IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

IEEE ON IEEE ENGINEERING IN MEDICINE AND BIOLOGY SOCIETY SECTION

Received August 8, 2020, accepted August 23, 2020, date of publication August 31, 2020, date of current version September 15, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3020592

Analysis for Disease Gene Association Using Machine Learning

MISBA SIKANDAR[1], RAFIA SOHAIL[1], YOUSAF SAEED[1], ASIM ZEB[2], MAHDI ZAREEI[3], (Senior Member, IEEE), MUHAMMAD ADNAN KHAN[4], ATIF KHAN[5], ABDALLAH ALDOSARY[6], AND EHAB MAHMOUD MOHAMED[7,8], (Member, IEEE)

[1]Department of Information Technology, The University of Haripur, Haripur 22620, Pakistan
[2]Department of Computer Science, Abbottabad University of Science and Technology, Havelian 22500, Pakistan
[3]School of Engineering and Sciences, Tecnológico de Monterrey, Zapopan 45201, Mexico
[4]Department of Computer Science, Lahore Garrison University, Lahore 54000, Pakistan
[5]Department of Computer Science, Islamia College Peshawar, Peshawar 25000, Pakistan
[6]Department of Computer Science, Prince Sattam Bin Abdulaziz University, As Sulayyil 11991, Saudi Arabia
[7]Electrical Engineering Department, College of Engineering, Prince Sattam Bin Abdulaziz University, Wadi Addwasir 11991, Saudi Arabia
[8]Electrical Engineering Department, Faculty of Engineering, Aswan University, Aswan 81542, Egypt

Corresponding author: Mahdi Zareei (m.zareei@ieee.org)

ABSTRACT To recognize the basis of disease, it is essential to determine its underlying genes. Understanding the association between underlying genes and genetic disease is a fundamental problem regarding human health. Identification and association of genes with the disease require time consuming and expensive experimentations of a great number of potential candidate genes. Therefore, the alternative inexpensive and rapid computational methods have been proposed that can identify the candidate gene associated with a disease. Most of these methods use phenotypic similarities due to the fact that genes causing same or similar diseases have less variation in their sequence or network properties of protein-protein interactions based on-premises that genes lie closer in protein interaction network that causes the similar or same disease. However, these methods use only basic network properties or topological features and gene sequence information or biological features as a prior knowledge for identification of gene-disease association, which restricts the identification process to a single gene-disease association. In this study, we propose and analyze some novel computational methods for the identification of genes associated with diseases. Some advance topological and biological features that are overlooked currently are introducing for identifying candidate genes. We evaluate different computational methods on disease-gene association data from DisGeNET in a 10-fold cross-validation mode based on TP rate, FP rate, precision, recall, F-measure, and ROC curve evaluation parameters. The results reveal that various computational methods with advanced feature set outperform previous state-of-the-art techniques by achieving precision up to 93.8%, recall up to 93.1%, and F- measure up to 92.9%. Significantly, we apply our methods to study four major diseases: Thalassemia, Diabetes, Malaria, and Asthma. Simulation results show that the proposed Deep Extreme Learning Machine (DELM) gives more accurate results as compared to previously published approaches.

INDEX TERMS Disease gene association, computational approaches, privacy, topological features, biological features, protein-protein interaction network (PPIN), electron-ion interaction pseudopotential (EIIP).

I. INTRODUCTION

A gene is the basic physical and functional unit of heredity that is responsible for different biological processes in an organism. The mutation in a single gene sequence may mutate a biological process and leads to a certain disease. The genes in the human body are not isolated they interact with one another, therefore, the mutation in a single gene may affect its interacting gene which may also play a part in the mutation of different biological processes and cause different diseases. Therefore, consideration of biological mechanisms and based on these mechanisms discovering the relationship between the diseases and genes is a serious challenge in modern biology and medicine.

The associate editor coordinating the review of this manuscript and approving it for publication was Kin Fong Lei.

Understanding the association between casual genes and their genetic disease is a fundamental problem regarding human health [1]. Technology is involved in the detection and monitoring of various human diseases such as Parkinson [2]. Also, the Internet of Medical Things (IoMT) is in focus for addressing human health [3]. Different experimental methods have been proposed to associate genes with a disease but these methods are expensive in terms of cost and time [4]. For that reason, alternative computational approaches are gaining popularity for disease-gene-association. These computational approaches identify or prioritize genes associated with a disease based on genes sequences and genes interactions. When genes involved in disease are not known then the prioritization task is done based on different features like computing the similarities between the known disease genes and a given gene [5], as shown in Figure 1. It illustrates the diseases (D1-D4), known causal genes (G1-G4), and the connection lines are used to represent the connections to the diseases. Features are evaluated using gene information (F1-F5) and the connection lines are used to represent the interactions with known causal genes.

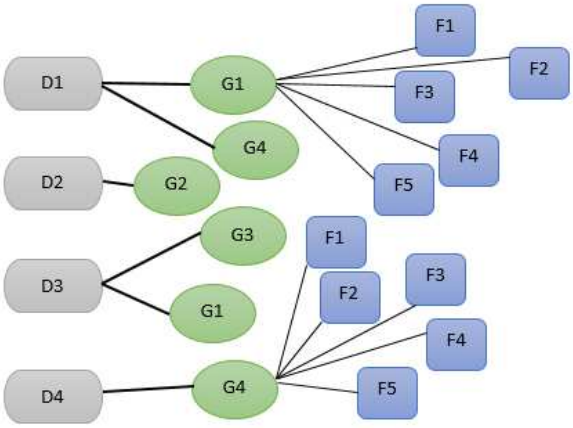


FIGURE 1. Illustration of the disease gene association.

While prioritizing a gene to be a candidate gene, a threshold is used to determine its likelihood to be involved in a disease. For the prioritization of candidate genes various approaches were proposed which previously used phenotypic similarities measure, focused on the heterogeneous network, and some on comprehensive and accurate protein-protein interaction data by utilizing known disease genes data [6]. In [7], PROSPECTR is presented that utilizes sequence-based features and identify the genes participating in Mendelian and oligogenic disorders. Similarly, [5] used the Random Walk examination which describes similarity in protein-protein interaction utilizing artificial linkage interval. It contains the first 100 genes situated nearby to the disease gene, based on genetic distance on the same chromosome. For the ranking of the genes, they also used PROSPECTR [5]. A method of candidate gene prioritization is defined in [8] that is exclusively grounded on the protein-protein interaction network (PPIN). Other than functional annotation of data and protein interaction data, another method for the selection of genes is proposed in [9] which utilized 1D discrete wavelet-transform-based choice of genes. This technique allows scores to genes between two classes for distinguishing samples. Using various levels of the wavelet transform it decompose gene expression signal. The genes are selected having maximum scores for the formation of a feature set and trial classification.

