International Journal of Engineering Research and Technology.
ISSN 0974-3154 Volume 11, Number 4 (2018), pp. 637-648

http://www.irphouse.com

G-HWRF: Gene Signature based Hierarchical Weighted Random Forest Clustering Technique for High Dimensional Human Disease Data Sets

N.K.Sakthivel[1], N.P.Gopalan[2], S.Subasree[3]

[1]Research Scholar, Bharath University, Chennai – 600 073, Tamil Nadu, India.

[2]Professor, Department of Computer Applications, National Institute of Technology, Tiruchirappalli, Tamil Nadu, India.

[3]Professor and Head, Department of Computer Science and Engineering, Nehru College of Engineering and Research Centre, Pampady- 680 588, Kerala, India.

ABSTRACT

Many Human Diseases are not taking place or occurring due to Gene only and instead diseases have occurred due to interact with various genomes together and this is causing diseases. Hence it is needed for analysing the entire genome sequences thoroughly for understanding the interactions that will help us for predicting various patterns of human diseases. A few Gene Signature based Clustering Techniques namely i. Hierarchical-Random Forest based Clustering (HRF-Cluster), ii. Genetic Algorithm-Gene Association Classifier (GA-GA) and iii. Weighted Common Neighbor Classifier (wCN) have been proposed recently by Researchers to predict the gene patterns of Diseases. This research work previously developed a Gene Signature based Hierarchical-Random Forest Cluster(G-HRF Cluster) and achieved better performance compared to the above mentioned Classifiers. However, from the experimental study, it is observed that G-HRF Cluster unable to maximize Classification Accuracy when very High Dimensional Data Sets used and analysed for Diseases Pattern Prediction. To address this issue, this paper proposed an efficient model called Gene Signature based Hierarchical Weighted Random Forest Clustering Technique (G-HWRF). That is the proposed model will construct weighted matrix from Hierarchical Weighted Random Forest, which will minimize Misclassification Rate. This Model was simulated and carefully studied. This research work revealed from its experimental results that the proposed Clustering Technique G-HWRF is performing well compared to that of our

previous work G-HRF Cluster with regard to Accuracy, Disease Pattern Classification/Prediction, Memory Usage, and Processing Time.

Keywords: Genetic Algorithm, Gene Association, Human Genome Prediction, Hierarchical Clustering and Weighted Random Forest.

I. INTRODUCTION

The DNA Microarrays was designed to measure the various transcript DNA and RNA stages and levels. The transcript DNA and RNA stages and levels have been derived from genome genes [1,2,3,4,5]. The prime objective of expression signature of Gene was used for predicting and identifying possible human disease patterns[1,2,5,7,8]. As group of Genes Association possibly creating and causing diseases, more researches have focusing bioinformatics for their research. This disease-gene association research is facilitating researchers to identify or predict disease genes.

Though there are several classification schemes proposed for predicting the Pattern or for studying various possible human disease patterns for improving classification accuracy, we still needed to propose efficient Models for achieving better classification accuracy and also needed effective models for predicting patterns from large datasets. This is the prime objective of this Research work. From the literature survey, it was observed that there were some models ie some diseased-Gene prediction methods were developed recently classify or predict Genes associated various diseases. To improve the classification accuracy, some methods were proposed group of Known Genes Diseases[1,5,6,7,14] that is employed for predicting various human diseases patterns by various Gene Pattern Prediction Methods. The Author Sakthivel and et. al.[1,2,3] identified recently proposed popular classifiers namely i. Genetic Algorithm-Gene Association Classifier (GA-GA)[12] ii. Hierarchical-Random Forest based Clustering (HRF-Cluster)[11], and iii. Weighted Common Neighbor Classifier (wCN)[13] and thoroughly studied. Experimental Analysis reported that these models unable to achieve fair gene pattern classification accuracy. Sakthivel and et. al. [1,2,3] developed an efficient and effective pattern classifier named Gene Signature based Hierarchical Random Forest (G-HRF)[1,2] for improving prediction/classification accuracy and it relatively performed better compared to that of HRF-Cluster[1,11], GA-GA[1,12] and wCN[1,13] Classifiers.

