International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-3, September 2019

Gene Classification using Effective Random Forest Bootstrap Technique for Predicting the Gene Abnormalities

A. Immaculate Mercy, M. Chidambaram

Abstract*: Gene classification is an increasing concern in the field of medicine for identifying various diseases at earlier stages. This work aims to specifically predict the abnormalities in human chromosome-17 by means of effective random forest bootstrap classification. The homo-sapiens dataset is initially preprocessed to remove the unwanted data. The enhanced data undergoes training phase where the appropriate and relevant features are selected by wrapper and filter methods. Based on the feature priorities, decision trees are formulated using random forest technique. The statistical quantities are estimated from the samples and a bootstrap sampling is designated. The effective bootstrap technique classifies the gene abnormalities in chromosome-17. The performance metrics are evaluated and the classification accuracy value is compared with the values of existing algorithms. From the experimental results, it is proved that the proposed method is highly accurate than the conventional methods.

***Keywords*: Genes, Gene Classification, Gene Properties, Random Forest Bootstrap Technique.**

I. INTRODUCTION

A diagnostic group of model is foretold with the expression array phenotype which is termed as microarray classification of data performing supervised learning task. A classifier model has been generated where the new data samples are classified as several predefined disease. Earlier gene classification methods are homology-based approaches and statistical approaches. But it still having issues to attain classification and accurate alignment. In homology based approaches, there is no possibility of acquiring information from synonymous mutation, structural features of proteins utilized for classifying the genes neglects synonymous mutation. So far feature selection is one of the main problem in Gene classification. Feature selection can be done with two methods that is specified as filter and wrapper. Filter model affords good generality and less computation cost. Wrapper model provides high classification accuracy. The feature subset evaluation is not similar for Wrappers and filters, inappropriate features are removed in filter approaches based on its general data features. In contrast, machine learning algorithms are applied toward feature subsets and cross validation is utilized for evaluating the feature subset scores.

Mostly, gene selection methods utilized microarray data analysis and focused on filter approaches, while some of them utilized wrapper approaches. The wrapper based approach offers accurate classification results than filter approach. Utilizing classifiers in wrappers support estimating the feature subset and it also achieves better classification accuracy in terms of using tailor made feature subset. The computation is difficult while combining with sophisticated algorithms like support vector machines. This is one of the disadvantages in wrapper approaches.

II. REALTED WORKS

Conventionally, several research has been done in gene classification to enhance the efficiency of classification. Mostly, genes are selected self-reliantly using statistical learning methods. The large statistical variance with higher dimension disturbs gene selection through lack of generalization ability. Employing deep learning techniques are difficult due to insufficient samples. Furthermore, there may be variations in certain subset of genes and classification performance among the different group of samples. Therefore, an efficient criterion is required to evaluate the quality and trustworthiness of the experimental results.

Inspired by deep learning approach, a new sparse model has been built for selecting gene efficiently. For gene selection process, gene grouping plays supreme role. The general group lasso method pioneered by Yadi Wang [1] has been applied for selecting cancer genes in groups. According to weighted gene co-expression network, a gene grouping heuristic method has been offered. The gene selection algorithm has been developed to implement a complicated group lasso calculation process. This method accomplishes superior classification performance on three cancer datasets than other state-of-the-art gene selection methods. Grouping of gene is difficult to detect the complex biological gene pathways. Though sparse group lasso identified the significant groups and genes with the selected genes, similar penalty coefficient applied to whole genes not concerned a relative importance. The sparse group lasso does not perform well if its group size is not uniform. An extension of group lasso was established [2] using adaptive sparse group lasso related to conditional mutual information.

Manuscript published on 30 September 2019
* Correspondence Author
A.Immaculate Mercy* Research scholar ,Department of computer science, A.V.V.M Sri Pushpam College, Poondi, Thanjavur, India. Mail id:mercyhelenphd17@outlook.com.
M.Chidambaram PG and Research ,Department of computer science, Rajah Serfoji Govt. Arts College, Thanjavur, India.

The data driven weights were constructed with the use of conditional mutual information in every single separated group. Importance of individual gene, correlation of remaining pair wise genes existing in every group has been evaluated.