A worldwide network-based technique is provided in [10] for the prioritization of disease genes and assuming protein complex links. They developed a method named PRINCE based on the prioritization function and its constraints that relate to its uniformity over the usage of prior information and network. They used the technique that not only predicts the gene associations but correspondingly protein complex associations by the disease of concern [10]. Gene prioritization is used for the identification of the disease gene but also for the identification of favorable candidates from various studies that generate gene lists [11]. Cofunction Networks (CFNs) is used based on mutual functional similarity to prioritize clusters of candidate genes from numerous disease-associated loci [12]. Recently, a new type of technique for prioritizing candidate genes associated with a certain disease is proposed. It is a knowledge-based methodology that studies gene-gene association tendency in diseases from acknowledged gene-disease association. Mutual information is used to quantify the strength of gene-gene association in a certain disease [13].

Most of the above-mentioned techniques emphase on prioritizing self-determining genes, in various circumstances mutations at different loci might lead to the alike disease [10]. However, these approaches perform well but they still have some limitations [6]. PROSPECTR is not able to achieve reliable, and sizeable set of genes in complex traits, this intended that algorithm performance could not be verified on the constituents of complex diseases [7]. A method in [5] can only be used for those genes whose protein-protein interactions are known or predicted. Similarly, a technique in [8] is applicable only for the known disease-related (seeds) genes. The wavelet-based method in [9] regularly outperforms only when the number of selected genes were more than 40. PRINCE depends on prior knowledge of phenotype that bounds the application such as diseases that are phenotypically related to diseases with identified fundamental genes. Secondly, PRINCE computation was not considered other related data such as genes that are differentially expressed in the disease state, it uses known disease-gene associations [10]. Knowledge-based approach chooses a substitute technique such as Know-GENE without using mutual information or the network propagation technique. It does not perform well for genes that do not exist in any of the diseases used in originating the mutual information [13]. Therefore, in this study, we proposed and analyze some novel computational methods for the identification of genes associated with diseases based on some advanced biological

features. Biological features are calculated based on gene sequence information. Furthermore, we also test our data on advance topological features. Topological features are calculated based on protein complexes. Most of these biological and topological features were not used in previous studies. As we extracted the biological feature set by applying the discrete wavelet transform on EIIP values of genes amino acids. However, [9] used 1D DWT for the choice of genes but they do not utilize EIIP values of genes amino acid sequence. Similarly, [14] used only degree connectivity and betweenness centrality in their computational pipeline for the prioritization of genes associated with the disease. They do not utilize some other topological features which we proposed and achieve remarkable results. Secondly, we use a supervised learning method for the identification of genes associated with genes that were not used before. Previously, most of the proposed methods used unsupervised prioritization techniques.

To analyze different computational techniques, the known disease genes are downloaded from DisGeNET, sequences of genes from UniProt, the binary protein interactions from HPRD (Human protein reference database), and the true human protein complexes are from Comprehensive Resource of Mammalian protein complexes (CORUM). By using all these data resources, we have extracted different biological and topological features. The extracted feature set is passed to different computational methods for classification using 10-fold validation mode. We analyze the power of the proposed methods by reviewing four major diseases: Malaria, Asthma, Thalassemia, and Diabetes. As more than one gene is responsible for a single disease so our approach identifies and associates it respectively by using advanced features. We analyze that by using an advanced feature set, most of the proposed computational methods perform well by achieving the highest accuracy demonstrated in Section IV. Furthermore, we compared our proposed methods with the previous methods and found that the proposed methods outperform by using advance topological and biological features.

Although numerous computational methods have been proposed for the disease-gene association, the proposed analysis introduces a new feature set along with existing one and gives insight into the behavior of various novel machine learning models that have not been tested previously. Instead of exploring the behavior of only a single machine learning model, this study investigates the prediction power of different machine learning models with the newly introduced feature set. It is shown that these models perform more accurately and precisely than the previous models with an advanced feature set. Furthermore, this study grouped different machine learning models behavior for disease-gene association in a single manuscript which will be helpful for future study and health community.

Our article comprises of the opening paragraphs, which provide an initial impression about the logic of our argument. Section II explains the suggested method and approaches. In Section III, proceeding paragraphs will provide implications for our research. Section IV will demonstrate results derived from the proposed methodology, and Section V concludes the paper.

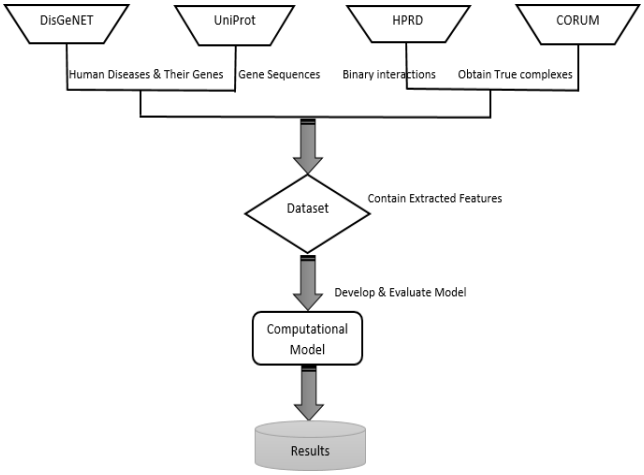


FIGURE 2. Schematic overview of the methodology.

II. MATERIAL AND METHODS

The flow chart of the research methodology is shown in Figure 2. It depicts datasets that are downloaded from different sources. Secondly, different biological and topological features are extracted by utilizing the datasets and finally, testing of different computational models are carried out using the feature set and the behavior of these models are analyzed in Figure 2

A. DATASETS AND SOURCES

The diseases and their respective genes are downloaded from DisGeNET (http://www.disgenet.org/) and the sequences of genes that are involved in given diseases are downloaded from UniProt (http://www.uniprot.org/). DisGeNET is one of the largest databases containing collections of genes involved in human diseases. UniProt is a database of the protein sequence. The binary protein interactions are downloaded from HPRD (http://www.hprd.org/) and the true human protein complexes are downloaded from CORUM (http://mips.helmholtz-muenchen.de/corum/). All of these are publically available datasets. For analysis and evaluation Weka (http://www.cs.waikato.ac.nz/ml/weka/) is used for data mining. Table 1 comprises our extracted dataset statistics.

TABLE 1. Dataset statistics.

