Journal of Engineering and Applied Sciences 14 (21): 8072-8079, 2019
ISSN: 1816-949X

IHDGAP: Deep Learning based Intelligent Human Diseases-Gene Association Prediction Technique for High Dimensional Human Diseases Data Sets

[1]N.K. Sakthivel, [2]N.P. Gopalan and [3]S. Subasree
[1]Department of Computer Science and Engineering, Bharath University, Chennai, Tamil Nadu, India
[2]Department of Computer Applications, National Institute of Technology, Tiruchirappalli, Tamil Nadu, India
[3]Department of Computer Science and Engineering, Nehru Institute of Engineering and Technology, Coimbatore, Tamil Nadu, India

Abstract: For decades, more and more experimental researches have collectively indicated that microRNA (miRNA) could play a vital role in many important biological processes and thus, it also the pathogenesis of human complex diseases. It is also noticed that the resource and time cost requirement for processing data in traditional biological method is more expensive and thus, more and more focusing have been paid to the enhancement of efective and accurate computational mechanisms for predicting potential associations between diseases. To focus towards this, researchers identified that gene is not responsible for many human diseases and instead, diseases occur due to interaction of different group of genomes that is responsible for different diseases. Hence, it is very important to analyze and associate the complete genome sequences and its associations to understand or predict various possible human diseases. To identify and predict the associations between diseases, this research work is proposed deep learning based Intelligent Human Diseases-Gene Association Prediction technique for high dimensional human diseases data sets (IHDGAP). This gene disease sequences prediction technique is proposed through deep learning method that will predict the association between the diseases. It employs Convolution Neural Network (CNN) algorithm which contains multiple number of hidden layers which is helping to predict gene patterns and its associations to predict human diseases. The proposed model, deep learning based Intelligent Human Diseases-Gene Association Prediction technique (IHDGAP) is implemented and analyzed carefully in terms of processing time, memory usage/utilization, accuracy, sensitivity, specificity and Fscore. From the experimental results, it is noticed that the proposed deep learning mechanism improves the performances of the proposed classifier in terms of accuracy, sensitivity, specificity and Fscore as compared with our previous model gene signature based Hierarchical Random Forest (G-HRF). However, it was noticed that the proposed model consumes relatively more memory and processing time as we use Convolution Neural Network (CNN) to predict gene associations.

Key words: Gene Hierarchical based Random Forest (G-HRF), Intelligent Human Disease Gene Association Prediction (IHDGAP), deep learning, Convolution Neural Network (CNN), association mining, prediction accuracy

INTRODUCTION

DNA microarrays designed to focus for measuring the transcriptional levels of DNA and RNA transcripts. The signature of gene expression in the biomedical field used to identify a few human disease patterns (Sakthivel *et al*., 2016, 2017; Zahari *et al*., 2011; Kukreti *et al*., 2006). Associating genes with genotypes or phenotypes is demanding research topic in bioinformatics which is called as disease-gene association research. This is also called as identification or prediction of diseased genes.

Literature review: From the literature survey, it was noticed that the identification and recognition of gene diseases have been a long goal of biomedical research. It helps researchers to understand the gene function, the interactions and pathways towards improvement and contributions of medical care. There may be more number of traditional methods of gene analysis mechanisms are available but all these methods are having its own unique disadvantages Even though the association analysis mechanism work well to a set of selected functional set of genes, the selection of genes are not straight forward. Thus, we unable to apply the specialized knowledge and

Corresponding Author: N.K. Sakthivel, Department of Computer Science and Engineering, Bharath University, Chennai, Tamil Nadu, India

hence, it is considered as a limitation of this association analysis. They are many network-based algorithmic approaches have been proposed and identified for classifying gene-disease associations but most of these methods simply focus to view the objects in gene-phenotype heterogeneous networks as the same type and it does not focus the different meaning behind the gene path. To address the above identified issues, this research work focuses to identify the association between the gene associations to predict various diseases.

