IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received December 16, 2019, accepted December 26, 2019, date of publication January 3, 2020, date of current version January 10, 2020.

Digital Object Identifier 10.1109/ACCESS.2019.2963797

IoMT-Based Association Rule Mining for the Prediction of Human Protein Complexes

MISBA SIKARNDAR[1], WAQAS ANWAR[2], AHMAD ALMOGREN[3], (Senior Member, IEEE), IKRAM UD DIN[1], (Senior Member, IEEE), AND NADRA GUIZANI[4]

[1]Department of Information Technology, University of Haripur, Haripur 22620, Pakistan
[2]Department of Computer Science, COMSATS University Islamabad, Lahore Campus, Lahore 54000, Pakistan
[3]Chair of Cyber Security, Department of Computer Science, College of Computer and Information Sciences, King Saud University, Riyadh 11633, Saudi Arabia
[4]Department of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907, USA

Corresponding author: Ahmad Almogren (ahalmogren@ksu.edu.sa)

The authors are grateful to the Deanship of Scientific Research, King Saud University for funding through Vice Deanship of Scientific Research Chairs.

ABSTRACT The inspiring increase in the Internet-enabling devices has influenced health industry due to the nature of these devices where they offer health related information swiftly. One of the prominent characteristics of these devices is to provide physicians with effective diagnosis of sensitive diseases. Internet of Medical Things (IoMT) is a means of connecting medical devices to computing nodes with the help of Internet for affording real-time communications between patients and clinicians to understand the interaction of human protein complexes. A secure and correct protein complex prediction plays an important job in perceiving the principal method of various cellular determinations and to elucidate the functionality of different un-annotated proteins. Different experimental schemes have been evolved to accomplish this task, however, these schemes have high error rates and are not efficient in terms of time, cost, privacy, and security. To tackle these limitations, numerous computational models have been developed that consider a protein complex as a dense sub-graph and utilize some basic topological properties such as density and degree statistics as a feature set for protein complex prediction. Different kinds of sub-graph structures, e.g., ring, star, linear, and hybrid have also been found in Protein-Protein Interaction Network (PPIN), therefore, more advance topological properties may be helpful to predict these structures. Moreover, the amino acid sequence of protein determines its formation, thus, the sequence information is important for predicting the interacting property among proteins in a secure way. In this study, we have computed basic as well as advance topological features by considering the interaction network of human protein complexes in the IoMT environment. In addition, biological features, i.e., discrete wavelet coefficients, length, and entropy from amino acid sequences of proteins have been computed. The supervised learning method based on association rules such as Partial Tree (PART) and Non-Nested Generalized Exemplars (NNGE) are trained to identify human protein complexes on the basis of integrated topological and biological properties. The 10-fold cross validation is exercised to measure the proposed methods. Experimental results show that association rule learners with integrated features outperform other complex mining algorithms, i.e., probabilistic Bayesian Network (BN), and Random Forest, in terms of accuracy and efficiency in addition to provide privacy.

INDEX TERMS Discrete wavelet transform, NNGE, PPI, PART, privacy, security.

I. INTRODUCTION

Internet of Medical Things (IoMT) is a perception of linking computing nodes to medical equipment with the help of Internet [1]–[3]. To supervise various lingering ailments, different equipment are used in hospitals and healthcare units to protect patients' privacy in a secure way [4], [5]. Nevertheless, with technological evolution, IoMT is becoming more feasible to set up a convincing control over a range of tools for working together in identifying and/or curing numerous health related issues [6]–[9]. In IoMT applications, safeguarding the security of patients' information, systems, and equipment in addition to the privacy of information and information processing, is critical [10]. While IoMT devices offer various aids,

The associate editor coordinating the review of this manuscript and approving it for publication was Victor Hugo Albuquerque.

they also foster life-threatening privacy and security issues. One of the most prominent among different issues is the secure identification of protein complexes. An organism is composed of proteins, which are tiny particles or biological molecules that are made up of amino acid residues. Proteins are different from each other due to variations in amino acid sequence determined by codon and which usually result protein folding into a specific three-dimensional structure that regulates its activity. Different molecular and biological processes in an organism are mediated through protein actions. Proteins interact with each other that result to the formation of protein complexes, which perform different biological functions. A molecular interaction of several proteins with one another at the same locality and time produces a protein complex [11]. Protein complexes mediate different types of functions within an organism, such as replication of DNA, response to stimuli, catalysing metabolic reactions, and molecules transportation to different locations within a cell [12].

The complete elucidation of PPIN in an organism will have important applications for science [13], for example, protein complexes control the mechanism leading to diseased and healthy states in an organism. Therefore, the molecular basis of a disease can be elucidated from protein interaction network, which helps in finding techniques for diagnosis, treatment, and prevention of a disease. Due to these reasons, elucidation of PPIN and protein complex identification is an important goal in the BioNLP field [13]. Protein complex identification is a challenging and the most important task in post genome era. Some of the challenges involve in protein complex detection are as follows:

- Protein-protein interaction data is noisy due to high rates of false negative and false positive.
- A protein may be a participant of more than one complex, i.e., protein complexes may overlap and involve in different biological functions.
- The representation of a protein complex structure such as clique, star, linear or hybrid.

Different experimental and computational methods have been devised to undertake these challenges with maximum accuracy, however, there still exists a bottleneck. The most preferable experimental methods are Tandom Affinity Purification with Mass Spectrometry (TAP-MS) [14] and Yeast to Hybrid (Y2H), but these methods have high error rates and are not efficient in terms of time and cost [15]. To tackle these limitations, numerous computational models have been developed that can be categorised into four groups, i.e., agglomerative, clique finding, traditional graph clustering, and core attachment methods [16]. In agglomerative methods, each single node or a sub-graph makes a cluster at initial stage where these clusters are merged and grow under certain limitations. The two examples of agglomerative methods are MCODE [17] and DPClus [18]. The MCODE selects a seed protein as a primary cluster based on high weight and then increases it. Likewise, the DPClus augments clusters that start from the seed nodes that are selected based on high weights. The examples of Clique finding methods are CFinder [19], CMC [20], and ClusterONE [21]. The CFinder identifies functional modules in the PPIN by utilizing the clique percolation method and detects the k-clique percolation cluster. The CMC predicts the complexes by generating maximal cliques on the basis of weights assigned to protein pairs. These weights are assigned through the iterative scoring method in CMC. Finally, from the generated maximal cliques, highly cohesive clusters are merged or removed on the basis of their interconnectivity. The ClusterOne method augments the cluster on the basis of seed vertex and finds the highly cohesive groups. The graph clustering methods use the premises where a protein complex in the PPIN is subject to a dense sub-graph. An example of traditional graph clustering method is Markove clustering (MCL) [22]. The Markov clustering splits the PPIN into several non-overlapping dense sub-graphs or clusters by simulating random walks within the graph. An architecture, named COACH [23], is proposed for complex detection, which is based on the premises where a protein complex has the core-attachment. The basic idea of COACH is such that it selects a sub-graph as a core, and then augments it.