Total number of genes	874
Total number of diseases	4
Number of sequences	874
Total number of complexes	1925
Number of complexes with no interactions found in HPRD	1205
Number of complexes with more than three interactions	500

B. FEATURE EXTRACTION

Subsequently, after downloading different datasets, useful biological and topological information is extracted to be utilized as a feature set. The specifics of feature sets are specified as:

a) *Biological Features:* The biological features are extracted by using the amino acid sequence information of the gene. Amino acids play a vital role and act as structure blocks of genes. A mutation in an amino acid sequence may lead to a certain disease thus genes causing certain diseases may have a similar amino acid sequence structure. That is the reason for biological feature computation. This feature set consists of length, entropy, and discrete wavelet features. Length and entropy were used previously for protein complex identification [19], [20] not for the disease-gene association. However, discrete wavelet features from EIIP values are not used previously in any study, to the best of our knowledge. We are the first one to utilize the EIIP values of amino acid for feature extraction by applying a discrete wavelet transform.

b) *Length:* Gene sequence is the combination of different amino acids, and different sequences have different lengths. We compute length by counting the number of amino acids in a sequence. So, the length is the total number of amino acids appearance in a sequence.

c) *Entropy:* The entropy can be estimated by computing the correct probabilities of a distinct probability and defined in equation 1.

$$E = -\sum_{ffln=1}^{ffl20} (p_n \times \log_2 p_n) \tag{1}$$

where p_n is the probability of an amino acid in a sequence.

d) Discrete Wavelet Transform (DWT): It is a tool that can be used to extract useful information from any data source without losing any data and with a distinction between important and non-important data with high speed [15]. Therefore, in this study, it is used for extracting useful information from gene sequences. For computing discrete wavelet features, initially, we replace the amino acids in a sequence with their EIIP values. The EIIP values designate the regular states of energy for all of the valence electrons in the specific amino acid [16]. Afterward, we apply discrete wavelet on these values which return approximation coefficient and detail coefficients for EIIP values of a sequence. These approximation and detail coefficients are utilized as a feature set to identify the underlying gene associated with a disease. There are separate EIIP values for each of the amino acids. The list of amino acids, codes, and EIIP values are given in Table 2.

TABLE 2. Amino acids.

Name	Representation	EIIP Values
Alanine	A	0.0373
Arginine	R	0.0959
Asparagine	N	0.0036
Aspartic acid	D	0.1263
Cysteine	C	0.0829
Glutamine	Q	0.0761
Glutamic acid	E	0.0058
Glycine	G	0.0050
Histidine	H	0.0242
Isoleucine	I	0
Leucine	L	0
Lysine	K	0.0371
Methionine	M	0.0823
Phenylalanine	F	0.0946
Proline	P	0.0198
Serine	S	0.0829
Threonine	T	0.0941
Tryptophan	W	0.0548
Tyrosine	Y	0.0516
Valine	V	0.0057

1. Topological Features: The topological features are extracted by utilizing the known human gene complexes from Corum and gene binary interactions from HPRD. When a gene interacts with other genes, it forms a complex which could be represented as a graph. A graph consists of vertices and edges and can be directed or undirected. The vertices are connected with each other by edges. If edges have direction from one vertex to another than it is said to be directed graph otherwise undirected. As a gene interact with other gene and forms a complex therefore genes could be represented as vertices and binary interaction between genes could be represented as edges. Here, we consider the gene interactions as undirected. The attainable protein-protein interaction (PPI) in the human genome provides a novel opportunity for discovering hereditary genes-disease by topological features from the PPI network [17]. It is important to analyze genes to prevent genetic problems [18], [23]. Moreover, the genes causing the same or similar disease may lie closer to each other.

Therefore, we utilized the topological structure of genes and compute a feature set known as topological features. The topological feature set consists of degree, eccentricity, Neighborhood Connectivity, Average Shortest Path Length, Betweenness Centrality, Closeness Centrality, Clustering Coefficient, Radiality, Topological Coefficient, and Stress Centrality [19]. This feature set is previously used for protein complex identification [19], [20] but not for the disease-gene association. Results reveal that with the use of this advanced topological feature set various computational methods to perform significantly.

After computing the feature set for different genes. We trained various computational models to classify these genes based on different features into four different disease classes, i.e, Thalassemia, Diabetes, Malaria, and Asthma. The prediction power of these computational models is analyzed

by different evaluation parameters discussed in section III. However, the performance is exhibited in section IV, i.e, results, and discussion.

C. PROPOSED DEEP EXTREME LEARNING MACHINE FRAMEWORK

The deep extreme learning machine (DELM) is well-known for forecasting health conditions, forecasting electricity use, transport and traffic control, etc. The DELM may be widely used in various contexts for classification and regression objectives since DELM learns efficiently. Extreme learning machine is a neural network system that only enables data to move one direction across several layers, but we have utilized the back-propagation approach in this proposed model during the training process as data flows back through the network. The weights of the network are constant during the validation process, where we import the trained model and estimate the real data. The DELM model integrates the input layer, several hidden layers, and one output layer.

In the DELM method, the n-th input node, the ith hidden node, and the pth output node can be considered as an, mi, and gm, correspondingly, while all the N input nodes, l hidden nodes, and P output nodes can be considered as $\grave{a} = \lceil \grave{a}_1, \grave{a}_2 \ldots\ldots \grave{a}_N \rceil^T \in R^N$, $\mathcal{M} = \lceil m_1, m_2 \ldots\ldots m_l \rceil^T \in R^l$, $\ddot{u} = \lceil 1, \ddot{u}_2 \ldots\ldots \ddot{u}_M \rceil^T \in R^P$, correspondingly. The DELM framework will, therefore, be characterized densely as;

$$\mathcal{M} = f\,(B_{\grave{a}} + c) \tag{2}$$

And

$$\ddot{U} = QB \tag{3}$$

where $B = [b_{in}] \in R^{lxN}$, $c = \lceil c_1, c_2 \ldots\ldots c_l \rceil^T \in R^l$, $Q = \left[q_{mi}\right] \in R^{pxl}$, and the activation function f ( ) could be used as sigmoid, linear Gaussian models, etc.

Assume there are just V diverse training records, and let $a_v \in R^N$ and $\ddot{u}_v \in R^P$ and symbolize the vth training input and the subsequent vth training output, correspondingly, where $v = 1, 2, \ldots\ldots V$. In the training dataset the input arrangement and output arrangement can be indicated as;

$$\grave{A} = \lceil \grave{a}_1, \grave{a}_2 \ldots\ldots \grave{a}_V \rceil^T \in R^{NxV} \tag{4}$$

and

$$\ddot{U} = \lceil \ddot{u}_1, \ddot{u}_2 \ldots\ldots \ddot{u}_V \rceil^T \in R^{PxV} \tag{5}$$

correspondingly. We can alternate (4) into (5) to obtain

$$\mathcal{M} = f\left(B\grave{A} + 1^T \otimes c\right), \tag{6}$$

where $\mathcal{M} = \lceil m_1, m_2 \ldots\ldots m_V \rceil^T \in R^{lxV}$ is the assessment grouping of all l hidden nodes, and $\otimes$ Kronecker product. Then we can override (6) and (5) in (3) to attain the real training implementation sequence.

$$\ddot{U} = Q\mathcal{M} \tag{7}$$

In DELM, the output weight Q is adaptable, while B (i.e., the input weights) and c (i.e., the biases of the hidden nodes) are randomly focused. Entitle the expected outcome as Y. Then DELM only minimizes the valuation inaccuracy;

$$E = Y - \ddot{U} = Y - Q\mathcal{M} \tag{8}$$

By verdict the least-squares explanation Q for the problem

$$\min_{Q} \|E\|_F^2 = \min_{Q} \|Y - Q\mathcal{M}\|_F^2 \tag{9}$$

where $\|.\|_F$ specifies the Frobenius norm.

