are i) Produce microarray ii) Measure transcriptome.

Revised Manuscript Received on 30 July 2019.

* Correspondence Author

T.Ragunthar*, Assistant Professor, Department of Computer Science & Engineering ,Sri Sairam Institute of Technology Chennai, India

S.Selvakumar, Professor, Department of Computer Science & Engineering ,GKM College of Engineering & Technology, Chennai, India

There are many technologies such as microarray, illumine bead array, nylon membrane, serial analysis of gene expression (SAGE), high-densities oligonucleotide arrays etc. used to express the level of genes.

The varying gene expression can be efficiently analyzed using microarray where all the genes of a particular organism are placed in different grooves on a slide. Microarrays are group of DNA spots on a solid surface, like glass or silicon in which hybridization of DNA samples can be made ordered arrangement of samples done using base pairing rules wherein matching familiar and unfamiliar DNA samples is followed, forms the microarray. Each microarray consists of thousands of pores known as probes. The two key terms for microarray synthesis are the blocking agent and the mask. The blocking agent prevents the binding of a nucleotide with some other nucleotide. This blocking agent can be removed using a laser. Masking leaves behind gapes in the microarray spots while the rest of places are masked and never be bind. On observing the colour of each probes in microarray using analyzer, the attributes of gene expression data are determined. Biological interpretation of gene expression data can be made using heatmaps. The heatmap can be combined with clustering techniques for grouping similar genes. Identifying similarly regulated genes can thus become easier.

B. Classification Of Genome Profiles Using Statistical Methods

The gene expression data usually has got very high dimensionality due to which biologists find it difficult to handle them [1]. Hence classification of such microarray data can be cumbersome. Also, there might be noisy data present in the gene expression dataset along with some irrelevant features. Statistical approaches could be an optimal solution to this problem[2]. In recent years, there have many statistical approaches with various level of complexity to analyze genotype data and detect variations in gene. In order to avoid the manual computation difficulties and errors that are likely to occur in such huge datasets it is advisable to automate the statistical computation. Such an approach can be obtained with the help of machine learning. This method would make the system learn through experience and later make the predictions based on the learning.

Machine learning is mainly classified into three algorithms namely supervised, reinforcement and unsupervised learning. Supervised learning is helpful in predicting the target resultant variable based on the input independent variables. Unsupervised learning does not have such target variable instead they form clusters to group similar data together. Past experience is used to predict the future based on trial and error approach in reinforcement learning.

Firstly, before handling the gene expression dataset for classification or clustering it is mandatory to reduce the dimensionality. There might be many irrelevant attributes present in the dataset along with noise and disturbances. Thus, pre-processing becomes mandatory.

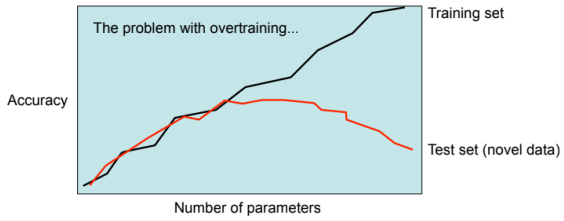

Fig 1: Problem with over training

The initial process to deal with the dataset is to prepare the data for further computations.It is quality of data that decides the strength of the model. A simpler model with clean data will generally outperform a complicated model basedon dirty or ambiguous data.One of the key role of data preparation is to ensure that information being prepared for accurate and consistent analysis, so the results of Business Intelligence (BI) and analytics applications will be valid. Data is often created with inaccuracies, missing values or other errors. Also, there exists different form of datasets. The process of performing verification, correcting inaccuracies and combining data setscontributes to major part of the data preparation phenomenon. Min-Max normalization technique specifies theformula(X - min(X))/(max(X) - min(X))to be applied to each value of features to be scaled. This method is traditionally used withK-Nearest Neighbors (KNN) Classification problems. To predict outliers Z-Score calculations can be made. Transformation of raw data into features suitable for modelling can be made through feature extraction. Feature selection removes unnecessary features while, Feature transformation helps to transform data to improve the accuracy of the algorithm.