To overcome the generalized group Lasso issues in learning procedures the GroupWise Majorization Descent (GMD) algorithm was suggested in paper [3]. A group lasso corrected least squares and large margin classifiers could be resolved by introducing concrete algorithms. The publicly available software named as group lasso has been utilized to computing the group lasso which consistently performed well than other simulation models. A structured penalized logistic regression model has been introduced to ponder particular correlation structure within the data and it is shown in [4]. The classification performance can be improved through the proper selection of gene. The feature selection and learning model was simultaneously performed for gene expression data analysis using structured penalized logistic regression model. To optimize the model, a coordinate descent model has been implemented. From the simulation studies it was proved that the highly correlated features were selected in order to improve the classification performance. A comprehensive gene has been identified to enhance the classification accuracy of gene and recognizing the gene list that are augmented in various pathways [5] with significant p values. a structured penalized logistic regression model was introduced to consider particular correlation structure within the data and it is shown in [4]. The classification performance can be improved through the proper selection of gene. The feature selection and learning model was simultaneously performed for gene expression data analysis using structured penalized logistic regression model. To optimize the model, a coordinate descent model has been implemented. From the simulation studies it was proved that the highly correlated features were selected in order to improve the classification performance. [6] established a new flexible neural network in which the transformation of multiple classification problems were done for every single forest. In order to deepen the flexible neural model, a cascade DFNT model has been explored. The objective of this method was to combine the fisher ratio and neighborhood rough set for dimensionality reduction to achieve better classification accuracy. This gene selection method achieved higher accuracy on RNA-seq gene expression data. Several existing research about machine learning scheme in gene classification [7-11] has been surveyed deeply to examine the best machine learning approach with higher classification accuracy. In [12], author inspected towards the comparison of experiments with LibSVMs, BaggingC4.5, C4.5, AdaBoostingC4.5 and Random Forest over microarray datasets. From the experimental approach, the author analyzed that these five methods attains better data preprocessing, gene selection and discretization and two statistical test has been utilized to comparing the accuracy of ten-fold cross validation tests on seven data sets in order to confirm the results. Author absorbed that the Wilcoxon signed rank test outperformed than other test for this purpose [13] suggested a dimensionality reduction technique for classifying the class features having high correlation in order to achieve higher accuracy. For selecting the features, k means algorithm was utilized as clustering approach. The similar characteristic features existing in single cluster were categorized to eliminate the redundancy of microarray data. A relief algorithm has been utilized for ranking the results to obtain finest score of every cluster. Next the best elements were selected as a features from each cluster for the classification process. Then the random forest algorithm was utilized. At the end, from the simulation it was demonstrated that the accuracy of proposed approach for these Lung Cancer, Colon and Prostate Tumor datasets outperformed than random forest approach without clustering. [14] examined a diagnosis of cancer with gene classification approach utilizing Improved-Binary Particle Swarm Optimization (IBPSO). This framework was examined with eleven different cancer microarray datasets and it achieved higher accuracy in terms of classifying samples. Moreover, fewer relevant genes were only selected which helps to diagnose the cancer earlier. The characteristic local optimum problem existing in traditional approach was also resolved with this proposed improved-BPSO. [15] proposed a novel RF classification approach with feature value searching and subspace feature sampling method. Random Forest (RF) algorithm to deal with high dimensional data for classification using subspace feature sampling method and feature value searching. In this framework, lower prediction error was only concerned utilizing novel subspace sampling method that sustains the randomness and diversity of forest. During the construction of decision trees, a greedy technique was utilized for handling the features holding cardinal category to splitting the nodes effectively. The computational time taken for constructing the RF model was reduced. Thus the proposed RF model achieves lesser rate of prediction error than traditional methods. [16] proposed a novel technique on the basis of important variable importance measures such as Gini and permutation measures for ranking the candidate predictors. There was no natural cut off for distinguishing important and non-important variables which was one of the main disadvantage. To address these problems several testing approaches were developed. One of the existing testing approach was permutation based where the repeated computation of forest was required. The computation time of permutation based approach was only efficient for lower dimensional settings not in higher dimensional settings. In order to overcome this dimensionality issue, a novel computational effective variable importance test has been implemented in which several variables are not carried over with information. Therefore the newly implemented modified version of permutation testing approach outperformed with the use of cross-validation procedures in it. [17] proposed a new approach for selecting the best gene subset through removing irrelevant data to achieve better performance in classification process. A unique high dimensional microarray data was projected as lower dimensional subspace with the use of projection matrix. A local manifold structure of unique data was temporarily preserved with the regularization term of Laplacian graph. To solve the issues, an iterative update algorithm has been developed as well. This method was superior to other state of methods concerning microarray data classification. [18] proposed a cuckoo search algorithm for classifying the binary datasets effectively. ELM is a learning algorithm which was used to train the single layer feed forward neural networks in classification field.

In order to achieve more stable model, cuckoo search was utilized to pre train the ELM model through picking the hidden biases and input weights. Here ICSELM was utilized for logically determining the output.

The effectiveness of proposed method was analyzed using several datasets and it proved as one of the efficient method comparing other methods. [19] presented classification of breast cancer utilizing a deep learning model. Breast cancer was classified with the use of class structure deep convolutional neural network. The workload and accuracy was improved by this method. The aforementioned survey investigated several machine learning algorithm for gene classification concerning accuracy, computation time etc. From the investigations it was found that the random forest classifier achieves better performance due to its decision tree approach. After discussing the gene selection and classification techniques, there is a need to focus on identifying the genes on chromosomes. Some of the research has been reviewed to identify the gene on chromosomes [20]. The author focused mainly on copy number variants over chromosome 21 [21].

An existing approach was surveyed above and found that there were several drawbacks in achieving classification accuracy using several gene classification approaches. When comparing other methods random forest achieved better performance of predicting the accuracy. The author inspired with random forest approach and decided to take a challenge of attaining more accuracy in microarray data classification.

III. PROPOSED WORK

The proposed method shows the gene classification approach using effective random forest bootstrap algorithm. In this approach, initially pre-processing is done for removing the irrelevant and redundancy data. After pre-processing, a pure data is acquired. The pure data is trained for selecting the best features with random subset selection. Then the feature selection process is carried out with random forest classifier until the tree growth condition gets satisfied. While growing trees for decision making, the data is split up for best prediction. In this stage there is no predictive ability if the importance score of predictive variable is negative or zero. Even the predictor variable with positive importance score have the difficulty to identify larger importance scores thereby it is improbable to say that this happened accidently. The variable importance is disturbed depending on different factors such as signal-to-noise ratio or the total amount of variables correlations among the data as well as forest specific factors and amount of randomly drawn candidate predictor variables for every split. Thus to regulate the high importance score there is no availability of common threshold value. At last the error is estimated in order to predict the abnormalities in human gene. Again the process is repeated from pure data until obtaining required number of trees. The bootstrap samples and randomized tree learn are the functions utilized in RF classifier for acquiring efficient learning time without overhead. In our research work the gene abnormalities on chromosome 17 are recognized with higher accuracy. If there is deletion or duplication of gene on selected chromosome 17, there may be chance of occurring the disorders namely Micro deletion Syndrome, Alexander disease, Andersen-Tawil syndrome, Aneurysmal bone cyst, Birt-Hogg-Dubé syndrome, Bladder cancer, Breast cancer, Hereditary neuropathy with liability to pressure palsies etc.