However, Gene Signature based Hierarchical Random Forest Clustering Technique (G-HRF) unable to perform well for very large Dimensional Data Sets used for diseases pattern classification and prediction. As this is one of the major issues, we have developed an effective model for maximizing the efficiency of the existing G-HRF Classifier. That is, this paper proposed a Weighted Random Forest Approach based Classifier called Gene Signature based Hierarchical Weighted Random Forest Clustering Technique G-HWRF. The detailed methodology of the developed Hierarchical Weighted Random Forest Clustering Technique (G-HWRF) is described here.

We have arranged and written this paper as follows. Gene Signature based HRF Cluster G-HRF is narrated in the Section 2. The proposed model, the Gene Signature based Hierarchical Weighted Random Forest Clustering Technique (G-HWRF) is narrated in Section 3. The simulated outputs of the developed model are presented in Section 4 and in the Section5, we have given Conclusion.

II. G-HRF : GENE SIGNATURE BASED HIERARCHICAL RANDOM FOREST CLUSTER

In this section, the existing classifier, called Gene Signature based Hierarchical Random Forest Cluster (G-HRF) was discussed.

This work was developed to Identify various Signatures of Genes to predict different Patterns of Genes with better accuracy[1,2,3,4,6]. The Procedure is discussed in detail below section.

A. G-HRF Procedure

The G-HRF was developed to predict various gene sets which have strong association with gene patterns and expressions. The Euclidean Distance Model is used for calculating various gene patterns' points distances and these points were considered to construct Clusters. This model is also able to combine various Clusters that are very close to its Points and Sizes.

The prime feature of this Model is used to eliminate outliers that are considered as Noises which will facilitate to reduce dimension that will maximize the Classification Accuracy and minimize misclassification as well.

The Closest Clusters that were constructed and formed by this model was further improved through Genetic Approach based Hierarchical Random Forest Model. It is noticed that, this model is achieving better accuracy with regard to prediction and classification as well.

$$\rho = 1 - \frac{6\sum d^2}{N(N^2 - 1)} \tag{1}$$

The detailed procedure, operations and architecture of the Gene Signature based Hierarchical Random Forest Cluster (G-HRF) is shown in the Figures Fig. 1(a) and Fig. 1(b).

Procedure Steps	Activities and Functions
1	*Collect Genome Sequence Training Data*
2	*Create Multiple Clusters through Euclidean Distance*
3	*Find Similar Clusters based on distance Calculated*
4	*Find Clusters with less points and merge together through Hierarchical Cluster*
5	*Validate through Hierarchical Random Forest*
6	*Minimize Misclassification Rate through GA-HRF*
7	*Maximize Area Under Curve (AUC) Measurement*
8	*Select Most Closest Cluster through GA-HRF*
9	*Remove Redundant Clusters through Spearman Rank Correlation Model*

Fig1a: The Operation Steps and Functionality of G-HRF Cluster

III. G-HWRF : GENE SIGNATURE BASED HIERARCHICAL WEIGHTED RANDOM FOREST CLUSTERING TECHNIQUE

In our previous G-HRF Model, the Gene Signatures involved for Diseases have been predicted and employed for Clinical Tests[1,2,3]. As our previous work unable to support for very High Dimensional Data Sets, this Research Work enhanced this existing model further for improving the efficiency of the developed classifier with regard to Classification/Prediction Accuracy, Execution / Processing Time, and Memory Usage for supporting High Dimensional Data Sets. The key idea behind the developed model is that it is developed to reduce subspace size which improves Prediction Accuracy significantly.

As the Hierarchical Weighted Random Forest[1,4,5,9] has the following Features, we have chosen for our work.