Feature selection is a primary step in machine learning as this would help to improve the performance of the model trained. The most relevant feature that truly contribute to the output computations can be easily sort out from a very huge data sample. Pros of this method includes reduced over-fitting, improving accuracy, reduction in training time required and reduced complexity of the model. Feature selection techniques are generally classified as Filter method, Wrapper method and Embedded method. Features are selected based on their correlation with outcome variable in filter methods. While in the case of wrapper class model, subset of features is used for training and based on the results of previous model addition or removal feature is made. Embedded algorithms are combination of both filter and wrapper class algorithms that built-in feature selection methods like LASSO and RIDGE regression. Hence the algorithms need to be mandatorily applied to the dataset for reducing the dimensionality of the gene profile thereby improving accuracy and performance of the model. Scaling also plays an important role in predicting the accuracy of results ass they improve the strength of the trained model. This is generally applied when the range of input varies widely. Hence, it is required to normalize or scale values for different features such that they fall under a common range.

Analyzing the subset obtain becomes the next key step. To predict the existence of a disease it is necessary to either classify or cluster them. These help to form a logical structure examining the similar groups. For those data with defined attributes contributing to supervised learning method, classification can be used. While clustering is unsupervised learning that groups data on similarity from the experience it learns. Some of classification algorithms include logistic regression, random forest, SVM, KNN etc. K-means, hierarchical, density-based are examples of clustering methods. To improve the performance measure namely the accuracy, sensitivity, kappa value obtained through classification ensemble methods are used. Bagging, Boosting or stacking can be made to compare the enhancement made on the classification model for better prediction outcomes. Majority voting, weighted voting are the different approaches defined for the working of such ensemble methods to improve the accuracies of classification algorithms.

II. RELATED STUDIES

Several approaches have been followed to apply data analytics to gene expression profiles. First applied by Barnard (1935) at the suggestion of Fisher (1936) Fisher's Linear Discriminant Analysis (FLDA) a method that does not uses parameters instead it computes a matrix called projection that modifies the dataset for maximizing the differences between classes termed as class separability was proposed. Class Separability is the ratio of between class scatter matrix and within class scatter matrix. Weighted voting of informative genes is used to classify Binary class data. This method classifies data using the correlation between the genes and the class parameters. The gene that has high correlation with that of the class parameter is called the informative gene. This method is proven efficient for gene expressions with large range and has low variations within a class and high variations between the classes. The sign of the parameter that denotes the level of correlation decides the class in which the gene is to be classified.

There are many artificial intelligence approaches also. Probabilistic induction: Naive Bayes method uses probabilistic induction to classify the data based on the independence among the attributes of the dataset: Neural networks was used for cancer classification by J. Khan. This method comprises of Principle Component Analysis followed by Relevant Gene Selection and finally ANN Prediction. The initial principle component analysis is used to evade the 'Overfitting' problem. This method concluded that the class labels produce bias in the data while included to the reduction process. Thus, it is better to exclude class labels in the process of reduction. Cross validation was done in three folds for the prediction procedure.

Decision tree also called the classification tree is one of the familiarly used classification methods. A decision tree is a collection of internal and leaf nodes. The nodes of decision trees that are internal denote split criterion attribute which splits into one or more split attributes. The leaf nodes denote separate class labels. The construction of decision trees usually involves two phases. In the first phase each, the dataset is split at each internal node.

In the second phase to avoid the over fitting problem, the tree is pruned.Sharmila et al. had motivated to identify the presence of any abnormal pattern in the gene expression data and used a method called Regular Expression based Pattern Matching for this [3]. Hu et al. proposed methods for analysing classification performances of cancer cell with the help of unsupervised and supervised learning methods[4]. Selection of small set of features from microarray data for cancer classification was proposed by A. Bharathi, A. M. Natarajan that involved the use of supervised learning approach called analysis of variance (ANOVA).

Dey et al. had developed an automated system for measuring a haematocrit level form RBC count and a comparison was made with the help of traditional method[5]. This automation is similar to that in laboratories, hence the system can overcome manual errors that are likely to occur.Krampis et al., presented the publicly accessible Cloud BioLinux similar to publicly available virtual machine (VM) to facilitate the on-demand infrastructures for the high-performance bioinformatics [6]. The pre-configured command line consists of a featured desktop interface with graphic software application, and a documentation with135 bioinformatics packages used for the display, sequence alignment, editing, assembly, clustering, and phylogeny. It is mandatory to define the functionality of every tool in the VM interface.