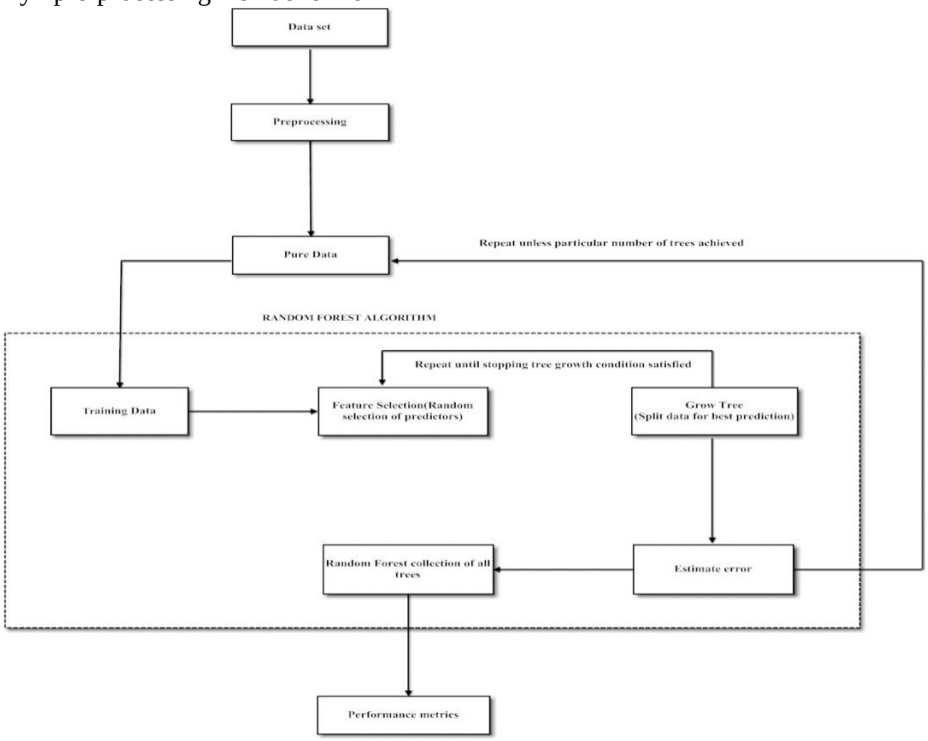

Fig. 1. Proposed flow diagram for gene classification by random forest algorithm

The proposed flow diagram is shown in Figure 1. Pre-processing is the initial step in proposed system. The imbalance in redundancy is removed by data pre-processing which removes unwanted noisy data. High dimensional, multisource data was reduced in RF algorithm. Benefits of Random forest algorithm are specified as managing nonlinear correlated data and robustness.

Random forest is group of classification and regression trees. It trains same size dataset known as Bootstraps. Random Forest adopts particular rules in combining trees, tree growth, testing, post-processing. Feature Selection for random forest is evaluated by Gini index.

Bootstrap samples are generated from training samples. Classification and regression trees are grown based on bootstrap samples. Finally new data is predicted by major values. Parameters used for random forest are

- Number of trees
- Descriptors for partitioning every node
- Features used for node partitioning.

Feature selection can be done in two methods named as filter and wrapper. Filter model provides good generality, less computation cost. Wrapper model provides high classification accuracy. Feature selection has been performed during the construction of classification rule. Last stage out of bag error estimation which is used for analyzing classification accuracy. Each and every attributes used in RF algorithm is described in detail as below.

A. Data Pre-processing:

Data preprocessing is the method where the irrelevant data are removed effectively from the dataset. Data cleaning is also performed to remove the noise. Missing of data also creates a big issue, so handling of missing data is considered for the effective preprocessing. Normalization is also done in the preprocessing stage to improve the quality of the data. The main intention of this preprocessing is to pick the minimum possible set of genes that achieves decent performance.

B. GINI Index

For each attribute, a binary split is done for the measurement of Gini index. An impurity of data i.e. either partition of data or group of training tuples is measured.

$$\text{Gini (n)} = 1-\sum_{j=1}^{n}(pj)^2 \quad (1)$$

Pj denotes the probability that a tuple in D belongs to Class Cj and is estimated by$|C_{j,D}|$ / |D|. For each acquired partition results the weighted sum of impurity is computed. For instance, in case of binary split of A partitioned D as D1 and D2, the D's Gini index partitioning is

$$\text{GiniA (D)} = \text{Gini (D1)} + \text{Gini (D2)} \quad (2)$$

Every single binary split is likely considered for each attribute. The splitting attribute for discrete valued attribute is considered from the subset which provides minimum Gini index i.e. an attribute having least Gini index are designated as splitting attribute. An attribute selection of Gini index is effective compared to other approaches. The best attribute selection can be easily done for constructing decision tree using Gini index.

C. Feature Selection

Feature selection technique used for selecting the optimal feature subset that produce better pattern characterization of several classes through eliminating inappropriate and redundant features. The feature selection strategy is categorized into wrapper and filter methods. Filter method utilized independent measure for estimating the features that are correlated whereas wrapper utilized particular learning algorithms to evaluate the feature values. The model preferred here is simple to understand. The variance of model is reduced and the computational cost of training model is also reduced. The best relevant features are only recognized is termed as feature selection.