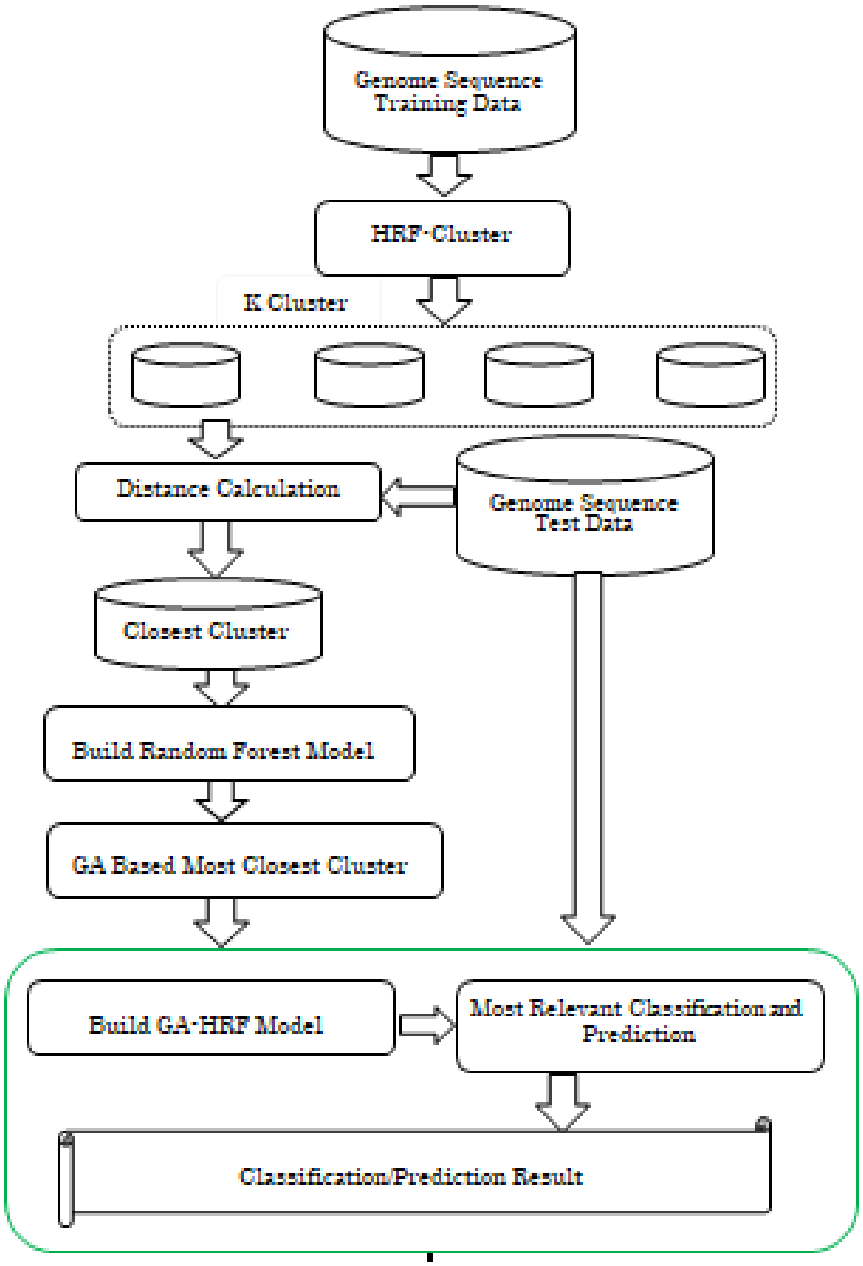

Fig 1b: Genetic Signature based Hierarchical Random Forest Cluster (G-HRF Cluster)

- Best Performance to address over-fitting
- Machine-Learning based Classification
- Cost-Effective Identification
- Supports well for Hierarchical Classification for High Dimensional Data Sets

The detailed Architecture and Operations of the Hierarchical Weighted Random Forest Clustering Technique (G-HWRF) is elaborately presented in below section.