MATERIALS AND METHODS

Gene signature based HRF cluster (G-HR): Identifying gene signatures Alanis-Lobato *et al.* (2014), Hu (2015), Conze *et al.* (2016) and Paul *et al.* (2017) for predicting the various gene patterns with highest accuracy is most essential and that could be employed to build high accuracy gene classifier/predictor for clinical tests and applications (Sakthivel *et al.*, 2017; Phongwattana *et al.*, 2015; Tahmasebipour and Houghten, 2014; Zeng *et al.*, 2017; Fan *et al.*, 2010). Thus, an efficient gene signature based HRF cluster called G-HR was proposed. This is our previous proposed model. The procedure is elaborately discussed in the following section. The architecture of the genetic signature based hierarchical random forest is shown in Fig. 1 to achieve better pattern prediction and classification accuracy.

G-HR procedure: The procedure of the G-HR Sakthivel *et al.* (2017), Phongwattana *et al.* (2015) method is follows. This can identify gene sets that are associated with genes expression and its subset clusters. It will form clusters based on the distances of points which can calculate with Euclidean Distance Model. This model was capable of merging clusters depends on its sizes. It is capable of eliminating noises and outlies, so that, the misclassification can be reduced which will help to maximize the classification accuracy. The closest cluster built by hierarchical random forest model (Alanis-Lobato *et al.*, 2014; Hu, 2015; Conze *et al.*, 2016) was further optimized through genetic algorithm based hierarchical random forest model.

As a whole this proposed model achieves higher classification accuracy

Algorithm 1:

- Step 1: Collect genome sequence training data
- Step 2: Create multiple clusters through Euclidean distance
- Step 3: Find similar clusters based on distance calculated
- Step 4: Find clusters with less points and merge together through hierarchical cluster
- Step 5: Validate through Hierarchical Random forest
- Step 6: Minimize misclassification rate through GA-HRF

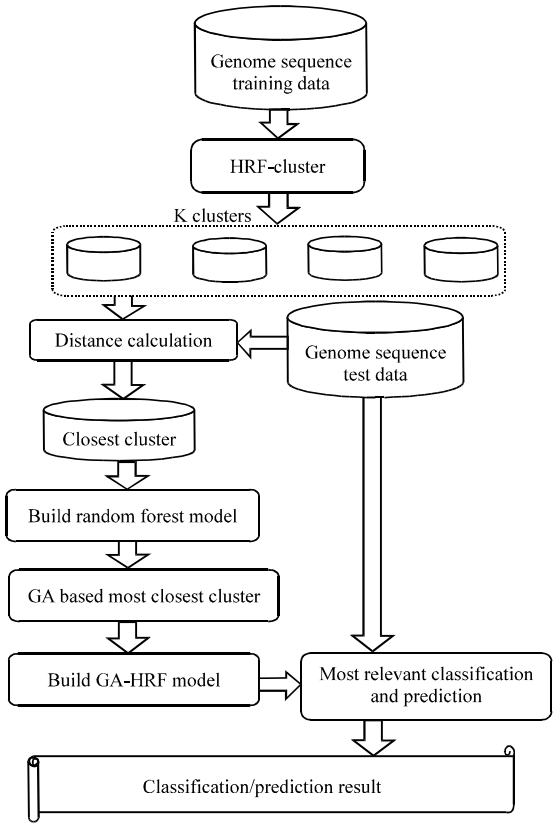

Fig. 1: Genetic signature based Hierarchical Random forest cluster (G-HR cluster)

- Step 7: Maximize Area Under Curve (AUC) measurement
- Step 8: Select most closest cluster through GA-HRF
- Step 9: Remove redundant clusters through Spearman Rank Correlation Model (Eq. 1):

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

RESULTS AND DISCUSSION

Deep learning approach and association predictions: The machine learning methods were introduced by researchers to improve classification accuracy for predicting target diseases patterns. In this study, the features of machine learning and its association predictions were discussed.