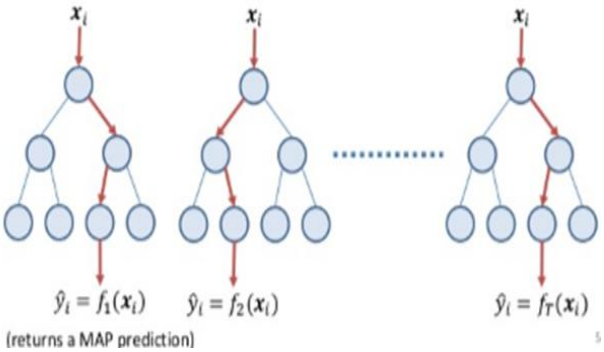

Fig. 2. Random feature selection - decision trees [22]

Random forest is typically utilized for feature selection technique due to its ranking that concerned with the enhancement of purity of nodes with the tree based strategies. This shows that there is reduction in impurity across every single node. The highest decrease in impurity of nodes occurs when trees gets started and at the finishing stage of trees a least decrease in node impurity occurs. Therefore through cutting the trees under specific node, a subset of the most significant features are created. The random feature selection decision trees shown in Figure 2.

D. Efficient Random Forest Bootstrap technique:

A group of randomized trees are constructed in random forest algorithm for the purpose of classification. The classification and regression tree algorithm is the typical one amongst the decision tree algorithms. The classification tree (CART) comprised of root node, leaf nodes, splitting nodes and edges. The optimal variable selection of nodes and the splitting of those nodes to making the pure child node are the significant facts that has to be considered while constructing a CART. An impurity of node is measured utilizing Gini index of classification algorithm. Assume node s and the assessed class probabilities, q(c/s) (C=1…C). The Gini index of node s is defined as:

$$\text{N(s)} = \sum_{c1\neq c2} q\left(\frac{c1}{s}\right) q\left(\frac{c2}{s}\right) = 1 - \sum_{c=1}^{c} q^2\,\text{(c/s)} \quad (3)$$

Allow 'o' be the splitting point of node s, where the node gets separated into two portions. In that proportion qR, t samples are allocated to sR and the sl acquired from the ql proportion i.e. sR+ sl = 1. Therefore, the decrease in Gini index is shown below.

$$\Delta N(o,t) = N(s) - q_R\, N\, s_R - q_l\, \text{N}\, s_l \quad (4)$$

The optimal splitting point s^* and optimal variable j^* yields highest decrease in Gini impurity gotten and it is shown below,

$$o^*, j^* = \arg max_{o,j} \Delta \text{N (o, s)} \quad (5)$$

The aforementioned function is called repeatedly via classification algorithm in order to generate a tree. A random decision forest is collaborative model in accordance with classification tree algorithm combining bagging and the random subspace theory. A group of decision trees are trained with the classification tree (CART) algorithm utilizing bootstrap samples and the subspaces of variables that are designated arbitrarily into every single non-leaf node. The trees existing in forest are

completely grown limitless till acquiring pure leaf nodes.

We split the RF algorithms into three functions named as bootstrap sample, Tree Generation and Randomized Tree Learn. An efficient random forest bootstrap algorithm is designated for achieving effective gene classification with high classification accuracy and less computation time. In this approach, the bootstrap samples are extracted for enhancing the performance of the model through decreasing the variance of model. The main advantage of this work is that the use of both bootstrap sample selection and the randomized tree learning reduces more number of irrelevant and redundant attributes effectively which also reduces the learning time of classifier without overhead.

Algorithm 1: Efficient Random Forest Bootstrap Algorithm

Precondition: A training set T: = (x_1,y_1). . . (x_n,y_n), features F, and number of trees in forest K.
Function Random Forest (T, F)
P ← ∅
For i ∈ 1. . . K do
$T^{(i)}$ ← A bootstrap sample from T
$T^{(i)}$← Tree Generation ($T^{(i)}$, F)
p_i ← RandomizedTreeLearn ($T^{(i)}$, F)
P ← P ∪ {p_i }
End for
Return H
End function
Function Tree Generation (T, F)
Create a root node for the T
If sample have the same target attribute value ← T, then
Return the single node tree
Else
Select attribute from best classify T based on an entropy-based measure
Set A the attribute for Root
For each legal value of attribute do
Add a branch below Root, corresponding to attribute

Let current T will be the subset of T that have attribute
If T ←null then
Add leaf node below the branch with target value← common value of T
Else
Below the branch, add a sub tree
Function RandomizedTreeLearn (T, F)
At each node:
F ← very small subset of F
Split on best feature in f
Return the learned tree
End function

$$R_r^{cc}(x,z) = \sum_{r1\ldots,,,rd,\sum k_j} \frac{r!}{r_1!\ldots\ldots,R_d}\left(\frac{1}{d}\right)^r \prod_{j=1}^{d} 1\left[2^r j x_j\right] = \left[2^r j z_j\right]$$
$$forallX\ , Z\boldsymbol{\varepsilon}[\mathbf{0},\mathbf{1}]^{\boldsymbol{d}}$$

The algorithm 1 shows the process of efficient random forest bootstrap method. The required parameters are given as T: = (x_1, y_1). . . (x_n, y_n), features F and number of trees in forest K. Initially bootstrap sample is taken from T and the tree generation is done with randomized tree learning specified as p_i. At each node, the best features are split and shown in the algorithm. Thus the abnormalities of human gene are predicted using random forest classifier. This random forest algorithm works under three functions namely a bootstrap sample from T, T^ ((i)) ← Tree Generation (T^ ((i)), F), p_i ← RandomizedTreeLearn.

IV. PERFORMANCE ANALYSIS

In performance metrics section, the classification accuracy of proposed method are deliberated with experimental results. The abnormalities in human gene are found with the parameters and the several methods are compared to prove that this method achieves better performance through higher classification accuracy.