For the problematic (9), the outstanding least norm least-squares solution is;

$$Q = Y\mathcal{M}^T \left(\mathcal{M}\mathcal{M}^T\right)^{-1} \tag{10}$$

To evade overfitting, the general Tikhonov regularization can be employed to amend Eq. (10) into

$$Q = Y\mathcal{M}^T \left(\mathcal{M}\mathcal{M}^T + v_0^2 I\right)^{-1}, \tag{11}$$

where $v_0^2 > 0$ indicates the regularization expression. Evidently, Eq. (11) is only the specific case of Eq. (10) with $v_0^2 = 0$. Consequently, we locate only Eq. (11) the regularization of Tikhonov for the DELM. Machine learning is a general approach for progressively mounting the number of hidden layers to the preferred accuracy. When this method is affected unswervingly in DELM, nevertheless, in Eq. (11) the opposite matrix process for standard ELM is compulsory when some or only another hidden node is enhanced, and the algorithm is therefore unaffordable to computation. The back-propagation process comprises weight initialization, feedforward propagation, back error propagation, and update of weight and uniqueness. An activation function like $g(x) =$ ***sigmoid*** occurs on each neuron in the hidden layer. This permits the sigmoid input feature and the DELM hidden layer to be constituted in this way;

$$\begin{aligned} E &= \frac{1}{2}\sum_j \left(sj - wp_j\right) 2 \\ sj &= \text{Expected output} \\ wp_j &= \text{considered output} \end{aligned} \tag{12}$$

Eq. (12) designates a back-propagation error that can be computed by splitting the sum of the square from the expected outcome by 2. The weight change is compulsory to decrease the common error. The rates of weight change for the output layer are presented in Eq. (13).

$$\begin{aligned} \Delta\mathfrak{H}_{i,j}^{l=6} &\propto -\frac{\partial \mathbf{R}}{\partial \mathfrak{H}^{l=6}} \\ \textbf{where } i &= 1, 2, 3 \ldots\ldots\ldots 10 \textbf{ (Neurons)} \end{aligned} \tag{13}$$

and j = **output Layer**

$$\Delta\mathfrak{H}_{i,j}^{l=6} = -\textbf{const}\frac{\partial \mathbf{R}}{\partial \mathfrak{H}^{l=6}} \tag{14}$$

inscribing Eq. (14) by exercising the chain rule technique

$$\Delta\mathfrak{H}_{i,j}^{l=6} = -\textbf{const}\frac{\partial \mathbf{R}}{\partial wp_j^l} \times \frac{\partial wp_j^l}{\partial Nh\mathfrak{H}_j^l} \times \frac{\partial Nh\mathfrak{H}_j^l}{\partial \mathfrak{H}_{i,j}^l} \tag{15}$$

The value of change weight can be attained after exchanging the values in Eq. (14) as presented in Eq. (15).

$$\Delta\mathfrak{H}_{i,j}^{l=6} = \mathbf{const}(\mathbf{sj}-wp_j) \times (wp_j^l(1-wp_j^l)\times wp_j^l) \quad (15a)$$

From wp to $\mathfrak{H}_6$

$$\Delta\mathfrak{H}_{i,j}^{l=6} = \mathbf{const} \ni_j wp_j^l \quad (15b)$$

The measurement for corresponding weight adaptation to the hidden weight can be seen in the next step. This is more complicated since by weighted connection it can lead to misinterpretation on any node.

From $\mathfrak{H}_6$ to $\mathfrak{H}_1$ or $\mathfrak{H}_n$

Where $\boldsymbol{n} = \mathbf{5}, \mathbf{4}, \mathbf{3}, \mathbf{2}, \mathbf{1}$

$$\Delta\mathfrak{H}_{i,n}^{l} \propto -\left[\sum_j \frac{\partial \mathbf{R}}{\partial \mathbf{wp}_j^l} \times \frac{\partial \mathbf{wp}_j^l}{\partial \mathbf{Nh}\mathfrak{H}_j^l} \times \frac{\partial \mathbf{Nh}\mathfrak{H}_j^l}{\partial \mathfrak{H}_n^l}\right] \times \frac{\partial \mathbf{wp}_n^l}{\partial \mathbf{Nh}\mathfrak{H}_n^{hl}} \times \frac{\partial \mathbf{Nh}\mathfrak{H}_n^l}{\partial \mathfrak{H}_{i,n}^l} \quad (16a)$$

$$\Delta\mathfrak{H}_{i,n}^{l} = \boldsymbol{R}\left[\sum_j \ni_j(\mathfrak{H}_{n,j}^l)\right] \times \mathbf{wp}_n^l(\mathbf{1}-\mathbf{wp}_j^l)\times \boldsymbol{Z}_{i,n} \quad (16b)$$

$$\Delta\mathfrak{H}_{i,n}^{l} = \boldsymbol{R}\ni_n \boldsymbol{Z}_{i,n} \quad (16c)$$

where

$$\ni_n = \left[\sum_j \ni_j(\mathfrak{H}_{n,j}^l)\right] \times wp_n^l(1-wp_n^l) \quad (16d)$$

The technique to develop the weight and bias among the output and the hidden layer is presented in Eq. (16e).

$$\mathfrak{H}_{i,j}^{l=6}(t) = \mathfrak{H}_{i,j}^{l=6}(t) + \boldsymbol{\lambda}\Delta\mathfrak{H}_{i,j}^{l=6} \quad (16e)$$

Eq. (17) indicate how updating the weight and bias between the input and the hidden layer.

$$\mathfrak{H}_{i,n}^{l}(t) = \mathfrak{H}_{i,n}^{l}(t+1) + \boldsymbol{\lambda}\Delta\mathfrak{H}_{i,j}^{l} \quad (17)$$

III. EVALUATION PARAMETERS

To evaluate the performance of various computational model's TP rate, FP rate, precision, recall, F-measure, and ROC area are used as evaluation parameters.

$$TPR = \frac{TP}{TP+FN} \quad (18)$$

a) *True Positive Rate (TPR):* It is the ability of any algorithm which correctly predicts those having the disease.

b) *False Positive Rate (FPR):* It is the ability of any algorithm which correctly predicts those not having the disease.

$$FPR = \frac{FP}{FP+TN} \quad (19)$$

c) *Precision:* It is the fraction between the relevant and retrieved elements.

$$Precision = \frac{TP}{TP+FP} \quad (20)$$

d) *Recall:* It is the fraction between relevant and total elements that are relevant and non-relevant.

$$Recall = \frac{TP}{TP+FN} \quad (21)$$

where TP is the number of disease genes that are correctly classified as disease genes, FN is the disease genes classified as non-disease genes, TN is the number of non-disease genes classified as non-disease genes and FP is the number of non-disease genes that classified as disease genes.

e) *F Score:* It is a harmonic mean of precision and recall.

$$Fmeasure = 2 \times \frac{REC \times PREC}{PREC + REC} \quad (22)$$

$$F_{0.5} = 1.25 \times PREC \times \frac{REC}{0.25 \times PREC + REC} \quad (23)$$

$$F_1 = 2 \times PREC \times \frac{REC}{PREC + REC} \quad (24)$$

$$F_2 = 5 \times PREC \times \frac{REC}{4 \times PREC + REC} \quad (25)$$

f) *ROC Area:* A ROC curve is a function that is plotted between true positive rate and false-positive rate for distinct cut-off points of a variable.