Several other classifications such as T-test was conducted along with SVM for this subset of data and the higher accuracy of ANOVA and SVM was proved. SudipMandal and Indrojit Banerjee used artificial neural network concept for detection of cancer. There exist a class discovery methodthat discovers the distinction between acute myeloid leukaemia (AML) and acute lymphoblastic leukaemia (ALL) automatically without previous knowledge**.** Diffuse large B-cell lymphoma resultant prediction was presented by Shipp M. A et al. using gene expression profiles and supervised machine learning approaches. Karaboga D used artificial bee colony(ABC) algorithm for obtaining optimum feature selection technique where in artificial bees discover the initial solution vectors randomly and improve by moving towards better solutions iteratively using neighbour search mechanism, abandoning poor solutions [7] . ABC-SVM proved to be much efficient approach yielding better results compared together wrapper class algorithms because of its convergence property[8]. Thus, several other researches were done top predict diseases from gene expression data mainly emphasising feature selection for pre-processing followed by clustering or classification that can be categorized as class discovery and prediction[9]. This plays a key role with gene expression data as it serves to be of great use for medical applications [10].

III. METHODOLOGY

3.1 DATASET

Gene expression omnibus (GEO) is a repository for storing curated gene expression datasets. The GDS230 dataset is used to retrieve the gene expression profiles of 160 samples with 12625 attributes defined for each observation. Also, those samples that suffers from the disease and those which are normal are defined explicitly in this dataset. To classify the dataset using supervised learning approaches, it is obligatory to have a predictor variable. Thus, based on the defined knowledge about a sample, we defined an additional column to the dataset with binary values, 0 for absence of disease and 1 if disease is present to the transposed GDS230 dataset. Hence, 160 X 12626-dimensional csv file is recovered.

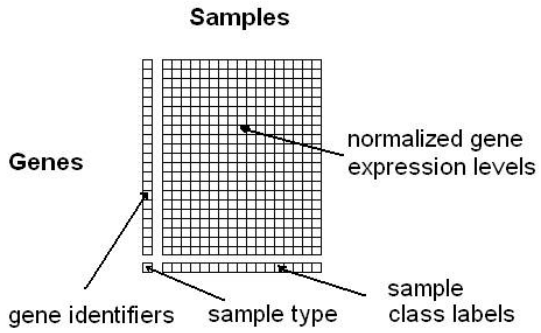

Fig 2: Structure of gene expression dataset

3.2 BORUTA ALGORITHM

Boruta is a wrapper class algorithm around random forest. It adds randomness to the dataset by shuffling data, thus forming shadow features. Boruta can handle interaction between variables as well as fluctuating nature of random forest measure. Random forest classifier is used on this altered dataset to compute the feature importance through Mean Decrease Accuracy. Those features with higher mean are taken to be important. Iteratively it compares its real value mean with that of shadow features to remove the unimportant features. The end of algorithm is reached if it reaches a specified random forest run limit or when all the features present gets rejected or confirmed. This algorithm follows all-relevant feature selection method that relates every feature that is associated with outcome predictor variable. Hence it becomes widely different from traditional approach that usually rely on small subset to find out minimal error on a selected classifier.

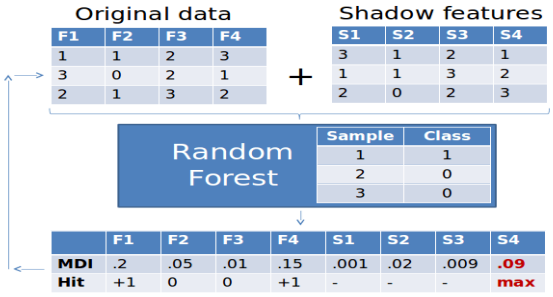

Fig 3: Computation of shadow features

After 99 iterations of the algorithm for the input dataset given, Boruta interpreted tentative important attributes. TentativeRoughFix function then confirmed the most important attributes out of these selected tentative features. Hence the final subset of feature selection resulted in 160 observations withmost important features. The dimensionality of the dataset was thus minimized.