Dataset description:

Homo sapiens dataset is utilized to predict the abnormalities of human gene such as Eukaryota, Metazoa, Chordata, Craniata, Vertebrata, Euteleostomi, Mammalia, Eutheria, Euarchontoglires, Primates, Haplorrhini, Catarrhini and Hominidae. Homo is the cell name that is used to find the abnormalities in DNA. The DNA sequence comprises genomic sequence, mainly completed clones that were sequenced as a part of the Human Genome project. PCR products and WGS shotgun sequence is added which is essential to fill gaps or correct errors.

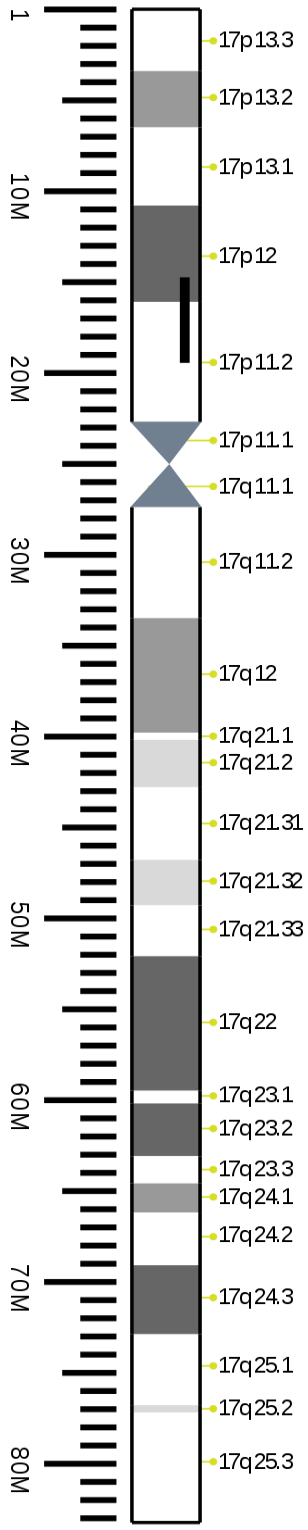

Fig. 3. Chromosome 17 – Gene [23]

The chromosome 17 contains 83 million DNA building blocks (base pairs) and signifies among 2.5 and 3 percent of the total DNA in cells. In genetic research, gene identification in every single chromosome is a dynamic research area. Several approaches have been introduced by the researchers for predicting the gene count on every chromosome, perhaps estimated gene number differs. Chromosome 17 comprised of 1,100 to 1,200 genes where the instructions can be learnt for generating protein which realizes different parts of body.

Gene classification is done with Random forest algorithm and its results are compared with other algorithm. Its performance was evaluated around 30 metrics such as accuracy, precision, recall, sensitivity, specificity. True Positive (TP), False Positive (FP), False Negative (FN), True Negative (TN), Positive (P), Negative (N), Accuracy, Error Rate, Balanced Accuracy, Specificity or True Negative Rate, Sensitivity or True Negative Rate, Sensitivity/ True Positive Rate/ Recall/ Detection Rate, Precision/ Positive Predict Value, Negative Predict Value, F1-Score, G Mean, False

Alarm Rate/ Fall out, False Negative rate/ Miss rate, Combined metric, False Detective Rate, Receiver Operating characteristics, Informedness, Markedness, Positive Likelihood Ratio, Negative Likelihood Ratio, diagnostic Odd Ratio, Jaccard Co-Efficient, Dice co Efficient, Matthews's Correlation Coefficient and Kappa Statistics. Confusion matrix is represented in Table 1.

Table 1. Confusion matrix

TP	FN
FP	TN

Binary classifying test performance was analyzed by sensitivity, specificity.

True Positive (TP):

$$TPR = \frac{\sum TP}{\sum (TP + FN)} \quad (6)$$

False Positive (FP):

$$FPR = \frac{\sum FN}{\sum (FP + TN)} \quad (7)$$

False Negative (FN):

$$FNR = \frac{\sum FN}{\sum (FN + TP)} \quad (8)$$

True Negative (TN):

$$TNR = \frac{\sum TN}{\sum (TN + FP)} \quad (9)$$

The True Positive (TP), False Positive (FP), False Negative (FN), True Negative (TN) classes of the proposed system were comparatively analyzed and observed a lot of true positive results which proves the efficiency of the projected system.

Positive and negative:

The positive value is obtained by,

$$\text{Positive P} = \text{TP} + \text{FN} \quad (10)$$

The negative value is obtained by,

$$\text{Negative N} = \text{FP} + \text{TN} \quad (11)$$

When comparing the positive and negative values of the proposed system large number of positive value have been observed.

Balanced Accuracy:

Balanced accuracy is measured as the average accuracy of each class individually.

$$\text{Balanced Accuracy} = (\text{TP} / \text{P}) + (\text{TN} / \text{N})) / 2 \quad (12)$$

The proposed system possesses high value of accuracy than the balanced accuracy.

ROC Curve:

The receiver operating characteristics is plotted against a false alarm rate and error rate. The false alarm rate and the Error rate is predicted by the given formula.

$$\text{False Alarm Rate} = \text{FP} / \text{N} \quad (13)$$

$$\text{Error Rate} = (\text{FP} + \text{FN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN}) \quad (14)$$

G-mean:

G-mean is also an average type, where it is used for calculating growth rates like growth of a population. This G-mean or Geometric mean will multiplies the given items. It is done only for positive values. It can be calculated as below,

$$G - mean = (\prod_{i=I}^{N} x_i)^{I/N} = \sqrt[n]{a_1 a_2 \dots a_n}$$

Where $G - mean$ is the product of nth root of numbers ranging from $a_1 \dots a_n$.