IV. RESULTS AND DISCUSSION

We analyze some novel computational methods for disease gene association by using advance biological and topological features. To evaluate the performance of these methods we used the FP rate (false positive), TP rate (True positive), recall, precision, F-measure, and ROC curve. To validate our evaluation, we use 10-fold cross-validation mode. Firstly, we test our data by using only biological features and it is shown in Table 3 that Random forest, Classification Via Regression, and Simple cart outperforms by achieving TP rate up to 93.1%, FP rate up to 11.8%, precision up to 93.8%, recall up to 93.1%, F-measure up to 92.9%, and ROC area 99.1%. The proposed DELM has a TP rate up to 96.95%, FP rate up to 14.67%, precision up to 94.35%, recall up to 93.45%, F-measure up to 93.45%, and ROC area up to 93.36%. It also observed that proposed DELM gives more accurate results as compared to Random forest, Classification Via Regression, and Simple cart.

We observed that by using biological features, we have achieved the highest accuracy, i.e., up to 90%-99% but the FP rate is quite less. To increase the FP rate, we tried another method and test our data by using only topological features and investigate the behavior. It is shown in Table 4 that the FP rate is increased significantly but the TP rate decreases. Moreover, with topological features, Random forest, PART, Logit boost and proposed DELM outperform as compared to other methods. As Random Forest has a TP rate up to 64.6%, FP rate up to 48.3%, precision up to 60.8%, Recall

TABLE 3. Computational approaches and their results on the biological feature set.

Name of Classifier	TP Rate	FP Rate	Precision	Recall	F measure	ROC	$F_{0.5}$	F_1	F_2
Random Forest	93.1	11.8	93.8	93.1	92.9	99.1	93.66	93.45	93.24
Naïve Bayes	78.0	11	82.0	78.0	70.0	91.0	81.17	79.95	78.77
Classification via clustering	55.1	59.6	42.1	55.1	47.5	47.6	44.18	47.73	51.90
Classification via regression	93.1	11.8	93.8	93.1	92.9	99.0	93.66	93.45	93.24
PART	79.7	33.4	83.2	79.7	77.1	89.5	82.48	81.41	80.38
Logit boost	86.3	15.1	85.9	86.3	85.8	94.2	85.98	86.10	86.22
Multiclass classifier	92.8	10.2	93.0	92.8	92.6	97.3	92.96	92.90	92.84
Conjunctive rule	66.4	54.7	55.1	66.4	56.6	58.0	57.04	60.22	63.78
Simple cart	93.1	11.8	93.8	93.1	92.9	96.7	93.66	93.45	93.24
Decision table	85.3	13.7	86.7	85.3	93.7	92.5	86.42	85.99	85.58
DELM	96.95	14.67	94.35	96.95	93.45	93.36	93.66	93.45	93.24

up to 64.6%, F-measure up to 50.6%, and ROC area up to 66.7%. PART have TP rate up to 64.5%, FP rate up to 50.7%, precision up to 59.4%, Recall up to 64.5%, F-measure up to 58.4% and ROC area up to 64.0%. Logit boost have TP rate up to 65.3%, FP rate up to 54.9%, precision up to 57.9%, Recall up to 65.3%, F-measure up to 56%, and ROC area up to 66.1%. Proposed DELM has TP rate up to 95.31%, FP rate up to 30.04%, precision up to 89.15%, Recall up to 95.31%, F-measure up to 92.13%, and ROC area up to 89.34%.

Furthermore, to achieve remarkable TP and FP rates both, we combine the biological features with topological features and test the data. The results are shown in Table 5, it is evident in Table 5 that the TP rate increases significantly as compared to the results of topological features only. Where FP rate varies with a little margin when compared to the results of only biological features. However, when compare FP rate to results of only topological features it decreases. Moreover, with combine biological and topological features Random forest, PART, Multiclass classifier, and proposed DELM outperforms. As shown in Table 5, Random Forest have the highest TP rate up to 93.0%, FP rate up to 10.8%, precision up to 93.4%, recall up to 93.0%, F-measure up to 92.8% and ROC area up to 99.0%. The PART has a TP rate up to 90%, FP rate up to 7.5%, precision up to 90.1%, recall up to 90%, F-measure up to 90.1%, and ROC area up to 91.0%. The Multiclass classifier has a TP rate up to 90.7%, FP rate up to 5.1%, precision up to 91.4%, recall up to 90.7%, F-measure up to 91%, and ROC area up to 94.2%. The proposed DELM has a TP rate up to 97.23%, FP rate up to 18.39%, precision up to 93.91%, recall up to 96.64%, F-measure up to 94.26%, and ROC area up to 88.2%.

It is revealed from results that Random Forest, Classification via Regression, and Simple Cart outperforms in case of classification with only biological features. The Random Forest, PART, and Logit Boost outperform in case of classification with only topological features. In the case of combining both biological and topological features, Random Forest, PART, and Multiclass Classifier outperform. All these algorithms make decisions based on regression analysis between attributes for classification so it can be concluded that the algorithms that use regression analysis give the best results for the disease-gene association. However, Random Forest outperforms in all the experimental setups due to its properties of never over fitting, identifications of important attributes for classification, and multiple trees with multiple levels.

Moreover, the FP rate with only biological features is quite less as compared to TP rate as shown in Table 3 where it is increased by using only topological features as shown in Table 4, however with topological features the FP rate increases significantly but TP rate decreases when compare to biological feature results. Therefore, to achieve a remarkable TP rate and FP rate both the biological and topological features are combined which increases the TP rate significantly as compare to topological feature result and the FP rate vary with a little margin as compared to with biological features shown in Table 3. So, it is concluded that computational approaches with combine biological and topological features achieve the best performance.

V. COMPARITIVE ANALYSIS OF NOVEL COMPUTATIONAL METHODS WITH RESPECT TO TRUE POSITIVE RATE AT DIFFERENT THRESHOLDS

Furthermore, to pretend the robustness of different novel computational methods with advanced feature sets we analyze them at different thresholds with respect to True Positive Rate (TPR). The Figure 3a, 3b, and 3c show number of diseases correctly associated with genes at TPR = 0.1-0.20, 0.21-0.30, 0.31-0.40, 0.41-0.50, 0.51-0.60, 0.61-0.70, 0.71-0.80, 0.81-0.90, and 0.91-1.

It is exhibited from Figures 3a, 3b, and 3c that the number of correctly associated diseases with genes by almost all the computational approaches are more at threshold (TPR) value greater than 0.70. Only a few methods have TPR less than 0.70. Thus, it is concluded that the use of advance biological and topological feature set significantly

TABLE 4. Computational approaches and their results on the topological feature set.