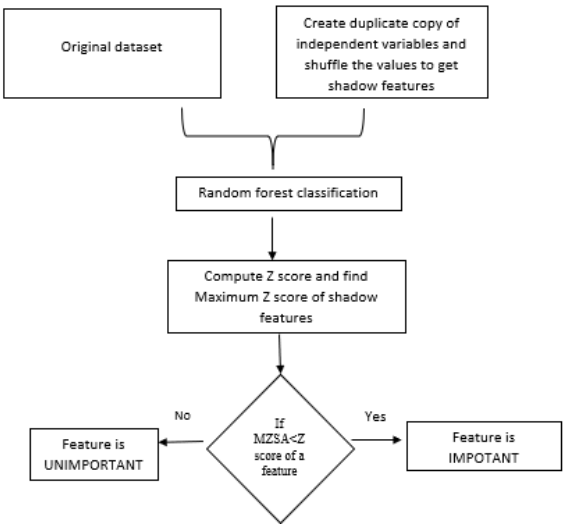

Fig 4: Workflow of Boruta

The new subset could then be classified by breaking the dataset into training and test set. Several supervised learning algorithms such as SVM, KNN, Random forest was tested for their accuracy.

3.3 SUPPORT VECTOR MACHINE

Support Vector Machine (SVM) are supervised learning approaches generally used for regression and classification [11]. It classifies the datapoints by finding a N-dimensional hyperplane where N denotes number of features available. A hyperplane is a straight line which divides the input variable space. In support vector machine, generally a hyperplane is selected to separate the datapoints into class 0 or class 1. These serve to be powerful tools with the aid of kernel ideas to large margin classifiers. If there exists a hyperplane, (w, b) where w is a vector and b is a scalar, two classes are linearly seperable such that for point x_i

$$\begin{cases} w^T x_i + b >= 0 \text{ for } c_i = 1 \\ w^T x_i + b <= 0 \text{ for } c_i = -1 \end{cases}$$

SVMs are widely used in engineering and research areas ranging from breast cancer diagnosis, recommendation system, detection of protein homologies, text categorization, database marketing, face recognition, etc. [12]. These contribute in forming the general framework of SVMs. Thus scope for applications of SVM is large. For classifying the gene expression data, production of hyperplane with maximized margin and the amount of training errors by SVM is modified.

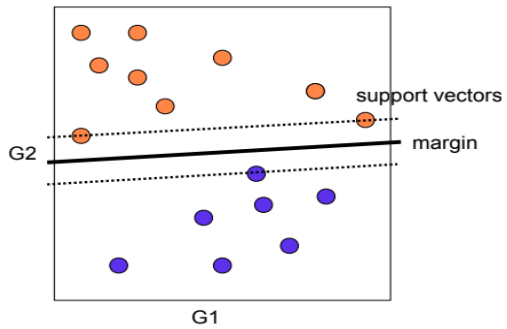

Fig 5: Linear SVM

The SVM will not workfor non-linear classification and for overcoming this limitations kernel are used .When a linear classifier $f^k : R^d \rightarrow \{-1, +1\}$ of the form

$$f^k(x)=sgn(g^k(x)) =sgn(x^T w^k+b^k)$$

is normalized such that the Euclidean norm $||w^k||^2$ is 1, $g^k(x)$ gives the Euclidean distance from x to the boundary of f^k[9].

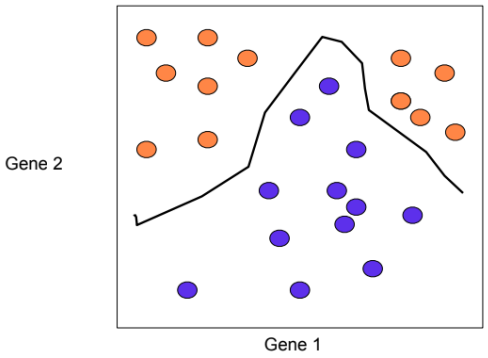

Fig 6:Non-Linear SVM

The trace of hyperplane in linear SVM is done by transformation of the problem with aid of linear algebra. The kernel plays vital role here. For **linear kernel** the equation for prediction between the input (x) and each support vector (xi) is calculated as follows:

f(x) = B(0) + sum(ai * (x,xi))

The above equation involves calculating the inner products of a new input vector (x) with all support vectors present in training data. The coefficients B0 and ai must be calculated from the training data using the learning algorithm. The **polynomial kernel** can be written as

$$K(x,xi)=1+sum(x*xi)\text{^}d$$

and **exponential** as

$$K(x,xi) = exp(-gamma * sum((x—xi^2))$$

To test the datasets SVM with linear kernel, neural networks SVM with RBF kernel, SVM with MLP kernel Decision tree, and Naïve Bayes are used.Several methods are used to compare the end result to improve the accuracy and compute the impact of k in k-fold cross validation [13].A hyper plane is defined by the equation, where bias is a point lying on the hyperplane and normal to it. For the case of linearly separable, a separating hyperplane can be defined for the two classes. These two equations can then be integrated. The training data points present on these two hyper planes are referred as support vectors and are typicalfor the establishment of the optimal separating hyperplane. Several statistical parameters were used for evaluation of the accuracy of the SVM.So they have, in general, better performance than hold out.