Deep learning approach: Deep learning Cheng *et al.* (2015), Chen *et al.* (2017, 2018) is a machine learning technique in which a model learns and predict the classification patterns directly from images, text or sound. Deep learning uses a neural network architecture. The term "deep" refers to the number of layers in the network if more layers, the deeper the network. Traditional neural networks contain only minimum number of layers while

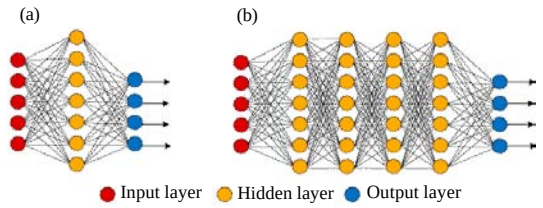

Fig. 2: Deep learning neural networks: a) Simple neural network and b) Deep learning neural network

deep networks can have hundreds of layers. Deep learning is especially, well-suited for better prediction/classification and a few applications are like classification and prediction of various diseases Chen *et al*., 2016; Li *et al*., 2017; Chen *et al*., 2016; Hoe and Rebeck, 2008; Shi *et al*., 2014), face recognition, text translation, voice recognition and advanced driver assistance systems, including, lane classification and traffic sign recognition.

As shown in Fig. 2, a deep learning neural network combines multiple nonlinear processing layers using simple elements operating in parallel and inspired by biological nervous systems (Alshalalfa and Alhajj, 2013). It consists of an input layer, several hidden layers and an output layer. The layers are interconnected via. Nodes or neurons with each hidden layer using the output of the previous layer as its input.

Identified problem: The gene signatures for predicting the various gene patterns with highest accuracy is most essential and that could be employed to build high accuracy gene classifier/predictor. This is needed for clinical tests and applications. G-HR is an efficient gene signature based clustering mechanism which is used to identify the multiple clusters to predict the accuracy of gene classification and predictions. It is capable of eliminating noises and outliers, so that, the misclassification can be reduced which will help to maximize the classification accuracy.

But however, it is predicted that if we associate group of genes that were responsible for diseases, then the diseases prediction accuracy will be better than that of G-HR. ie., our previous model unable to classify or predict gene data in better manner as we didn't group disease associated genes. To improve the classification/prediction accuracy further, it is needed to propose an efficient deep learning based gene association mechanism to predict human diseases.

To address the above mentioned issue, this research work proposed an efficient technique called, deep learning based Intelligent Human Diseases-Gene Association Prediction technique for high dimensional human diseases data sets (IHDGAP) that will classify and predict the gene patterns with better accuracy as compared with our previous model G-HR.

Intelligent Human Disease-Gene Association Prediction technique (IHDGAP): This research work is focused to improve the classification/prediction accuracy of classifier. It proposed and deployed deep learning based Convolution Neural Network (CNN) algorithm to enhance the performances of classifier in terms of processing time, memory usage/utilization, accuracy, sensitivity, specificity and Fscore.

Convolution Neural Network (CNN): A Convolutional Neural Network (CNN) is one of the most popular algorithms for deep learning. Like other neural networks, a CNN is composed of an input layer, an output layer and many hidden layers in between.

Feature detection layers are responsible for performing one of three types of operations on the data. It performs either convolution or pooling or Rectified Linear Unit (ReLU). Convolution puts the input data through a set of convolutional filters, each of which activates certain features. Pooling simplifies the output by performing nonlinear down sampling, reducing and identifying the number of parameters that the network needs to learn.

Rectified Linear Unit (ReLU) allows for faster and more effective training by mapping negative values to zero and maintaining positive values. These three operations are repeated perform over tens or hundreds of layers with each layer learning to detect different features and improve the accuracy. Figure 3 shows the architecture of CNN with its operations.

The architecture of deep learning based Convolution Neural Network (CNN) is shown in Fig. 3. It consists of six layers, one input layer, two convolutional layers and two sub-sampling layers and an output layer. As shown in the Architecture, the convolutional layers are labeled Ci and the sub-sampling layers are labeled Si, where i is the layer index. The Ci layer is obtained by performing a convolution of the previous layer data and adding a bias.