Informedness:

Informedness measures predictor's provided information for the definite condition, and postulates the possibility that a calculation is informed with respect to condition.

$$\text{Informedness} = \text{sen\_tpr\_rec\_dr} + \text{spec\_tnr} - 1 \quad (15)$$

Where sen_tpr_rec_dr and spec_tnr represents sensitivity and specificity of predictor values (true positive and true negative values).

Markedness:

Markedness measures predictor's provided marked for the definite condition, and postulates the possibility that a calculation is marked with respect to condition.

$$\text{Markedness} = \text{prec\_ppv} + \text{npv} - 1 \quad (16)$$

Where prec_ppv represents precision of positive value.

Jaccard Co-efficient:

The Jaccard coefficient measures resemblance in between finite sample sets, and is well-defined as the intersection size divided by the union size of the sample sets.

$$\text{Jaccard Co Efficient} = \text{tp} / (\text{tp} + \text{fp} + \text{fn}) \quad (17)$$

Dice Co-efficient:

The straight comparison of continuous measures with the ground truth segmentation is achieved by dice coefficient

$$\text{Dice co Efficient} = 2 * \text{tp} / (\text{fp} + 2 * \text{tp} + \text{fn}) \quad (18)$$

Positive and negative predict value:

These positive and negative predictor values are the measures of positive and negative results respectively having true positive and true negative results both diagnostically and statistically.

$$PPV = \frac{\sum TP}{\sum (TP + FP)}$$

$$\text{NPV} = TP / (TP + FP) \quad (19)$$

Kappa statistics:

The Kappa coefficient is a measure of inter-rater agreement that is utilized to measure qualitative documents and regulate agreement in between two raters.

$$\text{Kappa Statistics} = (2 * ((\text{tp} * \text{tn}) - (\text{fp} * \text{fn})))/(((\text{tp} + \text{fp}) * (\text{fp} + \text{tn})) + ((\text{tp} + \text{fn}) * (\text{fn} + \text{tn}))) \quad (20)$$

F1-Score:

It is defined as the harmonic average of the recall and precision.

$$F1\,Measure = 2 * \frac{precision.recall}{precision + recall} \quad (21)$$

Accuracy:

Classifier accuracy is probability of unlabeled instance class prediction.

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP} * 100\% \quad (22)$$

Sensitivity:

Sensitivity is referred as the true positive value to total positive occurrences. Sensitivity is also called as True positive rate. Formula for Sensitivity was given below:

$$\text{Sensitivity} = \frac{True\ Positive}{\text{True Positive+False Negative}} \quad (23)$$

Specificity:

Specificity is referred to as the true positive divided by total occurrences that can be positive. Specificity is also Known as True negative rate. Specificity was given by the following formula:

$$\text{Specificity} = \frac{True\ Positive}{\text{True Positive+False Positive}} \quad (24)$$

Precision:

Precision is calculated using the given formula:

$$\text{Precision} = \frac{|\{Relevant\ documents\} \wedge \{Retrived\ documents\}|}{|\{Retrived\ documents\}|} \quad (25)$$

Precision is the fraction of retrieved documents which relevant to the query. Other name of precision is exactness.

Recall:

Positive tuples fraction doing for classifier is positive. The perfect score of recall is 1.0. It is the document related to query divided by facts retrieved. Other name of recall is completeness. Precision and recall are inversely related. Recall is measured using the given formula:

$$\text{Recall} = \frac{|\{Relevent\ documents\} \wedge \{Retrived\ documents\}|}{|\{Relavant\ documents\}|} \quad (26)$$

The accuracy of classifier is also evaluated based on the precision and recall values.

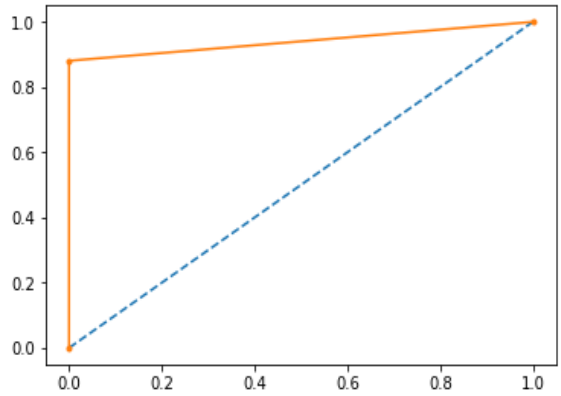

Fig. 4. ROC curve for proposed true positive and false positive

MCC metrics:

Mathew correlation coefficient is the measure which is determined with True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). Concerning TP, γ, π revision can be made.

$$\text{MCC}(\theta) = \text{TP TN} - \text{FP FN} / \sqrt{(TP+FP)(FP+FN)(TN+FN)(TN+FP)}$$

$$= \frac{TP-\gamma\pi}{\gamma(1-\gamma)\pi(1-\pi)} \quad (27)$$

In this case, when the small class is having label 1 than π resembles to minority class proportion. MCC is stated as

- Utilizing confusion matrix, a MCC be able to be measured
- Four quantities such as TP, TN, FP and FN are taken for the evaluation to achieve better classification performance.
- If the measures like *TP + FN*, *TP + FP*, *TN + FP*, or *TN + FN* is zero, MCC cannot be determined.
- MCC accepts [−1, 1] interval value, where 1 shows complete agreement, -1 shows complete disagreement, 0 shows no prediction those are uncorrelated with the ground truth.
-