A. G-HWRF Procedure

The proposed G-HWRF Clustering Technique is developed to predict various gene sets and patterns which have strong association with different gene patterns and expressions. The Architecture and Operations of the G-HWRF is shown in the Fig. 2.

Step 1: Forming Training Data(Genome Sequence)

Step 2: Construct Clusters based on Euclidean Distance Model

Step 3: Grouping Identical Clusters

Step 4: Merging Clusters with close associated sets through Hierarchical Clustering Approach

Step 5: Construct In_of_Bag Tree(IoB Tree)and Out_of_Bag Tree(OoB Tree)

Step 6: Combine IoB Classifiers and Validate Data Sets through Bootstrap

Step 7: Select Trees with maximum Classification Accuracy

Step 8: Construct Weighted Matrix through Hierarchical Weighted Random Forest as

$$w_{ij} = \frac{\mathrm{Acc}_{ij}}{\sum_{j=1}^{K} \mathrm{Acc}_{ij}} \qquad (2)$$

Step 9: Reduce Misclassification and Dimensionality by G-HWRF

Step 10: Measure Weighted Voting to Maximize (AUC Measurement as

$$wv_i = \sum_{j=1}^{K} \mathrm{W}_{ij} * \mathrm{v}_{ji} \qquad (3)$$

Step 11: Select Most Closest and Relevant Cluster through G-HWRF

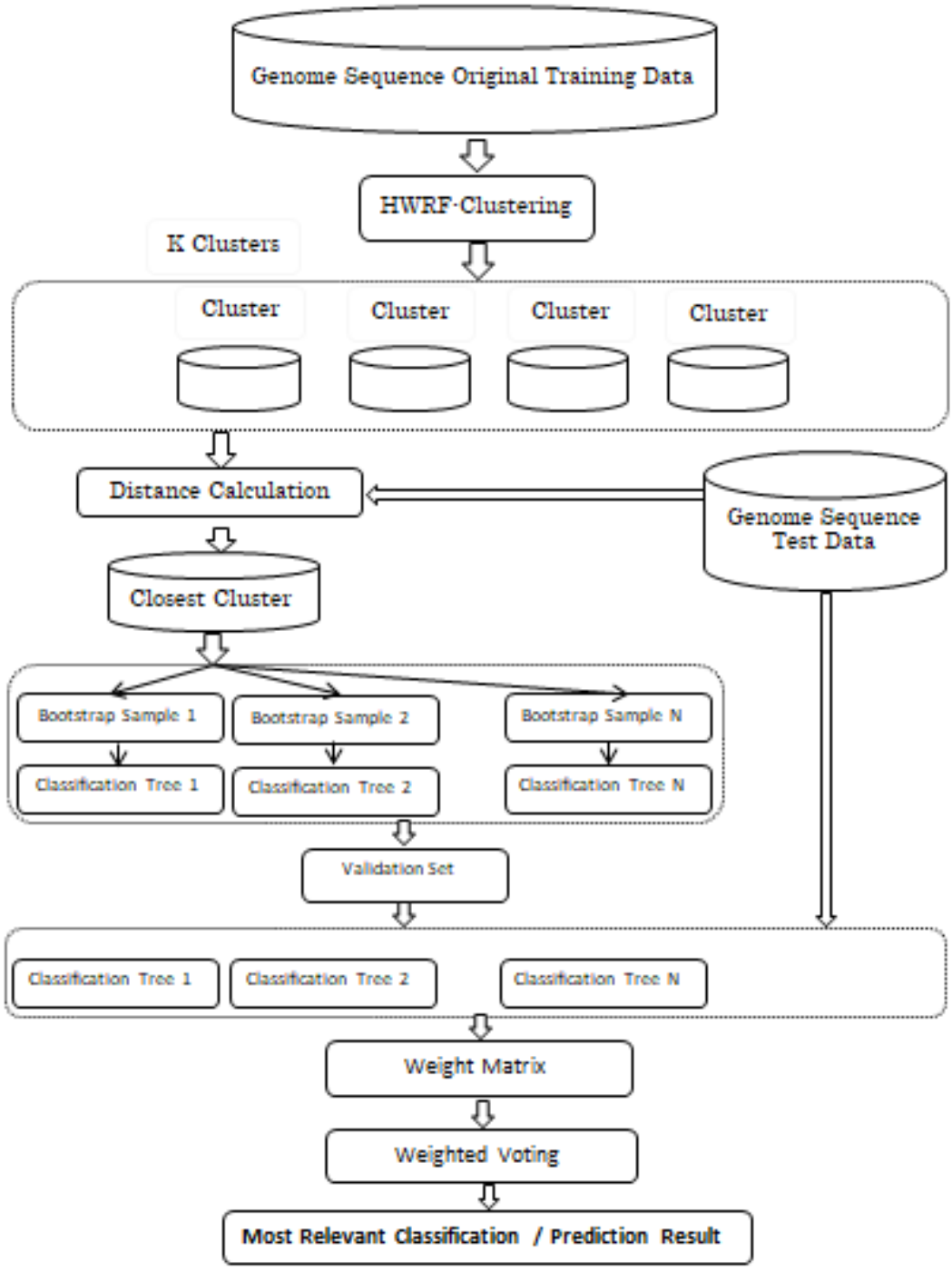

Figure 2: Genetic Signature based Hierarchical Weighted Random Forest Clustering Technique (G-HWRF)

IV. PERFORMANCE ANALYSIS