All these methods are unsupervised learning methods and most of them use only a few basic topological features of the PPIN. These methods neither utilize the available known true complexes as a prior knowledge nor the available biological information of proteins (biological features). If the available known true complexes and biological features are used as a prior knowledge, the protein complex identification from a PPIN can be improved. However, some supervised learning techniques are proposed by different researchers [[11], [16], [24]], which also utilize some biological features. [24] introduced a supervised learning technique, known as probabilistic Bayesian Network model, for the detection of protein complexes. In their proposed model, biological and topological features of the PPIN have been utilized. [16] proposed another supervised learning technique, i.e., Regression model, which is used to help in filtering and growing cliques using topological features. [11] proposed the Random forest by using topological feature vector for prediction of protein complexes. Shi et al. [25] introduced a semi-supervised learning method based on neural network that uses biological and topological feature vectors. The DyCluster [26] method utilized the gene expression data for detection of a protein complex. A number of computational approaches along with their frameworks and feature sets are depicted in Table 1. Results from prior research exhibit that the use of supervised learning methods with biological and topological features are more effective to detect protein complexes than using unsupervised learning techniques with only topological features [27]. The above mentioned supervised methods achieved adequate accuracy rates for the identification of complexes, however, these rates require to be surpassed because the more accurate they are the more reliable they will be, and more likely to be used by biologists

TABLE 1. Overview, features, and frameworks of existing methods.

Name	Category	Topological Features	Biological Features	Framework
MCode [17]	Agglomerative	Clustering coefficient statistics, degree statistics, density statistics,	NA	Unsupervised
DPCLUS [18]	Agglomerative	Degree statistics	NA	Unsupervised
CFinder [19]	Clique finding	Clustering coefficient statistics, graph size, weighted density statistics	NA	Unsupervised
CMC [20]	Clique finding	Density statistics	NA	Unsupervised
ClusterONE [21]	Clique finding	Cohesiveness based on edge weights	NA	Unsupervised
MCL [22]	Graph clustering	Density statistics	NA	Unsupervised
COACH [23]	Core attachment	Degree statistics, density statistics, neighborhood affinity	NA	Unsupervised
Probabilistic Bayesian Network [24]	NA	Degree statistics, density statistics, edge weight statistics, node size, degree correlation statistic, clustering coefficient statistics, topological coefficient statistics, first eigenvalues, density w.r.t. weight cutoffs,	NA	Supervised
Regression model [16]	NA	**For unweighted network:** Degree statistics, graph density, clustering coefficient statistics **For weighted network:** Graph density, edge weight statistics, degree statistics, topological change	NA	Supervised
Random forest [11]	NA	Density statistics, distance statistics, degree statistics, vertex betweenness statistics, singular values, clustering coefficient statistics, edge betweenness statistics, weighted features	NA	Supervised
Feed-Forward Neural network [25]	NA	Graph density, edge weight statistics, degree statistics, clustering coefficient, topological coefficient, topological change	Protein length, polarity of amino acids	Semi-supervised
DyCluster [26]	NA	NA	Biclustering used on gene expression data	Supervised

and scientists. Hence, keeping in view the significance of supervised learning models and various topological and biological feature vectors for the detection of protein complexes, we propose the association rule learners– supervised learning models, i.e., PART and NNGE, to detect protein complexes from a PPIN of humans by incorporating the basic and advanced topological and biological feature set of proteins.

The remainder of this paper is organized in the following fashion: Section II describes the material and methods including the extraction of topological and biological features and proposed methodology. Section III presents the evaluation scores in terms of f-measure, recall, and precision for the proposed models in contrast to two state-of-the-art techniques, i.e., Bayesian network and Random forest, on the basis of 10-fold cross validation. Nevertheless, the evaluation scores indicate that the proposed methods outperform the mentioned two techniques.

II. MATERIAL AND METHODS

In PPIN, for the classification of protein complexes, rather depending only on one topological structure, we have considered numerous factors for the protein complex prediction task. It can be observed from Table 1 that most of the previous methods consider only basic topological properties, i.e., degree and density, and predicts a few sub-graph topologies such as clique, line, and spoke. Besides, other topological structures of protein complexes, i.e., ring, ring with clique, clique with star, clique with tail, star with ring, and star with tail, may also be found in the PPIN, as shown in figure4a and figure 4b. To predict these structures, it is evident to incorporate advance topological properties, for example, eccentricity, radiality, neighborhood connectivity, topological coefficient, and stress along with density, size, clustering coefficients, and degree statistics in order to achieve higher accuracy. Therefore, we have computed basic as well as advance topological properties of protein complexes for predicting complexes of varying sub-graph topologies. Apart from topological structure considerations, we also consider the biological structure formation of proteins. As the biological, physical, and chemical properties differentiate the complex from the noncomplex [24], the characterization of protein complexes from protein biological behaviors may be accurate enough. Keeping in view the effect of biological behaviors of proteins on complex formation, we compute

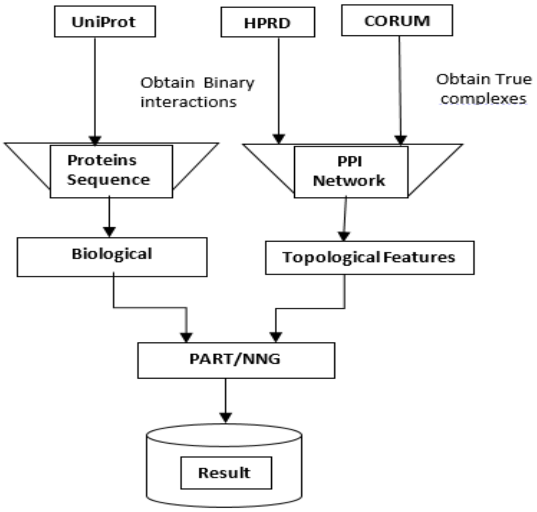


FIGURE 1. Flowchart of the proposed framework.

three types of biological features, i.e., DWT, length, and entropy. As the amino acid sequence is responsible for different biological behaviors in a complex, we utilize it to compute these biological features. Previously, only a few methods were used for biological properties to detect complexes, which achieved higher accuracy measures in contrast to those that only used topological properties of the PPIN. Inspired by these features, in this study, we compute and integrate advance biological and topological features, and test the proposed association rule-based learners, i.e., PART and NNGE, to classify human protein complexes in the IoMT environment. The obtained results reveal that the proposed framework outperforms Bayesnet and Random Forest techniques with respect to time complexity and accuracy. The proposed system is represented through a flowchart in Fig 1.