The aim of the convolution operation is to enhance the original characteristics and remove the noise information. The Si layer is obtained by calculating the average of four inputs and it will be multiplied with the average of training coefficient and these coefficient with bias and forward the result through a sigmoid function. The main focus of subsampling is to reduce the data processing without losing the useful information. Each neuron is defined as follows n (l, m, j) where, i , m and j denote the layer, map and neuron's position in the map, respectively. The value of a neuron is defined as $v^{i}_{m}(j)$ Eq. 2:

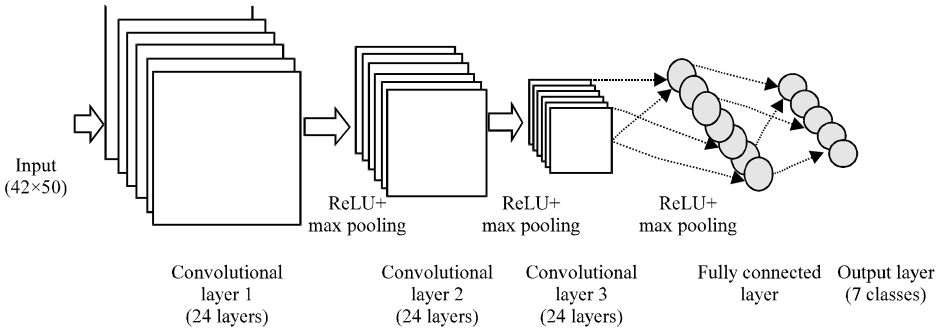

Fig. 3: Architecture of CNN

$$v_m^i(j) = f\left(x_m^l(j)\right) \tag{2}$$

where, f depends on the layer and $x_m^i(j)$, represents the scalar product between a set of input neurons in layer and the weight connections between these input neurons in the layer l-1 and the neuron number j in map m in layer l. Let define $v_m^i(j)$ first convolution layer as in Eq. 3:

$$x_m^l(f) = w(l,m,0) + \sum_{i=0}^{i<k} I_{i,j}\, w(l,m,i) \tag{3}$$

And define $x_m^i(j)$ for other convolution layers (Eq. 4):

$$x_m^l(j) = w(l, m, 0) + \sum_{i=0}^{i<k} v_m^{i-1}(j*k+i)\, w(l,m,i) \tag{4}$$

Where:

Ii,j = Stands for original input data in "Input Layer"

i = The index of each element in the kernel and the value of i is {0, 1, 2}

k = Denotes the size of the kernel and the value of k in this research is "3"

w(l, m, i) = Denotes the weight of each connection and is the weight of bias

The convolutional layer aims to find the most useful information for the classification.

Disease semantic similarity: The main aim of this method is to identify the relationship among different diseases that can be represented using Direct Acyclic Graph (DAG). Specifically an arbitrary disease D where, T(D) consisted of node D itself and all its ancestor node, E(D) is a corresponding edge set, consists of directed edges pointing to child nodes to parent nodes. The DAG can be calculated as in Eq. 5:

$$DAG(D) = (D, T(D), E(D)) \tag{5}$$

The semantic value of disease D is calculated as in Eq. 6:

$$DV(D) = \sum_{d \in T(D)} D_D(d) \tag{6}$$

$$@D_D(d) = 1 \text{ f d} = D @ D_D(d) = \max\left(\Delta, D_D(d)/d \in \text{ children of } d\right\} \text{if } d \neq D \tag{7}$$

where, Δ was the semantic contribution factor. For a given disease D, negative correction exist between D and another disease d and the contribution score of d of disease D. Disease locating in the same layer would contribute the same score to semantic value of disease D. The Semantic Similarity (SS) is calculated between disease d(i) to disease d(j) is calculated as in Eq. 8:

$$SS(d(i), d(j)) = \frac{\sum_{t \in T(d(i)) \cap T(d(i))} \left(D_{d(i)}(t) + D_{d(j)}(t)\right)}{DV(d(i)) + DV(d(j))} \tag{8}$$

Computation of scoring matrix: The main aim of identifying the semantic similarity is to identify the association between the diseases. If the association is exist are set to 1 and the elements which represent all unknown associations are set to 0. For each element a new adjacency matrix is calculated. The changed values are ranked from all samples. After getting ranks for all samples, the threshold calculation is done. If the rank is less than the threshold value, the prediction is negative otherwise the prediction is positive. For each threshold True Positive Rate (TPR-sensitivity) and False Positive Rate (FPR-specificity) can be calculated Eq. 9 and 10:

$$\text{Sensitivity} = \text{True positive}/(\text{True positive} + \text{False negative}) \tag{9}$$

$$\text{Specificity} = \text{True negative}/(\text{True negative} + \text{False positive}) \tag{10}$$