3.4 K-NEARESTNEIGHBOUR ALGORITHM

K-NN is a lazy learning technique where the function is locally approximated and all execution is postponed until classification [14].

The final classification is decided by majority of its neighbors. The average target value of the nearest problems are the key factors for numeric class problems. Themajor features of this method include calculation time, predictive power and interpretation of output with ease. Steps involved in KNN is as follows: Load the data. Initialize the value of k. To obtain the predicted class, iterate recursively from 1 to total number of training data points. Compute the distance between every row of training data and the test data. The most popular method Euclidean distance is used as distance metric. Other metrics such as Chebyshev, cosine, etc. Based on the distance value, the calculated distances is sorted in ascending order. From the sorted array, get the top k rows. Identify the most frequent class of these rows. Return the predicted class.

KNN's main disadvantage is it becomes slower as the data volume increases, thus making it an impractical choice in environments where predictions need to be made rapidly.There are better algorithms which can produce more accurate classification and regression results. KNN can be useful in solving problems that have solutions that dependon identifying similar objects. An example of this is recommender systems which is an application of KNN-search. It improves the Performance of Decision Trees on Test Set and also to reduce the variance, thus eventually avoid Overfitting. The idea focusses on building many Trees in such a way to make the Correlation between the Trees smaller.

The performance of thek-NN classifierhugely depends on the distance metric. The Euclidean distance is most commonly used. The diagnosis based on historical data can be made with ease based on this data. It emphasises on computing the probability of occurrence of a particular ailment with the help of a unique algorithm. The accuracy of such diagnosis can thus be improved by KNN. The algorithm can thus act as an aid to enhance the automated diagnoses, that includes diagnosis of multiple diseases which shows similar symptoms.

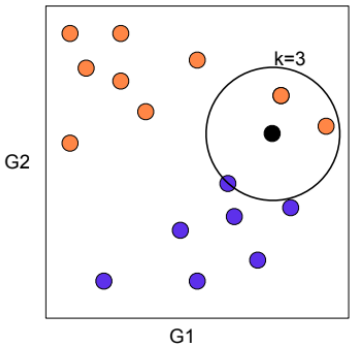

Fig 7: KNN

3.5 RANDOM FOREST CLASSIFICATION

Random forests are generally ensemble methods in which base learners are grown as trees and combine then predictions by averaging. Random forests have good practical performance, specifically in high-dimensional settings. Random Forests are similar to Ensemble technique called Bagging but these have a different tweak in it. Here, the Trees generated on different bootstrapped samples from training data are decorrelated. And then we simply reduce the Variance in the Trees by averaging them. Random forest can be used for bothregression as well as classification problems. Random Forest is also as a handy and very easy to use algorithm, since it's default hyperparameters often produce a good prediction result[14]. The number of hyperparameters neededis also not too high and could be understood with ease.

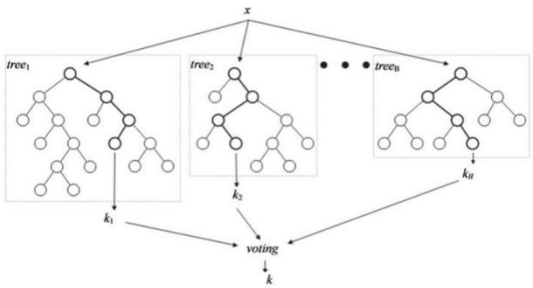

Fig 8: Random forest classification

The random forest algorithm finds application in different fields, like Banking, Stock Market, E-commerce and Medicine. In bank it is used to identify frequent users or those who have debt for long duration of time. Fraud customers who try to scam the bank can also be detected. A stock's behaviour inthe future can also be predicted. Whereas in the healthcare domain it is used to predict the correct combination of components in medicine and to examine a patient's medical history to predict diseases. It also finds its use in E-commerce technology.