A. TOPOLOGICAL FEATURES

A protein complex has proteins where these proteins interact with each other, therefore, it can be represented as a graph. A protein itself can be characterized as a node where the interaction between two proteins can be characterized as an edge. As a protein complex is subject to a graph, we use topological structure properties as a feature set for a protein. We treat the protein complex as an undirected graph and compute different topological properties as a feature set. The computed topological features for the classification of protein complexes are i) average shortest path length, ii) topological coefficient, iii) neighborhood connectivity, iv) clustering coefficient, v) degree, vi) eccentricity, vii) closeness centrality, viii) radiality, ix) stress, and x) betweenness centrality [28].

B. BIOLOGICAL FEATURES

The biological features are extracted through amino acid sequence data of a gene. Amino acids play central roles and act as the structure blocks of genes. These features are computed because the mutation in an amino acid sequence may lead to a certain disease and genes causing a certain disease may have similar amino acid sequence structures. This feature set consists of discrete wavelet features in addition to length and entropy, which are used in [28]. Each of these features is described in the following subsections.

1) DISCRETE WAVELET TRANSFORM

The discrete wavelet features are computed via DWT. The DWT is used due to its interesting properties such as compact support, dilating relation, and vanishing moment. In brief, the compact support is the guaranty of localization of wavelets, which means that managing a data region using wavelets does not disturb the data out of that region. The vanishing moment is the guaranty of distinction between the important and non-important information, while processing wavelet and dilating relation guaranty the speedy wavelet algorithms [29]. It is the requirements of hierarchical representation and manipulation, localization, efficiency, and feature selection in different tasks in data mining, which have made wavelets a very powerful tool.

The DWT is a method that is used in digital signal processing for investigating digital signals in time and frequency domains. Since a protein amino acid sequence is an alphabetical sequence, this can be treated as a signal by converting it into numerical values. These values can be utilized to obtain a useful feature vector for the identification of protein complexes by applying the DWT. Therefore, to compute the discrete wavelet features, the frequency parameter is used for the numeric conversion of protein sequences on the basis of premises that proteins in a same complex have less or more similar sequence information. Thus, the frequency of each amino acid is counted from a protein amino acid sequence. A vector of length 20 is obtained for a single protein and then a DWT is applied on this vector. The DWT returns the detailed and approximation coefficient values for each protein. The detailed and approximation coefficient values are utilized as a feature set. Proteins belonging to the same complex got similar values for the detail and approximation coefficients, as compared to proteins belonging to other complexes. Previously, DWTs were used for solving different problems such as deoxyribonucleic acid (DNA) clustering [30], G-protein-coupled receptor classes prediction [31], and protein analysis. However, the DWTs have not been used in any study so far. Thus, we are the first to use it for the prediction of protein complexes.

2) LENGTH

The length is the frequency of each amino acid in a protein sequence, which is determined by counting the frequency of each amino acid in a protein.

3) ENTROPY

The entropy is computed as

$$E = -\sum_{i=1}^{20}(p_i \times log_2 p_i) \tag{1}$$

where p_i is the probability of amino acid in a sequence. The entropy calculates the information disorder and a protein belonging to the same complex may have less information disorder in comparison to proteins belonging to other complex classes. For this reason, the entropy is utilized in the proposed framework.

C. MODEL CONSTRUCTION

Association rule based learners have a potential power of prediction that execute efficiently on a number of features. Nonetheless, we have modeled protein complexes by using association rule learners, i.e., PART and NNGE. Firstly, topological features vector has been computed using the binary interaction data of proteins that participate in a complex. Secondly, biological features vector is examined by analyzing the proteins amino acid sequences. The computed topological and biological feature vectors are integrated and divided into train and test sets by applying 10-fold cross validation. The training set is utilized by PART and NNGE to generate rules for the prediction of protein complexes. The working of PART and NNGE is described in the following subsections.

a: PART

A variety of approaches has been investigated for inducing a rule set to make predictions. Among these, two dominant approaches are C4.5 [32] and RIPPER [33]. Both approaches generate a generalized set of rules for the classification purpose. In addition, both of them involve two steps for inducing the generalized set of rules. In the first step, they induce an preliminary rule set, while in the second step, these rules are accustomed or abandoned via a comprehensive elevating policy. For example, C4.5 produces an unpruned decision tree and then transmutes it into a set of rules. A rule is produced for each trail from parent-to-child node. Then, a rule-ranking strategy is used to simplify each rule separately. Finally, rules are deleted from the rule set till the rule set's error rate decreases on the training instances. The RIPPER works on separate-and-conquer strategy for rule generation. It generates only one rule at a time and discards the instances covered by this rule from the training set. It repeatedly derives new rules until no more instance is left in the training set. A state-of-the-art technique that combines the benefit of C4.5 and RIPPER is PART [34], which implements the separate-and-conquer scheme of RIPPER and combines it with C4.5– a decision tree approach to avoid the global optimization needed for rule generation. To create a rule, PART shapes a restricted decision tree on the specified set of occurrences. A rule is created from the built partial decision tree. The leaf with the largest coverage is induced as a rule and the created partial decision tree is discarded, which avoids global optimization. The instances are also removed from the training set that are covered by the induced rule. This process is repeated until no more instances left in the training set. The motivation behind proposing PART for the prediction of protein complexes is its simplicity. It does not require global optimization due to which its time complexity is reduced and the performance becomes efficient [34].