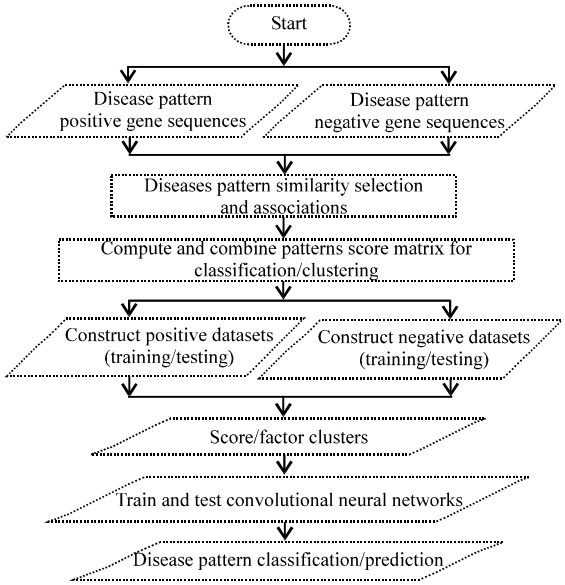

Fig. 4: Proposed architecture of Intelligent Human Diseases-Gene Association Prediction (IHDGAP)

Proposed IHDGAP architecture: The proposed architecture of the deep learning based Intelligent human diseases-gene association prediction technique (IHDGAP) is shown in the Fig. 4. The proposed model was designed to predict the various human diseases based on gene association prediction. The various steps that involved in the proposed model is clearly shown in Fig. 4.

As a whole, this proposed model achieves higher classification accuracy by using Convolution Neural Network (CNN) through deep learning mechanisms. The following steps are used to improve the accuracy of proposed model Intelligent.

Human Diseases-Gene Association Prediction (IHDGAP) Algorithm 2:

- Step 1: Collect genome sequence training data from database
- Step 2: Consider both the positive gene sequences and negative gene sequences and classify the input patterns into the mentioned two gene sequences classification
- Step 3: Identify the similarity of gene sequences and close association among gene sequences
- Step 4: Compute the pattern score matrix for the selected gene sequences
- Step 5: From the pattern score matrix, construct effective positive and negative datasets for training and testing as well
- Step 6: Calculate the new scores for the newly constructed datasets
- Step 7: Input the constructed datasets to train and test Convolution Neural Network (CNN) repeat for dataset optimization for achieving higher accuracy with association rules
- Step 8: Compare the accuracy periodically
- Step 9: Record the disease pattern classifi cations/prediction

Performance analysis: The experimental set up and simulations are carried out by this research work by using

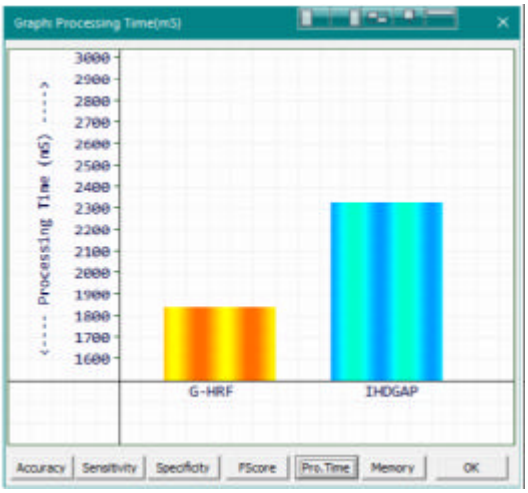

Fig. 5: Processing time vs. classifiers (proposed IHDGAP and our previous G-HRF)

Fig. 6: Memory usage vs. classifiers (proposed IHDGAP and our previous G-HRF)

the genome sequence data sets, Master.MER. This was downloaded from NCBI for thorough study. Simulations are conducted to examine the performances and classification and prediction abilities of the proposed Intelligent Human Disease-Gene Association Prediction technique (IHDGAP) and the results were compared with our previous model, gene signature based HRF cluster (G-HR) (Fig. 5 and 6).