IV. IMPLEMENTATION

The input dataset was initially fed into Boruta algorithm so as to feature select the huge lot of attributes present. After tracing 99 iterations Boruta algorithm yields a tentative 16 important attributes. On passing these through tentative rough fix function, confirmed set of important attributes are thus retrieved. This subset of data is cross-validated to split into training and test set with 75% split ratio.Each of these subsets are normalized to yield accurate results.

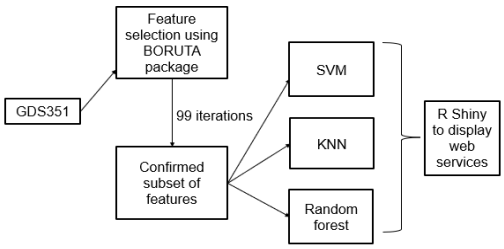

Fig 9: Architecture diagram for proposed system

The classification accuracy of the SVM, KNN, Random forest algorithm was checked. It was found that random forest classification was found to yield more accurate results. Front end design could be done using Rshiny library. This supports UI and server modules. It deals with the outer layout of the website that would display browse button for searching the file tobe loaded and displays the computed result. Server runs the backend process of training the model using the defined dataset. The enter input is retrieved from text box and tested with model designed. The output is later displayed on the website screen if the gene expression data fed suffers from a disease or if it is a normal gene.

Thus disease prediction would become much easier with the genome profiles. Human intervention and manual errors can also avoided to a larger extent hence facilitating diagnosis phenomenon.

V. RESULT

Confusion matrix could display the specificity, accuracy, sensitivity of the dataset fed in depending on the comparison made with the trained model. True positive (TP), True negative (TN), False negative(FN), False Positive(FP) values forms the matrix. The predictor variable is computed by comparing the attribute values in the training dataset.

Table 1 Classification accuracy

CLASSIFICATION ACCURACY	GDS230
SVM	0.8126
KNN	0.8654
RANDOM FOREST	0.907

Table 2 Sensitivity

SENSITIVITY	GDS230
SVM	0.555
KNN	0.643
RANDOM FOREST	0.756

Table 3 Specificity

SPECIFICITY	GDS230
SVM	0.764
KNN	0.825
RANDOM FOREST	0.865

Thus Random forest classification algorithm serves to give the most optimal results. Also this could be improved using ensemble methods of bagging, boosting and stacking. However, for such disease oriented prediction stacking would be the best choice to work with. Hence when this dataset is stacked it would result in much accurate results and hence diagnosis would become more valid. The below graphs depicts the bar graph of the accuracy, specificity and sensitivity observed from each of the 3 algorithms. This could provide a clear insight of the optimal algorithm for such disease prediction using gene expression dataset. More the value of sensitivity and specificity greater would be the accuracy of prediction and thus analysis could be effective.

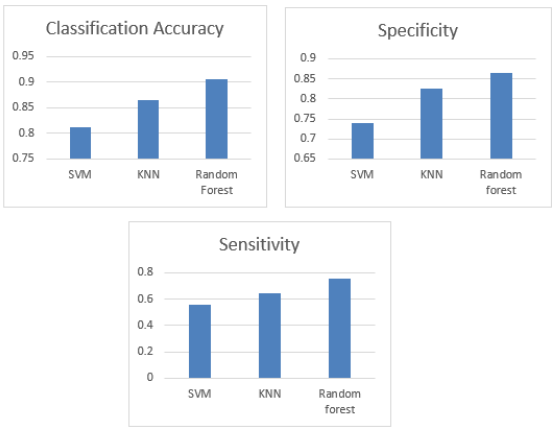

Fig 10: Bar graph for classification accuracy, specificity and sensitivity

VI. CONCLUSION AND FUTURE WORKS

It would be much efficient to make use of cloud technology to make this predictive analysis [15]. As storing such huge datasets on a single system requires high processing and also updating of sample data every time is cumbersome, it is rather preferable to bring in the concept of cloud computing. This would definitely render greater performance throughput as the data tested can be easily saved to the training dataset every time and hence the prediction accuracy in future would become much optimal. Thus, gene expression data could be effectively maintained and processed using statistical methods to make the analysis about diseases much easier. The condition of a cell expressed by mRNA content can thus serve to be of great help to determine if a cell is a normal or variant one. This could obviously be the helping aid in medical industry to handle data much quicker without the human intervention and also preventing the manual errors that could possibly occur.