Moreover, it is evident from the literature [11], [28] that tree based approaches are gaining popularity for making predictions. Thus, if PART– a hybrid approach that induces a rule set on the basis of decision tree, is used, it may lead to better prediction results. However, in this study, it is shown that PART outperforms other existing approaches, i.e., Random Forest– a tree base approach, and Probablistic Baysein Network with respect to accuracy and time complexity.

b: NNGE

The NNGE is an instance based learner and hybrid approach that combines the idea of Nearest Neighbor classifier with the rule based classifier. It generates a set of generalized exemplars or hyper-rectangles. The generalized exemplars are a set of instances that can be interpreted as a rule for classification purposes. For generating a generalized example, it borrows the distance function element from the nearest neighbor and computes the similarity between the generalized example and an example from the training set. It is not necessary that the similarity between a hyper-rectangle and an example is alike, i.e., it could be partial depending upon a certain distance function. However, the distance function used by the NNGE to compute the similarity is given in equation 2 [35].

$$D(E,H) = W_h \sqrt{\sum_{j=1}^{n}(W_j \times \frac{E_j - H_j}{maxval_j - minval_j})^2} \tag{2}$$

In case of numerical attributes

$$(E_j - H_j) = \left\{ \begin{matrix} E_j - H_{upper} & E_j > H_{upper} \\ H_{lower} - E_j & E_j < H_{lower} \\ 0 & otherwise \end{matrix} \right\} \tag{3}$$

In case of nominal attributes

$$(E_j - H_j) = \left\{ \begin{matrix} 0 & E_j \in H_j \\ 1 & E_j! \in H_j \end{matrix} \right\} \tag{4}$$

E_j is the j^{th} feature value of example, H_j is the j^{th} feature value of hyper-rectangle, W_h is the weight of hyper-rectangle, and W_j is the weight of hyper-rectangle j^{th} feature. The $maxval_j$ and $minval_j$ are the upper and lower bounds for the j^{th} feature value of example, and H_{upper} and H_{lower} are the upper and lower bounds for the j^{th} feature value of hyper-rectangle. However, the utilized features weighting scheme is IB4.

The motivation behind proposing the NNGE for the prediction of protein complexes is that it is a simple and powerful predictor that can speed up the classification process [35]. Due to the model adjustment and generalization steps, it does not generate overlapped hyper-rectangles and therefore improves the accuracy and augments the classification process. Hence, it is proven through the produces results that NNGE outperforms the existing schemes (i.e., Random Forest and Probabalistic Bayesian Network) in terms of accuracy and time complexity.

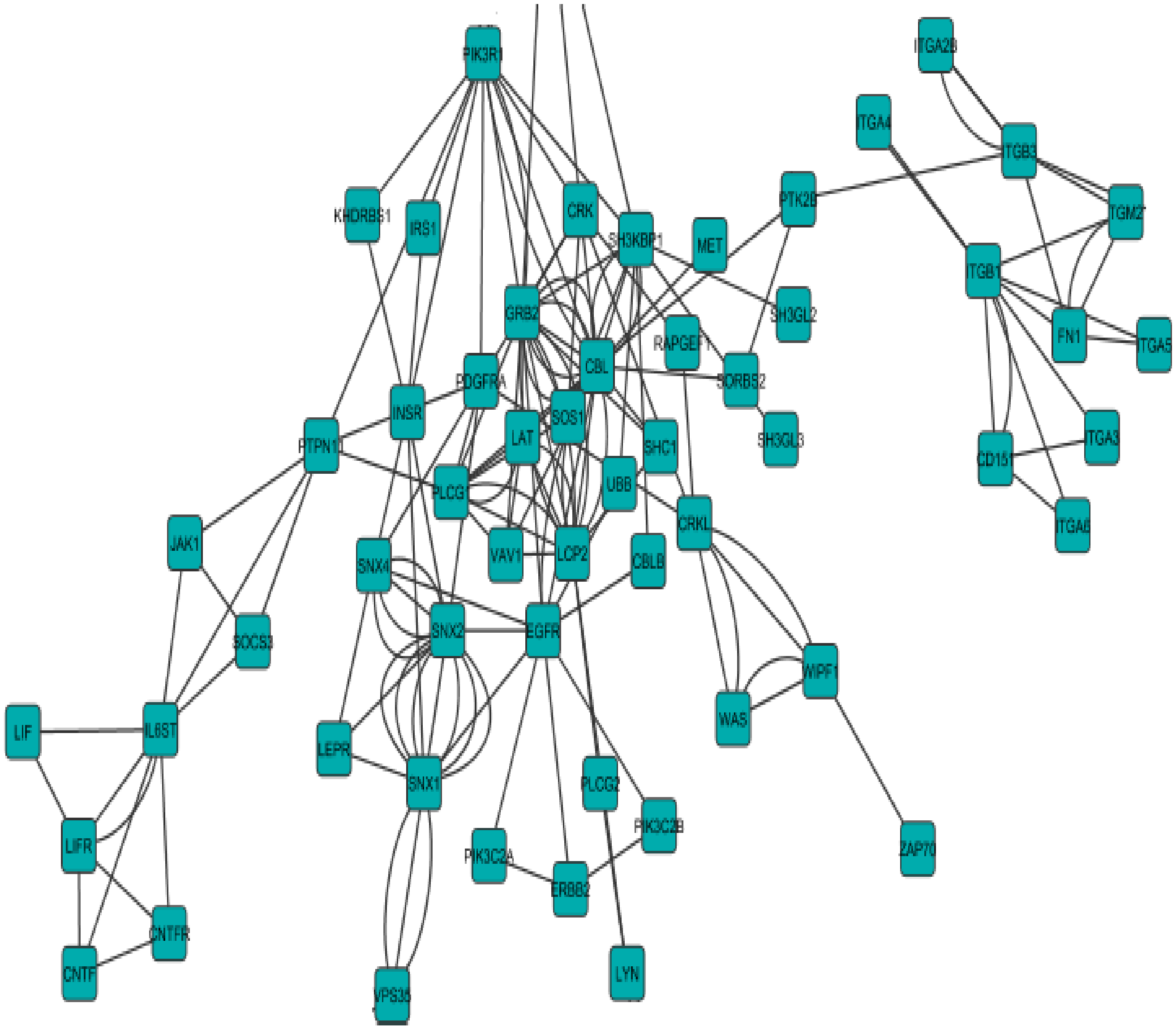


FIGURE 2. A few elements of PPIN of human protein complexes for the CORUM dataset.