The proposed model was implemented and simulated effectively with the help of Database and Data Sets, Master.MER[1,2,3]. This Data Sets were downloaded from NCBI for thorough analysis of the proposed model.

The proposed model is simulated to conduct experimental study and analysis as well to review the performances and prediction capabilities of the developed Gene Signature based Hierarchical Weighted Random Forest Clustering Technique (G-HWRF) by comparing with our previous model, Hierarchical-Random Forest based Clustering (G-HRF).

The BioWeka Tool was used to validate data to analyse the developed Classifier with regard to Prediction Classification, Processing Time, and Memory Usage.

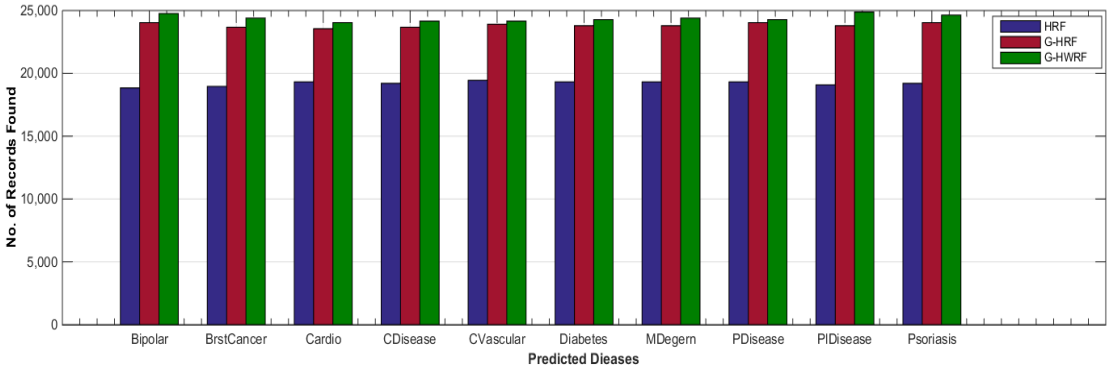

Figure 3: Performance Analysis (Pattern Prediction) of proposed G-HWRF

Table 1: Performance Analysis (Pattern Prediction) of the developed G-HWRF

Disease Pattern	**Predictors**	
	G-HRF	**G-HWRF**
Bipolar Disorder	48068	49517
Breast Cancer	47343	48792
Cardiomyopathy	47101	48068
Celiac Disease	47343	48309
Cerebral Vascular Disease	47826	48309
Diabetes	47585	48551
Macular Degeneration	47585	48792
Parkinson's Disease	48068	48551
Pericardial Disease	47585	49758
Psoriasis	48068	49275

The proposed Gene Signature based Hierarchical Weighted Random Forest Cluster Technique, G-HWRF was implemented, executed and analysed carefully with regard to Classification Accuracy, Prediction Accuracy, Processing Time and Memory Usage.