REFERNCES

1. Bennet, J., Ganaprakasam, C., Kumar, N.: A hybrid approach for gene selection and classification using support vector machine. International Arab Journal of Information Technology (IAJIT) 12, 695–700 (2015)
2. Lakhani, Sunil R., and Alan Ashworth. "Microarray and histopathological analysis of tumours: the future and the past?." Nature Reviews Cancer 1.2 (2001): 151-157.
3. Sharmila, L., Sakthi, U., Geethanjali, A., Sagadevan, S.: Regular expression based pattern matching for gene expression data to identify the abnormality gnome. In: 2017 Second International Conference on Recent Trends and Challenges in Computational Models (ICRTCCM), pp. 301–305. IEEE (2017)
4. Hu, Y.; Ashenayi, K.; Veltri, R.; O'Dowd, G.; Miller, G.; Hurst, R.; Bonner, R.; "A comparison of neural network and fuzzy c-means methods in bladder cancer cell classification", IEEE International Conference on Neural Networks, IEEE World Congress on Computational Intelligence, 1994, Vol. 6, Pp. 3461 – 3466.
5. Dey, R., Roy, K., Bhattacharjee, D., Nasipuri, M., Ghosh, P.: An automated system for measuring hematocrit level of human blood from total RBC count. In: 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 2273–2279. IEEE (2016)
6. Krampis, K., Booth, T., Chapman, B., Tiwari, B., Bicak, M., Field, D., Nelson, K.E.: Cloud BioLinux: pre-configured and ondemand bioinformatics computing for the genomics community. BMC Bioinform. 13(1), 42 (2012)
7. Karaboga, D., Basturk, B.: A powerful and efficient algorithm for numerical function optimization: artificial bee colony (ABC) algorithm. J. Glob. Optim. 39, 459–471 (2007)
8. Shanthi, S., Bhaskaran, V.M.: Modified artificial bee colony based feature selection: a new method in the application of mammogram image classification. Int. J. Sci. Eng. Technol. Res. 3(6), 1664–1667 (2014)
9. Todd R., et al. "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring." science 286.5439 (1999): 531-537.
10. D. Slonim, P. Tamayo, J. Mesirov, T. Golub, and E. Lander. Class prediction and discovery using gene expression data. In Proc. 4th Int. Conf. on Computational Molecular Biology (RECOMB), 2000, pages 263–272.

11. Mayoraz, E., &Alpaydin, E. (1999). Support vector machines for multi-class classification. Engineering Applications of Bio-Inspired Artificial Neural Networks, 833 842.doi:10.1007/bfb0100551
12. Kwang In Kim, Keechul Jung, Se Hyun Park, & Hang Joon Kim. (2002). Support vector machines for texture classification. IEEE Transactions on Pattern Analysis and Machine Intelligence,24(11),1542-1550.doi:10.1109/tpami.2002.104617.
13. De Lacerda, E. G. M., de Carvalho, A. C. P. L. F., &Ludermir, T. B. (n.d.). A study of cross-validation and bootstrap as objective functions for genetic algorithms. VII Brazilian Symposium on Neural Networks, 2002. SBRN 2002. Proceedings. doi:10.1109/sbrn.2002.1181451
14. Min-Ling Zhang ,Zhi-Hua Zhou: ML-KNN: A lazy learning approach to multi-label learning. Pattern recognition Volume 40, Issue 7, July 2007, Pages 2038-2048,Elsiever.
15. Ragunthar, T & Selvakumar, S , 'A wrapper based feature selection in bone marrow plasma cell gene expression data', Cluster Computing, https:doi.org10.1007/s10586-018-2094-2,2018

AUTHORS PROFILE

T. Ragunthar is an Assistant Professor in the Computer Science & Engineering at Sri Sai Ram Institute of Technology where he has been a faculty member since 2009. Ragunthar is serving in the field of academics for almost 10 years with eminent lectures and project guidance. He completed his B.Tech. in Information Technology at K.S.R College of Technology and M.E in Computer Science &Engineering at Bannari Amman Institute of Technology.

S. Selvakumar received Doctor of Philosophy in Computer Science and Engineering from the Anna University. He received the Master Degree in Computer Science & Engineering from the Madurai Kamaraj University. He received the Master of Business Administration from the same University. He is working as Professor in GKM College of Engineering & Technology, Chennai. His current research interests includes Big Data Analytics, Software Engineering,