Name of Classifier	TP Rate	FP Rate	Precision	Recall	F measure	ROC	$F_{0.5}$	F_1	F_2
Random Forest	64.6	48.3	60.8	64.6	50.6	66.7	61.52	62.64	63.80
Naïve Bayes	57.0	39.4	54.5	57.0	55.2	59.7	54.98	55.72	56.48
Classification via clustering	57.8	41.0	50.1	57.8	53.5	58.4	51.47	53.68	56.08
Classification via regression	65.9	55.3	65.3	65.9	56.3	63.2	65.42	65.60	65.78
PART	64.5	50.7	59.4	64.5	58.4	64.0	60.35	61.85	63.41
Logit boost	65.3	54.9	57.9	65.3	56.0	66.1	59.24	61.38	63.67
Multiclass classifier	64.0	55.2	62.1	64.0	54.8	61.4	62.47	63.04	63.61
Conjunctive rule	63.6	61.6	51.9	63.6	50.9	53.0	53.88	57.16	60.86
Simple cart	64.5	55.6	54.9	64.5	55.3	63.0	56.58	59.31	62.32
Decision table	64.3	56.2	61.3	64.3	55.0	58.6	61.88	62.76	63.68
DELM	95.31	30.04	89.15	95.31	92.13	89.34	90.32	94.26	94.48

TABLE 5. Computational approaches and their results on combined biological and topological features set.

Name of Classifier	TP Rate	FP Rate	Precision	Recall	F measure	ROC	$F_{0.5}$	F_1	F_2
Random Forest	93.0	10.8	93.4	93.0	92.8	99.0	93.32	93.2	93.08
Naïve Bayes	69.6	18.1	74.5	69.6	70.9	86.1	73.47	71.97	70.53
Classification via clustering	58.6	41.5	50.2	58.6	54.0	58.6	51.68	54.08	56.70
Classification via regression	89.5	11.4	89.3	89.5	89.3	95.1	89.34	89.4	89.46
PART	90.0	7.5	90.1	90.0	90.1	91.0	90.08	90.05	90.02
Logit boost	86.1	15.6	85.8	86.1	85.7	94.2	85.86	85.95	86.04
Multiclass classifier	90.7	5.1	91.4	90.7	91.0	94.2	91.26	91.05	90.84
Conjunctive rule	66.4	54.7	55.1	66.4	56.6	58.0	57.04	60.22	63.78
Simple cart	88.8	11.2	88.6	88.8	88.6	91.4	88.64	88.70	88.76
Decision table	85.9	12.2	87.0	85.9	85.6	94.8	86.78	86.45	86.12
DELM	97.23	18.39	93.91	94.62	94.26	88.2	94.05	94.26	94.48

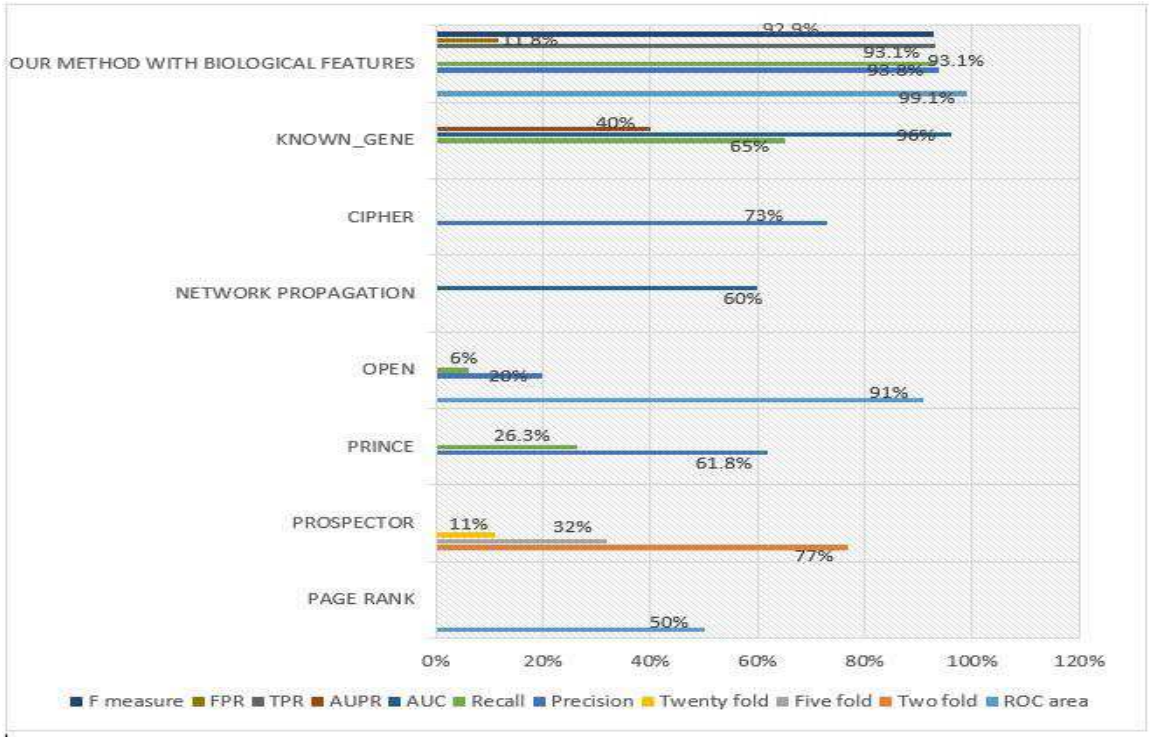
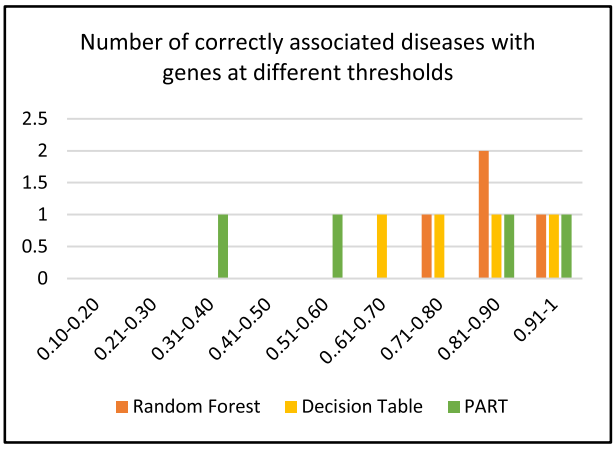
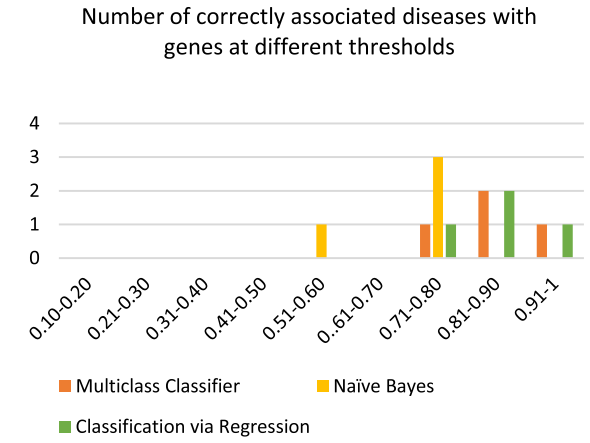


FIGURE 5. Comparison of various computational methods with previous methods by using advance biological features.