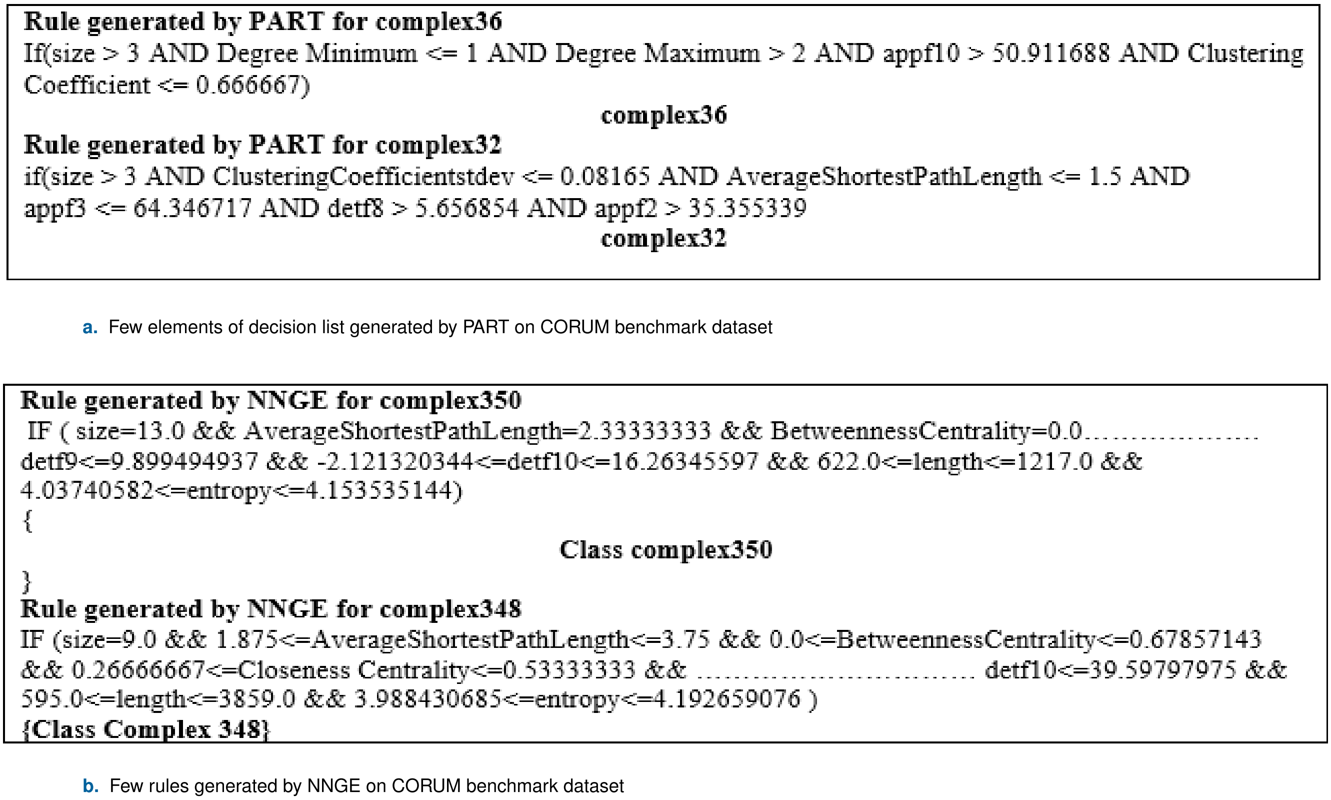
D. COMPLEX PREDICTION AND NEW SUB-GRAPH TOPOLOGIES

Some of the rules generated by PART and NNGE during the training phase on CORUM benchmark datasets are shown in Figure 3a and 3b, respectively. Different features are represented by variables, feature values are compared by comparison operators and combined using AND and OR connectors. For predicting a specific complex class for a given protein, during the test phase, each generated rule was checked one by one. A complex class was assigned to each feature vector of protein on the basis of satisfaction of conditions given in the rules. Statistics regarding the rules generated by each approach are given in Table 2.

TABLE 2. Association rule learner statistics.

NNGE		PART	
Hyper-rectangles	405	Number of rules	401
Exemplars	603		

From Table 2, it is clear that the total number of exemplars generated by NNGE are 603 including 405 hyper-rectangles while the number of rules generated by PART are 401. On the basis of these hyper-rectangles and rules generated by PART, each complex class is predicted. Some of the new hybrid complex topologies that were not covered previously, i.e., (a) Hybrid of Clique and Ring, (b) Clique with Tail, (c) Star with Tail, (d) Hybrid of Clique, Star, and Ring, (e) Hybrid of Ring and Star, and (f) Hybrid of clique and star, which match on a real complex set predicted by the association rule learners with the integrated feature vector (see Figure 4a). Figure 4b shows (a) Clique, (b) Star, (c) Linear, (d) Ring, and (e) Hybrid of clique and clique shape predicted complexes, which match with the real complex set.



Rule generated by PART for complex36

If(size > 3 AND Degree Minimum <= 1 AND Degree Maximum > 2 AND appf10 > 50.911688 AND Clustering Coefficient <= 0.666667)

complex36

Rule generated by PART for complex32

if(size > 3 AND ClusteringCoefficientstdev <= 0.08165 AND AverageShortestPathLength <= 1.5 AND appf3 <= 64.346717 AND detf8 > 5.656854 AND appf2 > 35.355339

complex32

a. Few elements of decision list generated by PART on CORUM benchmark dataset

Rule generated by NNGE for complex350

IF (size=13.0 && AverageShortestPathLength=2.33333333 && BetweennessCentrality=0.0.................... detf9<=9.899494937 && -2.121320344<=detf10<=16.26345597 && 622.0<=length<=1217.0 && 4.03740582<=entropy<=4.153535144)

{

Class complex350

}

Rule generated by NNGE for complex348

IF (size=9.0 && 1.875<=AverageShortestPathLength<=3.75 && 0.0<=BetweennessCentrality<=0.67857143 && 0.26666667<=Closeness Centrality<=0.53333333 && .... detf10<=39.59797975 && 595.0<=length<=3859.0 && 3.988430685<=entropy<=4.192659076)

{Class Complex 348}

b. Few rules generated by NNGE on CORUM benchmark dataset

FIGURE 3. a. Few elements of decision list generated by PART on CORUM benchmark dataset. b. Few rules generated by NNGE on CORUM benchmark dataset.