This research considered 10 different genome genes data sets categories for predicting possible diseases and each category has 50,000 records and in total there are 500000 records used for performance analysis of the proposed model. The experiment was repeated number of times and for classifying and predicting possible

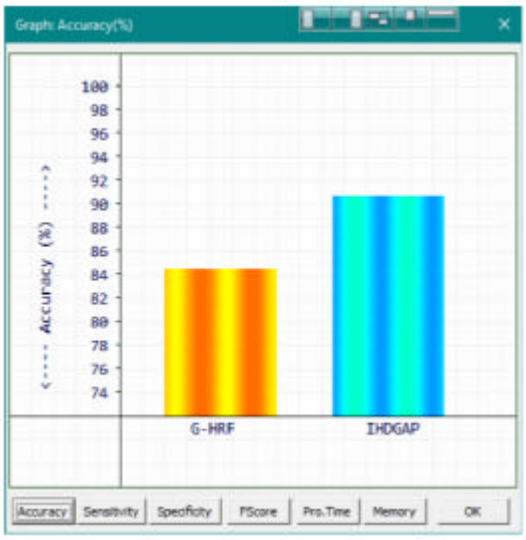

Fig. 7: Classification accuracy vs. classifiers (proposed IHDGAP and our previous G-HRF)

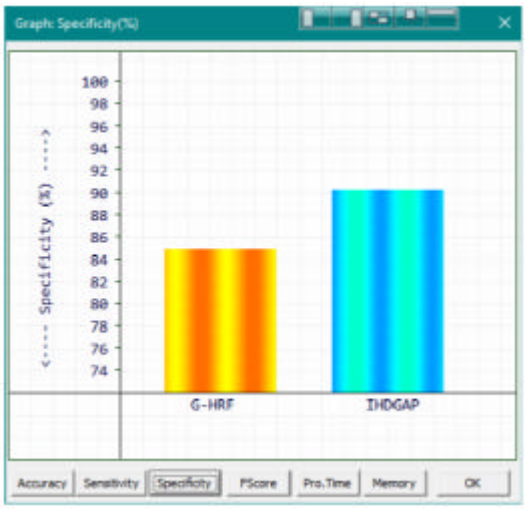

Fig. 9: Specificity vs. classifiers (Proposed IHDGAP and our previous G-HRF) between IHDGAP vs. G-HRF

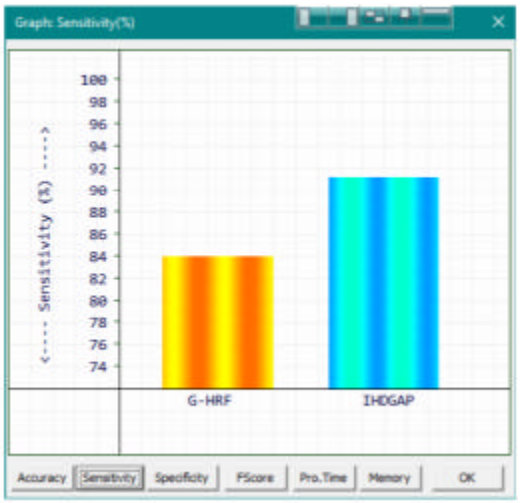

Fig. 8: Sensitivity vs. classifiers (Proposed IHDGAP and our previous G-HRF) between IHDGAP vs. G-HRF

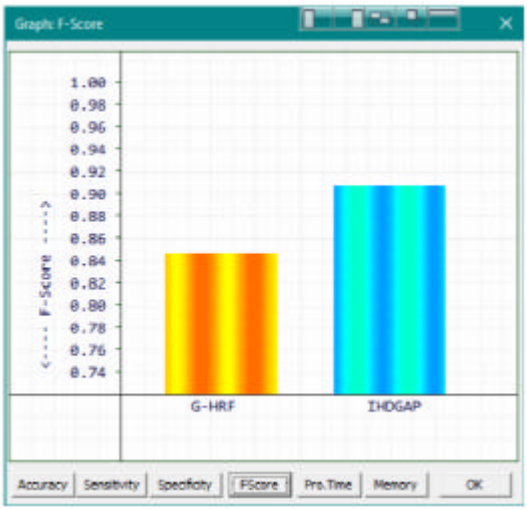

Fig. 10: Fscore vs. Classifiers (Proposed IHDGAP and our previous G-HRF) between IHDGAP vs. G-HRF

diseases were recorded. The performances of the above discussed genome classifiers have been studied in terms of in terms of processing time, memory usage/utilization, classification accuracy, sensitivity, specificity and Fscore.