Table 2. Elemental metric's definition used to articulate the evaluation metrics

Metric	Definition	Description	Metric	Definition	Description
TP	P(Y=1, $\theta = 1$)	True positive(appropriately identified)	FN	P(Y=1, $\theta = 1$)	False negative
TN	P(Y=0, $\theta = 0$)	True negative (acceptably disallowed)	FP	P(Y=0, $\theta = 0$)	False negative
TPR	TP/(TP +FN)	True positive rate	TNR	TN/(FP +TN)	True negative rate
Precision	TP/(TP +FP)	Positive prediction value	Recall	TP/(TP +FN)	True positive rate

Table 3. Elemental metric's definition used to articulate the evaluation metrics

Metric	Expression
MCC	TP TN – FP FN / $\sqrt{(TP+FP)(FP+FN)(TN+FN)(TN+FP)}$
AUC	Area under ROC curve
Accuracy	TPR+TNR/2
F1 measure	2/(1/precision + 1/recall)

Performance comparison with other methods:

Accuracy for gene classification by random forest algorithms are compared to other algorithms and listed in Table 4 which shows the accuracy of gene classification in various algorithms. The techniques such as GL, SGL, KNN, Linear SVM and MLP are compared with each other to analyze the effectiveness of proposed method. The GL attains 81.2% of classification accuracy, SGL acquires 82.1%, KNN yields 81.6% and MLP achieves 87%. These methods acquires lesser classification accuracy than proposed method which yield 97.6%. The existing linear SVM method attains 97% which is better than other existing approach. The proposed method is superior in classification accuracy when compared to other methods.

Table 4. Accuracy of gene classification in various algorithms

Gene Classification	
Proposed(Random forest)	97.6
GL	81.2
SGL	82.1
KNN	81.6
Linear-SVM	97.0
MLP	87.0

Figure 5 portrays gene classification accuracy of various algorithms. In the graph, we depict that the proposed random classifier achieves 97.6% which is 0.6% higher than the linear SVM. The other approaches attains only 80% which is much lower than the proposed work. Thus, the proposed method is superior to other state of art methods.

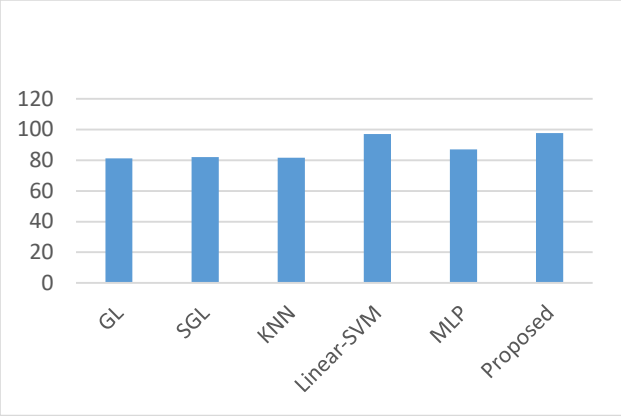

Fig.5. Gene classification accuracy comparison of various algorithms

Table 5 shows the Wrapper approach based on effective random forest bootstrap classifier. The significant features (cell types) selected in this approach are given as Eukaryota, Metazoa, Chordata, Craniata, Vertebrata, Euteleostomi, Mammalia, Eutheria, Euarchontoglires, Primates, Haplorrhini, Catarrhini and Hominidae. For 13 cell types, the accuracy achieves 97.61 which is much better than other approaches.

Table 5. Wrapper approach based on effective random forest bootstrap classifier

Significant selected features	Feature count	Accuracy % (ERFBC)
Eukaryota, Metazoa, Chordata, Craniata, Vertebrata, Euteleostomi, Mammalia, Eutheria, Euarchontoglires, Primates, Haplorrhini, Catarrhini and Hominidae	13	97.61

Table 6 shows the comparison of wrapper based approach. The ERFBC attains more accuracy with 13 features than the genetic algorithm based classification.

Table 6. Comparison of wrapper based approach

Techniques	Significant selected features	Amount of features	Accuracy
Genetic algorithm (Wrapper based approach) [24]	N2=51,575,633, 870,1244,1310, 1742,1952	8	88.70968
Efficient random forest bootstrap classifier	Eukaryota, Metazoa, Chordata, Craniata, Vertebrata, Euteleostomi, Mammalia, Eutheria, Euarchontoglires, Primates, Haplorrhini, Catarrhini and Hominidae	13	97.61

Thus, from the experimental results, it is proved that ERFBC outperforms than genetic algorithm. Homosapiens dataset chromosome 17 to find the abnormalities of the gene in the cell types such as Eukaryota, Metazoa, Chordata, Craniata, Vertebrata, Euteleostomi, Mammalia, Eutheria, Euarchontoglires, Primates, Haplorrhini, Catarrhini and Hominidae. The unique chromosome is considered to achieve better accuracy in identifying the disorder.

V. CONCLUSION

The abnormalities in human chromosome-17 are explicitly speculated by using effective random forest bootstrap technique. The unwanted data from homo-sapiens dataset are eliminated by preprocessing. The suitable and related features are then chosen from the improved dataset and ranked based on their priorities. The decision trees are framed on the basis of random bootstrap samples. Finally, the classification of abnormalities in chromosome-17 is effectuated by the proposed method. The accuracy value of proposed technique is contrasted with that of all the other prevailing algorithms. It is observed that the effective random forest bootstrap system performs gene classification in less time with high accuracy.