For simulating the proposed model, we considered about 10 different Data Sets to predict various patterns of human diseases. This Data Sets have 50,000 records each and totally 5,00,000 records. These Data Sets were downloaded from NCBI[13]. These records have been employed for simulations. For the purposes of records extraction and data validation, we developed Interface in VC++ and configured to BioWeka

Table 2. Performance Analysis of proposed G-HWRF

Classifiers / Parameters	Processing Time (ms)	Memory Usage (B)	Accuracy (%)
HRF-Cluster	75581	1227842	68.24
G-HRF	73546	1227842	84.86
G-HWRF	69767	1207378	92.06

The Simulation Results of the proposed model in terms of Pattern Prediction with maximum No. of Records is shown in the Fig. 3 and Table 1. From the results, it is revealed that G-HWRF, the developed Classifier, predicts diseases patterns well from more records listed as compared with that of the previous Classifier (G-HRF).

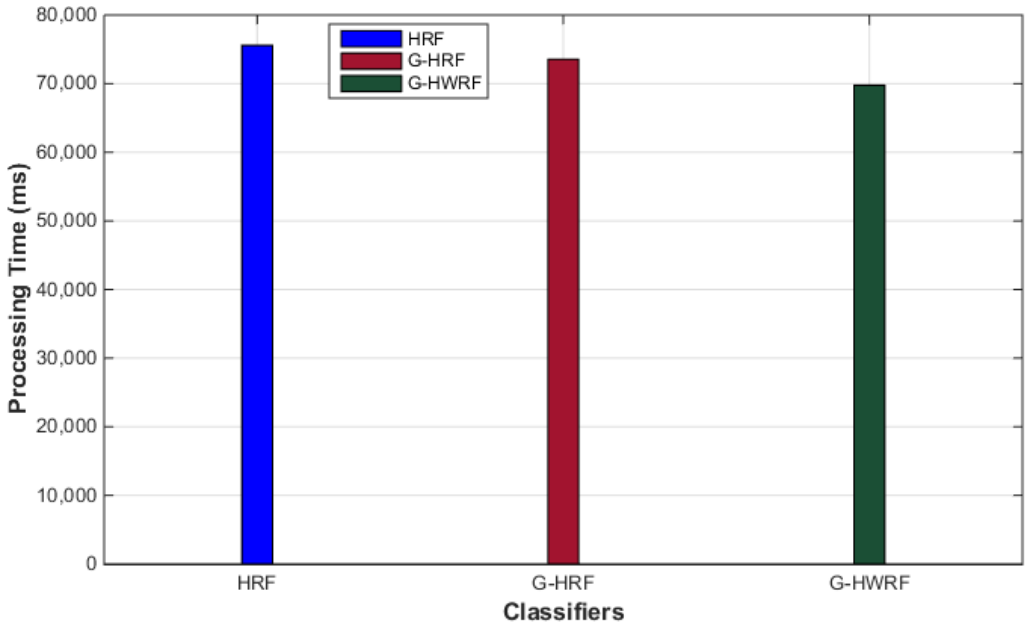

Figure 4: Processing Time (ms) vs Classifiers

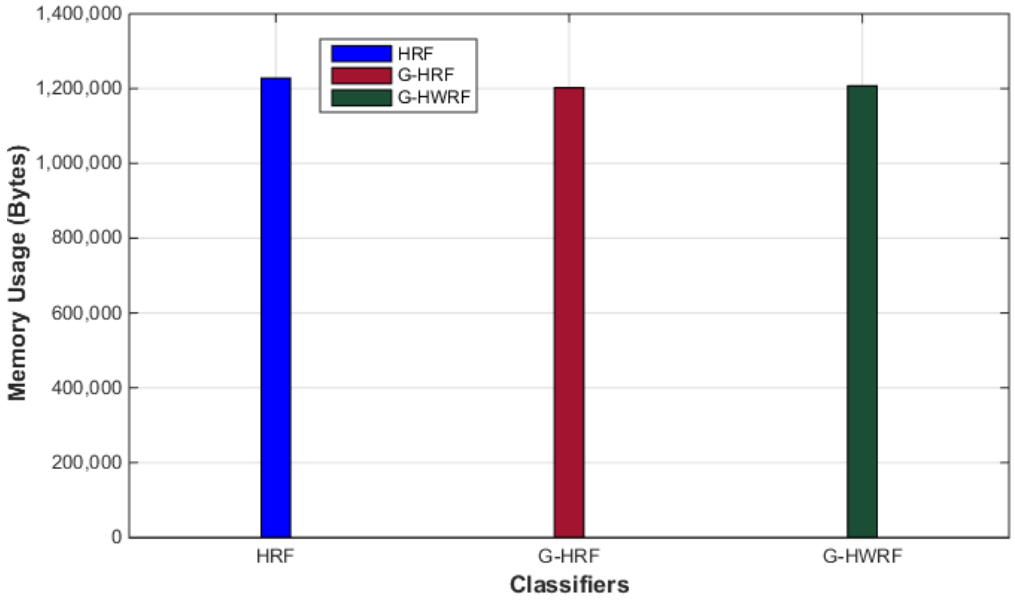

Figure 5: Memory Usage (Bytes) vs Classifiers

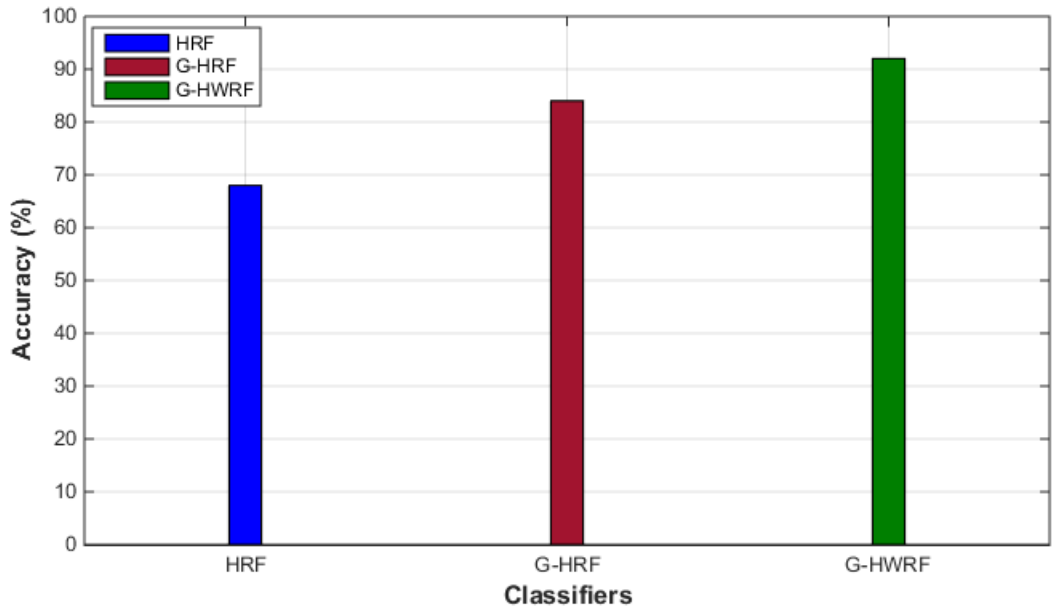

Figure 6: Pattern Prediction Accuracy vs Classifiers

From the Table 2, Figures Fig. 4, Fig. 5, and Fig.6, it was revealed that the proposed G-HWRF is performing well when the model involved for predicting pattern from Very High dimensional Data as compared with our previous work G-HRF with regard to Classification Accuracy, Processing Time and Memory Usage.