This research work has developed the interfacing tool with the help of VC++ programming language with R programming for computation to extract and validate the gene expressions which are downloaded from NCBI. The validated data is fed into BioWeka for analysing the proposed genome classifiers in terms of processing time, memory usage/utilization, accuracy, sensitivity, specificity and Fscore.

The experimental results of the proposed model, Intelligent Human Diseases Prediction Technique through Gene Association Prediction mechanism (IHDGAP) is compared with our previous model called gene signature based HRF Cluster (G-HRF) in terms of processing time, memory usage/utilization, accuracy, sensitivity, specificity and Fscore and analyzed thoroughly. From the results, it was noticed that the proposed classifier is performing well which are shown in Fig. 7-10.

From the Fig. 5 and 6, it was clearly observed that the processing time and memory usage of our proposed, model, IHDGAP is relatively high as compared with

Table 1: Performance analysis of the proposed IHDGAP vs. our previous G-HRF)

Parameters/predictors	G-HRF	IHDGAP
Accuracy (%)	84.71	90.96
Processing time (msec)	1853	2339
Sensitivity (%)	84.26	91.43
Specificity (%)	85.17	90.51
Fscore	0.85	0.91
Memory usage (B)	1250740	1262604

G-HRF mechanism as the proposed model uses CNN for training the data to achieve better accuracy. From the Fig. 7, it was clearly noticed that the Classification accuracy of the proposed model IHDGAP is better than that of our previous classifier, G-HRF. From Fig. 8 and 9, it is observed that the proposed model IHDGAP is performing well in terms of sensitivity and specificity as compared with our previous model called gene signature based HRF cluster (G-HRF).

This is also noticed that our proposed model is reduced misclassification as compared with our previous model G-HRF. That is the prediction scores of true positive and true negative high and false positive and false negative are very low. From the Fig. 10, it is clearly noticed that the achieved Fscore of the proposed model IHDGAP is better than that of G-HRF. That is it is clearly established that the proposed model is classifying and predicting diseases in better manner.

The performances analysis of the proposed Intelligent Human Diseases Prediction technique through Gene Association Prediction mechanism (IHDGAP) in terms of processing time, memory usage/utilization, accuracy, sensitivity, specificity and Fscore are shown in Table 1.

CONCLUSION

This research work is proposed deep learning based Intelligent Human Diseases-Gene Association Prediction technique for High Dimensional Human Diseases data sets (IHDGAP) to maximize the classification accuracy. The proposed model was implemented and results were studied thoroughly. The comparative experimental results established that the proposed classifier out performs our previous model, gene signature based Hierarchical Random Forest (G-HRF) in terms of classification accuracy, specificity, sensitivity and Fscore. However, it was noticed that the proposed model consumes relatively more memory and processing time as this research work employs Convolution Neural Network (CNN) to predict gene associations. The proposed model capable of classifying and predicting the disease patterns more accuratelyas we use deep learning method called convolution neural network algorithm and find out more positive gene sequences and negative gene sequences through the help of association mining.

ACKNOWLEDGEMENT

The researcher acknowledges his guide and co-researcher for supporting.

REFERENCES

Alanis-Lobato, G., C.V. Cannistraci and T. Ravasi, 2014. Exploring the genetics underlying autoimmune diseases with network analysis and link prediction. Proceedings of the 2nd Middle East Conference on Biomedical Engineering, February 17-20, 2014, IEEE, Doha, Qatar, ISBN:978-1-4799-4799-7, pp: 167-170.

Alshalalfa, M. and R. Alhajj, 2013. Using context-specific effect of miRNAs to identify functional associations between miRNAs and gene signatures. BMC. Bioinf., 14: 1-13.

Chen, L., B. Liu and C. Yan, 2018. DPFMDA: Distributed and privatized framework for miRNA-disease association prediction. Pattern Recognit. Lett., 109: 4-11.

Chen, X., C.C. Yan, X. Zhang and Z.H. You, 2016. Long non-coding RNAs and complex diseases: From experimental results to computational models. Briefings Bioinf., 18: 558-576.