III. EXPERIMENTS AND RESULTS

A. REFERENCE DATASETS

True human protein complexes are provided by the CORUM database –Comprehensive Resource of Mammalian protein complexes [36], where as the binary protein interactions are provided by the Human Protein Reference Database (HPRD) [37]. The repeating and self-connected interactions are removed from proteins. After preprocessing, the final dataset contains 500 complexes with 2973 instances. The amino acid sequences of human proteins are taken from Uniprot [38]. The dataset statistics are provided in Table 3. Figure 2 presents the few elements of PPI network of human protein complexes constructed from HPRD binary protein interactions. The size of each complex in the final dataset is restricted to be equal to or more than three.

B. PERFORMANCE MEASURES

Generally, in the literature, to evaluate the model performance precision, recall and F-measure parameters are utilized. These parameters are also utilized in the proposed study to evaluate the model. Precision is a fraction of the predicted known protein complexes to all identified complexes. Whereas, recall is the fraction of predicted protein complexes to all known protein complexes, where the harmonic mean of precision and recall is F-measure [39]. Mathematically,

$$Precision = \frac{True\ positive}{False\ positive + True\ positive} \quad (5)$$

$$Recall = \frac{True\ Positive}{False\ Negative + True\ Positive} \quad (6)$$

$$F - measure = 2 \times \frac{Recall \times Precision}{Recall + Precision} \quad (7)$$

where True positive is the protein complexes predicted as complexes, False Positive is the non-protein complexes predicted as complexes, False Negative is protein complexes predicted as non-protein complexes [40]. In the statistical prediction, independent data set test, K-fold cross validation test, and jackknife cross-validation are usually used to assess the prediction capability of the model. As it is clear from equations 28-32 in [41] and revealed in a series of studies [42]–[45], that jackknife cross-validation is most effective one among the three methods, and can give in a unique result. But, to save computational time, the proposed models are evaluated by using the 10-fold cross-validation method, where in the 10-fold cross validation, the data set is divided into 10 equal subsets, each time nine subsets are used for training and one subset for testing.

C. RESULTS AND DISCUSSION

In order to show the effectiveness of rule based identification of protein complexes, it is compared with other state-of-the-art techniques in the HPRD network with the CORUM dataset. Table 4 shows the comprehensive comparison results with Probabilistic Bayesian Network and Random forest. This table also shows that the association

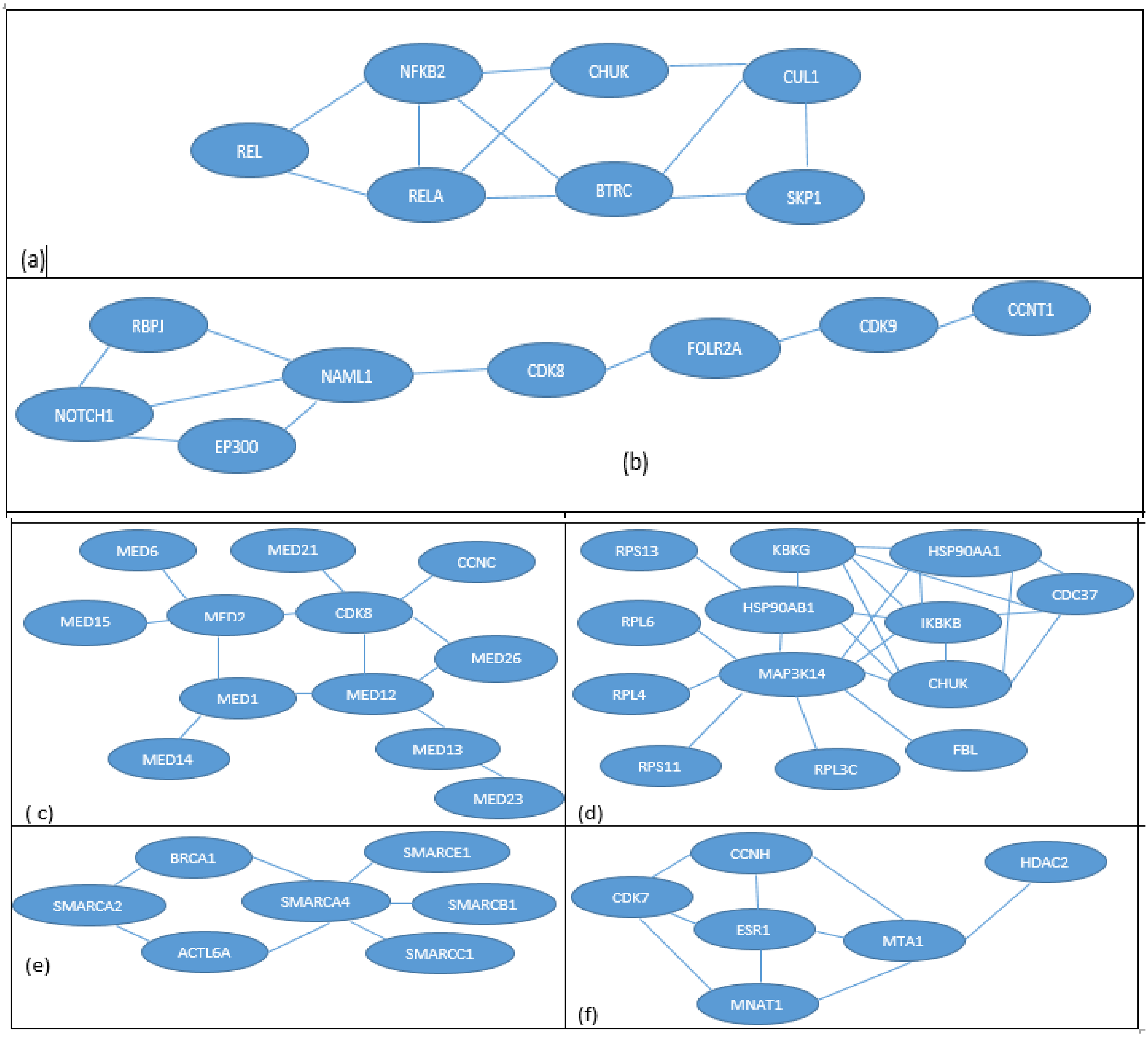


FIGURE 4. a. Example of different topological structures (a) Hybrid of Clique and Ring (b) Clique with Tail (c) Star with Tail (d) Hybrid of Clique, Star, and Ring (e) Hybrid of Ring and Star (f) Hybrid of clique and star.

rule based approaches, such as NNGE and PART, outperform Probabilistic Bayesian Network and Random forest. The PART gives the highest precision, recall and F-measure in contrast with all other methods. It shows that the association rule based approaches can achieve better performance than the Probabilistic Bayesian Network and Random forest.