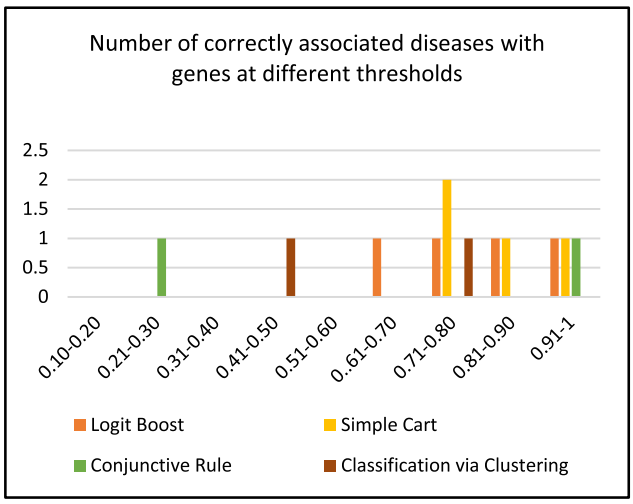
influenced the performance of different computational methods.



(a)



(b)



(c)

FIGURE 3. a. Comparative analysis of novel computational approaches at different thresholds. b. Comparative analysis of novel computational approaches at different thresholds. c. Comparative analysis of novel computational approaches at different thresholds.

VI. COMPARITIVE ANALYSIS OF COMPUTATIONAL COST OF NOVEL COMPUTATIONAL METHODS

We also analyze the computational time taken by each computational method to build a model for classifying genes associated with the diseases. Figure 4 shows the computational cost analysis. It is shown in Figure 4 that Conjunctive Rule takes less time to associate the genes with diseases and Multiclass Classifiers take more time than all other methods. But accuracy matters more than time in biological problems, and Multiclass Classifier yields more accurate result as compare to Conjunctive Rule. Conjunctive Rule has less accuracy than other methods as exhibited in Table 3, Table 4 and Table 5. However, Random Forest supersede all others methods concerning accuracy as elucidated in Table 3, Table 4, and Table 5. To conclude, Random forest is efficient with respect to accuracy while Conjunctive Rule is efficient concerning time complexity.

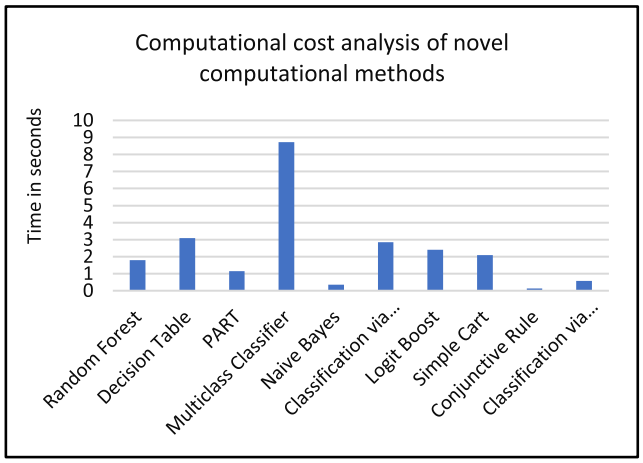


FIGURE 4. Computational cost analysis of novel computational approaches.

VII. COMPARISON WITH OTHER METHODS

We also compared our findings with previous methods and concluded that various computational methods with advance biological and topological features supersede the previous methods. We analyzed the performance of various methods based on different parameters that are FP rate, TP rate, recall, precision, F-measure, and ROC area. However, the previous methods use only one or two evaluation parameters, like PAGE RANK [8] algorithm used ROC area that was 50% accurate. The PROSPECTOR [7] achieved 77% accuracy with two folds, 32% accuracy with five-folds, and 11% accuracy with twenty folds. The PRINCE [10] algorithm was evaluated based on precision and recall, the precision was 61.8% and recall was 26.3%. OPEN [21] achieved precision up to 20%, recall up to 2-6% and ROC area up to 91%. NETWORK PROPAGATION [22] has an Area Under the Curve (AUC) up to 60%, and CIPHER [23] achieved 24.7% precision. KNOWN-GENE [13] have 65% recall, 40% AUPR and 96% AUC. However, comparing all these performances with our analyzed methods with advance biological and topological features, it is evident that our analyzed methods outperform by having an average accuracy of 90%. The comparison of various analyzed computational

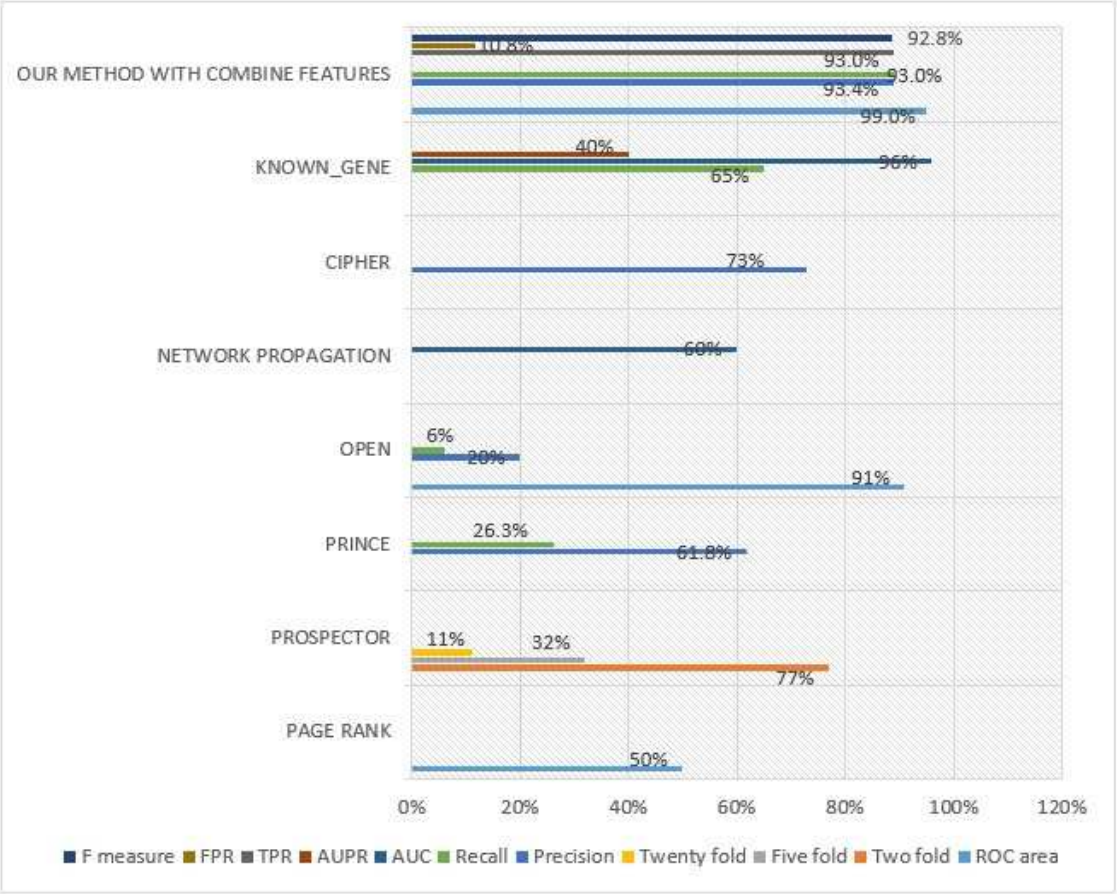


FIGURE 6. Comparison of various computational methods with previous methods by using both biological features and topological features.

methods using biological features, and combined biological and topological features with previous methods are shown in Figure 5, and Figure 6 respectively.