Chen, X., C.C. Yan, X. Zhang, Z.H. You and L. Deng *et al.*, 2016b. WBSMDA: Within and between score for MiRNA-disease association prediction. Sci. Rep., 6: 1-9.

Chen, X., Y.W. Niu, G.H. Wang and G.Y. Yan, 2017. Hamda: Hybrid approach for MiRNA-disease association prediction. J. Biomed. Inf., 76: 50-58.

Cheng, S., M. Guo, C. Wang, X. Liu and Y. Liu *et al.*, 2015. MiRTDL: A deep learning approach for miRNA target prediction. IEEE/ACM. Trans. Comput. Boil. Bioinf., 13: 1161-1169.

Conze, P.H., V. Noblet, F. Rousseau, F. Heitz and R. Memeo *et al.*, 2016. Random forests on hierarchical multi-scale supervoxels for liver tumor segmentation in dynamic contrast-enhanced CT scans. Proceedings of the 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI), April 13-16, 2016, IEEE, Prague, Czech Republic, ISBN:978-1-4799-2349-6, pp: 416-419.

Fan, H., F. Zhang, Y. Xu, X. Huang and G. Sun *et al.*, 2010. An association study of DRD2 gene polymorphisms with schizophrenia in a Chinese Han population. Neurosci. Lett., 477: 53-56.

Hoe, H.S. and G.W. Rebeck, 2008. Functional interactions of APP with the apoE receptor family. J. Neurochem., 106: 2263-2271.

Hu, W., 2015. High accuracy gene signature for chemosensitivity prediction in Breast Cancer. Tsinghua Sci. Technol., 20: 530-536.

Kukreti, R., S. Tripathi, P. Bhatnagar, S. Gupta and C. Chauhan *et al*., 2006. Association of DRD2 gene variant with schizophrenia. Neurosci. Lett., 392: 68-71.

Li, J.Q., Z.H. Rong, X. Chen, G.Y. Yan and Z.H. You, 2017. MCMDA: Matrix completion for MiRNA-disease association prediction. Oncotarget, 8: 21187-21199.

Paul, D., R. Su, M. Romain, V. Sebastien and V. Pierre *et al*., 2017. Feature selection for outcome prediction in Oesophageal Cancer using genetic algorithm and random forest classifier. Computerized Med. Imaging Graphics, 60: 42-49.

Phongwattana, T., W. Engchuan and J.H. Chan, 2015. Clustering-based multi-class classification of complex disease. Proceedings of the 2015 7th International Conference on Knowledge and Smart Technology (KST), January 28-31, 2015, IEEE, Chonburi, Thailand, ISBN:978-1-4799-6048-4, pp: 25-29.

Sakthivel, N.K., N.P. Gopalan and S. Subasree, 2016. A comparative study and analysis of DNA sequence classifiers for predicting human diseases. Prceedings of the International Conference on Informatics and Analytics (ICIA-16), August 25-26, 2016, ACM, Pondicherry, India, ISBN:978-1-4503-4756-3, pp: 1-5.

Sakthivel, N.K., N.P. Gopalan and S. Subasree, 2017. G-HR: Gene signature based HRF cluster for predicting human diseases. Intl. J. Pure Appl. Math., 117: 157-161.

Shi, C., X. Kong, Y. Huang, P.S. Yu and B. Wu, 2014. HeteSim: A general framework for relevance measure in heterogeneous networks. IEEE. Trans. Knowl. Data Eng., 26: 2479-2492.

Tahmasebipour, K. and S. Houghten, 2014. Disease-gene association using a genetic algorithm. Proceedings of the 2014 IEEE International Conference on Bioinformatics and Bioengineering, November 10-12, 2014, IEEE, Boca Raton, Florida, USA., ISBN:978-1-4799-7502-0, pp: 191-197.

Zahari, Z., L.K. Teh, R. Ismail and S.M. Razali, 2011. Influence of DRD2 polymorphisms on the clinical outcomes of patients with schizophrenia. Psychiatric Genet., 21: 183-189.

Zeng, X., Y. Liao, Y. Liu and Q. Zou, 2017. Prediction and validation of disease genes using HeteSim Scores. IEEE/ACM. Trans. Comput. Biol. Bioinf., 14: 687-695.