1) ROBUSTNESS OF METHOD

In order to show the proposed method's robustness, the true positive rate (TP) of PART and NNGE is compared with those of the Random Forest and Probabilistic Bayesian Network on the basis of different threshold (i.e., $t = 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8$ values), as shown in Figure 5. It is elucidated from Figure 5 that at $t = 0.2$, 0.4, and 0.7, the PART outperforms the other methods, while at $t = 0.3$ and 0.8, the NNGE surpasses the other methods. Overall, the association rule based learners excel Random Forest and Probabilistic Bayesian Network.

2) INFLUENCE OF BIOLOGICAL FEATURES

To measure the effectiveness of biological features in improving the performance of different methods, we have conducted the same experiments by using the Probabilistic Bayesian

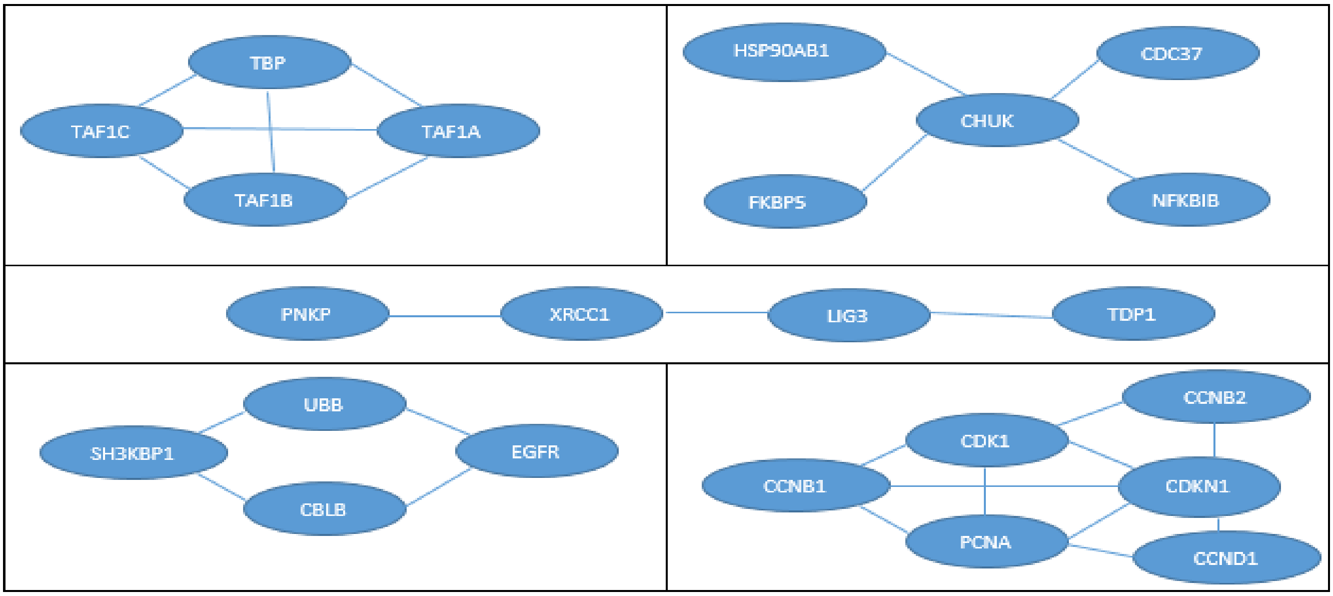


FIGURE 4. *(Continued).* **b. Example of different topological structures (a) Clique (b) Star (c) Linear (d) Ring (e) Hybrid of clique and clique.**

TABLE 3. Dataset statistics.

Hybrid of Clique and Ring	79
Clique with Tail	5
Star with Tail	15
Hybrid of Clique, Star and Ring	25
Hybrid of Ring and Star	36
Hybrid of clique and star	3
Linear	122
Star	14
Ring	96
Clique	39
Hybrid of clique and clique	4
Total No. of complexes	1925
No. of complexes with no interactions found in HPRD	1205
No. of complexes with more than three interactions	500
No. of Instances	2973
No. of Interactions	4395
No. of Proteins	2524
No. of Sequences	2524

TABLE 4. Comparison on CORUM benchmark dataset.

Methods	CORUM (Complexes 500)		
	Precision	Recall	Fmeasure
Probablistic Baysein Network	0.45	0.46	0.46
Random Forest	0.50	0.50	0.50
NNGE	0.47	0.47	0.47
PART	0.52	0.53	0.51

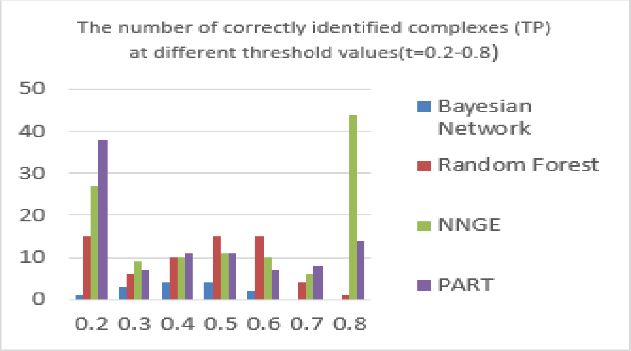


FIGURE 5. Robustness of proposed method.

Network, Random forest, NNGE, and PART on a benchmark dataset in two scenarios, i.e., (i) only considering topological features, and (ii) integrating topological and biological features of complexes. Table 5 shows precision, recall, and F-measure rates of the Probabilistic Bayesian Network, Random Forest, NNGE, and PART on CORUM dataset by using only topological features, whereas Table 6 exhibits precision, recall, and F-measure rates of the Probabilistic Bayesian Network, Random Forest, NNGE, and PART on CORUM dataset using integrated features.

The results indicate that in case of Probabilistic Bayesian Network and NNGE, using only topological features and then integrated features, no difference was observed in the precision, recall, and F-measure. While in case of Random forest and PART, topological features alone gave low precision, recall, and F-measure rates. However, when topological features were combined with biological features, a significant difference was observed. By the addition of more features, computational cost was increased on one hand but the number of correctly identified protein complexes was also increased, as shown in Table 5(b). However, the correct identification is more critical for different applications of complex detection like in the diagnosis of different diseases. Thus, the integrated feature vector is important and useful in increasing the performance efficiency. It is evident from Table 5(b) that PART outperforms the existing prediction methods.