VIII. CONCLUSION AND FUTURE WORK

Understanding the association between underlying genes and genetic disease is a fundamental problem regarding human health. Different computational methods have been devised to tackle this problem. But these methods used limited information for the identification of genes associated with a disease that restricts their performance. Therefore, in this study to improve the performance of computational approaches we use the advanced biological and topological features and analyze different computational methods. The advanced biological features are extracted from genes sequences information and advance topological features are extracted from the network among the genes. It is revealed from results that by using advance biological and topological features the performance of various computational approaches is increased significantly as compared to previous methods. Moreover, the DELM method gives more accurate results as compared to previously published approaches.

In the future instead of using EIIP values of Amino acid, biological properties as polarization, ionization, and hydrophobicity can be utilized. Similarly, weighted features of gene interaction networks can also be used to increase the accuracy of computational approaches. Further possible advancement could be achieved by integrating more verified features. Moreover, the privacy of biological and topological features can be considered in the future to conceal the features from external alteration.

Hardware, Availability, and Performance: The computational experiments were executed on Haier win8.1PC, Intel(R) Core (TM) i3-4010 processor, and 1.70 GHz, 64 operating system, and x64-based processor. The average runtime for inferring protein complexes or completing the cross-validation iterations was a maximum of 5 minutes and a minimum of 2 to 3 minutes. The datasets and code described herein are accessible upon request.

REFERENCES

[1] S. Navlakha and C. Kingsford, "The power of protein interaction networks for associating genes with diseases," *Bioinformatics*, vol. 26, no. 8, pp. 1057–1063, Apr. 2010.

[2] A. Almogren, "An automated and intelligent parkinson disease monitoring system using wearable computing and cloud technology," *Cluster Comput.*, vol. 22, no. S1, pp. 2309–2316, Jan. 2019.

[3] I. Ud Din, A. Almogren, M. Guizani, and M. Zuair, "A decade of Internet of Things: Analysis in the light of healthcare applications," *IEEE Access*, vol. 7, pp. 89967–89979, 2019.

[4] R. M. Piro and F. Di Cunto, "Computational approaches to disease-gene prediction: Rationale, classification and successes," *FEBS J.*, vol. 279, no. 5, pp. 678–696, Mar. 2012.
[5] S. Köhler, S. Bauer, D. Horn, and P. N. Robinson, "Walking the interactome for prioritization of candidate disease genes," *Amer. J. Hum. Genet.*, vol. 82, no. 4, pp. 949–958, Apr. 2008.
[6] Y. Li and J. C. Patra, "Genome-wide inferring gene–phenotype relationship by walking on the heterogeneous network," *Bioinformatics*, vol. 26, no. 9, pp. 1219–1224, May 2010.
[7] R. E. Adie, "Speeding disease gene discovery by sequence based candidate prioritization," *BMC Bioinf.*, vol. 6, no. 1, pp. 1–13, 2005.
[8] B. J. Chen, "Disease candidate gene identification and prioritization using protein interaction networks," *BMC Bioinf.*, vol. 10, no. 1, pp. 1–14, 2009.
[9] D. M. Z.-H. D. Adarsh Jose, "A gene selection method for classifying cancer samples using 1D discrete wavelet transform," *Int. J. Comput. Biol. Drug*, vol. 2, no. 4, pp. 398–411, 2009.
[10] O. Vanunu, "Associating Genes and Protein Complexes with Disease via Network Propagation," *PLoS Comput. Biol.*, vol. 6, no. 1, pp. 1–9, 2010.
[11] Y. Moreau and L.-C. Tranchevent, "Computational tools for prioritizing candidate genes: Boosting disease gene discovery," *Nature Rev. Genet.*, vol. 13, no. 8, pp. 523–536, Aug. 2012.
[12] M. Tan, G. Musso, T. Hao, M. Vidal, C. A. MacRae, and F. P. Roth, "Selecting causal genes from genome-wide association studies via functionally coherent subnetworks," *Nature Methods*, vol. 12, no. 2, pp. 154–159, Feb. 2015.
[13] H. Zhou and J. Skolnick, "A knowledge-based approach for predicting gene–disease associations," *Bioinformatics*, vol. 32, no. 18, pp. 2831–2838, Sep. 2016.
[14] F. Browne, H. Wang, and H. Zheng, "A computational framework for the prioritization of disease-gene candidates," *BMC Genomics*, vol. 16, no. 9, pp. 1–10, Dec. 2015.
[15] T. Li, Q. Li, S. Zhu, and M. Ogihara, "A survey on wavelet applications in data mining," *ACM SIGKDD Explor. Newslett.*, vol. 4, no. 2, pp. 49–68, Dec. 2002.
[16] M. S. Mabrouk, "A study of the potential of EIIP mapping method in exon prediction using the frequency domain techniques," *Amer. J. Biomed. Eng.*, vol. 2, no. 2, pp. 17–22, Aug. 2012.
[17] J. Xu and Y. Li, "Discovering disease-genes by topological features in human protein–protein interaction network," *Bioinformatics*, vol. 22, no. 22, pp. 2800–2805, Nov. 2006.
[18] Y. Qura-Tul-Ein, "DNA pattern analysis using finite automata," *Int. Res. J. Comput. Sci.*, vol. 1, no. 2, pp. 1–4, 2014.
[19] A. Sikandar, W. Anwar, U. I. Bajwa, X. Wang, M. Sikandar, L. Yao, Z. L. Jiang, and Z. Chunkai, "Decision tree based approaches for detecting protein complex in protein protein interaction network (PPI) via link and sequence analysis," *IEEE Access*, vol. 6, pp. 22108–22120, 2018.
[20] W. A. A. A. I. N. G. Misba Sikandar, "IoMT-based association rule mining for the prediction of human protein complexes," *IEEE Access*, vol. 1, pp. 1–12, 2020.
[21] R. C. Deo, G. Musso, M. Tasan, P. Tang, A. Poon, C. Yuan, J. F. Felix, R. S. Vasan, R. Beroukhim, T. De Marco, P.-Y. Kwok, C. A. MacRae, and F. P. Roth, "Prioritizing causal disease genes using unbiased genomic features," *Genome Biol.*, vol. 15, no. 12, pp. 1–19, Dec. 2014.
[22] S. Qian, "Identifying disease associated genes by network propagation," *BMC Syst. Biol.*, vol. 8, no. 1, pp. 1–7, 2014.
[23] X. Wu, "Network-based global inference of human disease genes," *Mol. Syst. Biol.*, vol. 4, pp. 1–11, 2008.