REFERENCES

1. Wang, Y., X. Li, and R. Ruiz, *Weighted general group lasso for gene selection in cancer classification.* IEEE transactions on cybernetics, 2018. **49**(8): p. 2860-2873.
2. Li, J., W. Dong, and D. Meng, *Grouped gene selection of cancer via adaptive sparse group lasso based on conditional mutual information.* IEEE/ACM transactions on computational biology and bioinformatics, 2017.
3. Yang, Y. and H. Zou, *A fast unified algorithm for solving group-lasso penalize learning problems.* Statistics and Computing, 2015. **25**(6): p. 1129-1141.
4. Liu, C. and H. San Wong, *Structured Penalized Logistic Regression for Gene Selection in Gene Expression Data Analysis.* IEEE/ACM transactions on computational biology and bioinformatics, 2017.
5. Wu, H.-C., X.-G. Wei, and S.-C. Chan, *Novel Consensus Gene Selection Criteria for Distributed GPU Partial Least Squares-based Gene Microarray Analysis in Diffused Large B cell Lymphoma (DLBCL) and related findings.* IEEE/ACM transactions on computational biology and bioinformatics, 2017.
6. Xu, J., et al., *A Novel Deep Flexible Neural Forest Model for Classification of Cancer Subtypes Based on Gene Expression Data.* IEEE Access, 2019. **7**: p. 22086-22095.
7. Chen, K.-S., et al., *Homologous recombination of a flanking repeat gene cluster is a mechanism for a common contiguous gene deletion syndrome.* Nature genetics, 1997. **17**(2): p. 154.
8. Nguyen, T., et al., *Medical data classification using interval type-2 fuzzy logic system and wavelets.* Applied Soft Computing, 2015. **30**: p. 812-822.
9. Sahmadi, B., et al. *A modified firefly algorithm with support vector machine for medical data classification*. in *Computational Intelligence and Its Applications: 6th IFIP TC 5 International Conference, CIIA 2018, Oran, Algeria, May 8-10, 2018, Proceedings 6*. 2018. Springer.
10. Manogaran, G., et al., *Machine learning based big data processing framework for cancer diagnosis using hidden Markov model and GM clustering.* Wireless personal communications, 2018. **102**(3): p. 2099-2116.
11. Jiang, Y., et al., *Seizure classification from EEG signals using transfer learning, semi-supervised learning and TSK fuzzy system.* IEEE Transactions on Neural Systems and Rehabilitation Engineering, 2017. **25**(12): p. 2270-2284.
12. Hu, H., et al. *A comparative study of classification methods for microarray data analysis*. in *Proceedings of the fifth Australasian conference on Data mining and analystics-Volume 61*. 2006. Australian Computer Society, Inc.
13. Aydadenta, H., *A Clustering Approach for Feature Selection in Microarray Data Classification Using Random Forest.* Journal of Information Processing Systems, 2018. **14**(5).
14. Jain, I., V.K. Jain, and R. Jain, *Correlation feature selection based improved-binary particle swarm optimization for gene selection and cancer classification.* Applied Soft Computing, 2018. **62**: p. 203-215.
15. Wang, Q., et al., *An efficient random forests algorithm for high dimensional data classification.* Advances in Data Analysis and Classification, 2018. **12**(4): p. 953-972.
16. Janitza, S., E. Celik, and A.-L. Boulesteix, *A computationally fast variable importance test for random forests for high-dimensional data.* Advances in Data Analysis and Classification, 2018. **12**(4): p. 885-915.
17. Tang, C., et al., *Gene selection for microarray data classification via subspace learning and manifold regularization.* Medical & biological engineering & computing, 2018: p. 1-14.
18. Mohapatra, P., S. Chakravarty, and P.K. Dash, *An improved cuckoo search based extreme learning machine for medical data classification.* Swarm and Evolutionary Computation, 2015. **24**: p. 25-49.
19. Han, Z., et al., *Breast cancer multi-classification from histopathological images with structured deep learning model.* Scientific reports, 2017. **7**(1): p. 4172.
20. Satsangi, J., et al., *Two stage genome–wide search in inflammatory bowel disease provides evidence for susceptibility loci on chromosomes 3, 7 and 12.* Nature genetics, 1996. **14**(2): p. 199.
21. Rambo-Martin, B.L., et al., *Analysis of copy number variants on chromosome 21 in Down syndrome-associated congenital heart defects.* G3: Genes, Genomes, Genetics, 2018. **8**(1): p. 105-111.
22. Kimura, A., *Global Refinement of Random Forest.* Slideshare, 2015.
23. *Chromosome 17.* Wikipedia, 2018.
24. Pavithra, D. and B. Lakshmanan. *Feature selection and classification in gene expression cancer data*. in *2017 International Conference on Computational Intelligence in Data Science (ICCIDS)*. 2017. IEEE.

AUTHORS' PROFILE

Immaculate Mercy. A is pursuing her Ph.D. in Computer Science in Bharathidasan University. She holds a Master's Degree in Computer Applications, Masters in Computer Science Engineering and M.Phil. in Computer Science. She has over 16 years of experience in teaching at the Under Graduate and the Post Graduate level and 8 years of industry experience, which involved in software development and e-content writing. She also has a proven experience in Corporate Training. Her research work focuses on Data Science, Prediction Analysis, Data Mining, Big Data Analytics and Cloud Computing.

Chidambaram. M pursued his Master's in Computer Science from Bharathidasan University, M.Phil. from Bharathidasan University, M.B.A from Periyar University and Ph.D. from Vinayaka Mission in Computer Science. He has over 21 years of experience in teaching at the undergraduate and post graduate level. He has guided over 25 scholars towards M.Phil. degree and 8 scholars towards Ph.D. degree. He has published more than 40 research papers in various National and International journals, 8 conference papers .He is currently working as an Assistant professor in Computer science in RSGC, Thanjavur. His areas of Research are Cloud Computing, Grid Computing, Data Mining and Big Data Analytics.