V. CONCLUSIONS

The Gene Signature based Hierarchical Weighted Random Forest Clustering Technique (G-HWRF) is simulated and studied thoroughly. For comparative analysis and thorough study, we compared this developed model with our previous version, Gene Signature based Hierarchical Random Forest Clustering Technique (G-HRF). These two models were studied carefully with regard to Classification Accuracy, Memory Usage, Memory Utilization, and Execution Time with 5,00,000 Human Genome Patterns. From the simulated outputs, it was established that the classification efficiency of the developed model, Gene Signature based Hierarchical Random Forest Clustering Technique (G-HWRF) is performing better compared to our previous work (G-HRF) Cluster with regard to Classification Accuracy, Memory Usage, Memory Utilization, and Execution Time for Very High Dimensional Data Sets.

REFERENCES

[1] N. K. Sakthivel, N. P. Gopalan, and S. Subasree, Parallel Framework based Gene Signature-Hierarchical Random Forest Cluster for Predicting Human Diseases, International Journal of Engineering & Technology, No. 7, Pp. 12-16, 2018.

[2] N. K. Sakthivel, N. P. Gopalan, and S. Subasree, G-HR : Gene Signature based HRF Cluster for Predicting Human Diseases, International Journal of Pure and Applied Mathematics, Vol. 117, No. 9, Pp. 157-161,2018.

[3] N. K. Sakthivel, N. P. Gopalan and S. Subasree, A Comparative Study and Analysis of DNA Sequence Classifiers for Predicting Human Diseases, ACM International Conference on Informatics and Analytics, ICIA-16,2016.

[4] Yue Liu, Zhiqiang Ge,Weighted random forests for fault classification in industrial processes with hierarchical clustering model selection, Journal of Process Control 64, Pp. 62–70, 2018.

[5] Baoxun Xu1, Joshua Zhexue Huang, Graham Williams and Yunming Ye1, Hybrid weighted random forests for classifying very high-dimensional data, The Computer Journal, 2015.

[6] Jonathan H. Chan, Clustering-Based Multi-Class Classification of Complex Disease, 7th IEEE International Conference on Knowledge and Smart Technology (KST2015) Pp. 25-29, Chon Buri, Thailand, 2015.

[7] Gregorio Alanis-Lobato, Exploring the Genetics Underlying Autoimmune Diseases with Network Analysis and Link Prediction, Middle East Conference on Biomedical Engineering (MECBME),2014.

[8] Wei Hu, High Accuracy Gene Signature for Chemosensitivity Prediction in Breast Cancer, Tsinghua Science And Technology, 530-536, Volume 20, Number 5, October, 2015.

[9] Conze, and et. al., Random Forests on Hierarchical Multi-Scale Supervoxels for

Liver Tumor Segmentation in Dynamic Contrast-Enhanced CT Scans, IEEE 13th International Symposium on Biomedical Imaging (ISBI), April, 2016.

[10] Desbordes Paul and et. al., Feature selection for outcome prediction in esophageal cancer using genetic algorithm and random forest classifier, Computerized Medical Imaging and Graphics,2016.

[11] Thiptanawat Phongwattana, Worrawat Engchuan and Jonathan H. Chan, "Clustering-Based Multi-Class Classification of Complex Disease," 7th IEEE International Conference on Knowledge and Smart Technology (KST2015) Pp. 25-29, Chon Buri, Thailand, 2015.

[12] Koosha Tahmasebipour and Sheridan Houghten, Disease-Gene Association Using a Genetic Algorithm. 14th IEEE Computer Society conference on Bioinformatics and Bioengineering", Pp. 191-197, 2014.

[13] Gregorio Alanis-Lobato, Exploring the Genetics Underlying Autoimmune Diseases with Network Analysis and Link Prediction, Middle East Conference on Biomedical Engineering (MECBME), 2014.

[14] http://www.biolab.si/supp/bi-cancer/projections/