TABLE 5. a. Precision, Recall, and F-measure by considering only topological features. b. Precision, Recall, and F-measure by considering biological and topological features.

a. Precision, Recall, and F-measure by considering only topological features.

	CORUM Topological Features		
Methods	**Precision**	**Recall**	**F-measure**
Probabilistic Bayesian Network	0.45	0.46	0.46
Random Forest	0.45	0.45	0.45
NNGE	0.47	0.47	0.47
PART	0.47	0.48	0.47

b. Precision, Recall, and F-measure by considering biological and topological features.

	CORUM Integrated Features (Topological, Biological)		
Methods	**Precision**	**Recall**	**F-measure**
Probabilistic Bayesian Network	0.45	0.46	0.46
Random Forest	0.50	0.50	0.50
NNGE	0.47	0.47	0.47
PART	0.52	0.53	0.51

3) COMPUTATIONAL COST

The computational cost of proposed methods is also analyzed mathematically and empirically, which give comparative results with regards to Random Forest on the CORUM Benchmark dataset. Mathematically, the computational cost with respect to Big(O) for Random Forest is O(mtree × ntree × nlogn), where mtree is the number of trees and ntree is the number of attributes that have to be sampled at each node. Similarly, O(a × nlogn) is a time complexity for PART, where a is the number of attributes and n is the number of instances. Moreover, for NNGE, the time complexity is O(d × n), where n is the number of instances and d is the time to compute distance or similarity between an instance and a hyper-rectangle. Analytically, it is proven that NNGE and PART have the lowest time complexity as compared to Random Forest.

Furthermore, the empirical analysis also shows that NNGE has the lowest computational cost with respect to execution time, as exhibited in Table 7. The machine specifications are also provided in Table 6 on which the empirical analysis has done.

The results indicate that NNGE achieved better performance than Random Forest and Probabilistic Bayesian Network in computational cost for predicting protein complexes, whereas PART has more computational cost than Random Forest. However, it is evident from the analysis that as the sample size increases, the time complexity of Random Forest may increase due to the number of instances, number of trees, and number of attributes. Where the time complexity of PART may decrease because of the number of instances and number of attributes. Secondly, Random Forest needs more memory or heap size to perform its computations as compared to PART and NNGE.

TABLE 6. Empirical comparison of computational cost for random forest, NNGE and Part on CORUM benchmark dataset with integrated features.

Methods	**Correctly Classified Instances**	**InCorrectly Classified Instances**	**Computa-tional Cost(sec)**	**Machine Specifications**
NNGE	1389	1584	9.72	Processor: intel® core$^{TM}i3-$ 4005u CPU@ 1.7GHz RAM: 4.00GB System Type: 64 bit
Random Forest	1492	1481	23.72	
PART	1581	1392	29.65	

Hence, it is concluded that NNGE and PART outperform Random Forest in terms of computational requirements such as space and time complexity. While the NNGE and PART excel Random Forest in terms of time complexity and accuracy, respectively, as exhibited in Table 6. Therefore, for the identification of protein complexes in the PPIN, association rule learners can achieve better results with respect to accuracy and computational cost.

4) THE EFFECT OF DIFFERENT FEATURES ON ACCURACY AND COMPUTATIONAL COST

The same experiments were further applied for the identification of complexes by dividing the features set into two groups, i.e., baseline and advanced feature sets, in order to check different features' effects on accuracy and computational cost. The baseline features incorporate those features that are commonly used in the literature for the detection of complexes, while the advanced feature set encapsulates those features that were introduced in this study. The baseline features comprise size, betweenness centrality, average shortest path length, closeness centrality, clustering coefficient, density, and degree. The advanced feature set include eccentricity, neighborhood connectivity, radiality, stress, topological coefficient, discrete wavelet coefficients, length, and entropy. The Random Forest, PART, and NNGE were compared, where PART and NNGE achieved better results than the Random Forest in terms of accuracy and time complexity, as presented in Table 8.

Table 7 presents the overall results of baseline features, advanced features, and the integration of both baseline and advanced, where the integrated features gave better prediction performance.

After performing the same experiment with different angles, it is concluded that NNGE surpasses other schemes with regard to computational cost. In addition, PART achieved superior accuracy as compared to other schemes.

TABLE 7. The effect of different features on accuracy and computational cost using CORUM benchmark dataset.

Methods	CORUM Baseline Features			CORUM Advanced Features			CORUM Baseline +Advance Feature		
	Correctly Classified Instances(TP)	Incorrectly Classified Instances(FP)	Computational Cost(sec)	Correctly Classified Instances(TP)	Incorrectly Classified Instances(FP)	Computational Cost	Correctly Classified Instances(TP)	Incorrectly Classified Instances(FP)	Computational Cost
Random Forest	1342	1631	9.45	456	2517	24.04	1492	1481	23.72
NNGE	1445	1528	3.48	414	2559	5.88	1389	1584	9.72
PART	1462	1511	2.78	492	2481	48.78	1581	1392	29.65

IV. CONCLUSION AND FUTURE WORK

To examine the principal mechanism of various cellular functions and to elucidate the functionality of different un-annotated proteins, the correct prediction of protein complexes plays an important role. Several complex detection algorithms have been proposed, which utilize only basic topological properties of protein rather than using advance topological and biological properties. The association rule learners such as PART and NNGE were proposed with advance topological and biological feature sets to detect complexes from PPIN. The results indicate that the association rule learners can supersede the existing schemes for achieving better accuracy and low computational cost.

In the future, some other computational methods can be developed that can help incorporating the benefits of various computational methods such as combining Bayes with rules. Moreover, some other important biological properties, e.g., ionization, polarization or hydrophobicity of proteins, and topological properties, e.g., weighted features, may be integrated in the feature set to achieve a remarkable accuracy.

REFERENCES

[1] I. Ud Din, M. Guizani, S. Hassan, B.-S. Kim, M. K. Khan, M. Atiquzzaman, and S. H. Ahmed, "The Internet of Things: A review of enabled technologies and future challenges," *IEEE Access*, vol. 7, pp. 7606–7640, 2019.

[2] 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.

[3] K. A. Awan, I. U. Din, M. Zareei, M. Talha, M. Guizani, and S. U. Jadoon, "HoliTrust—A holistic cross–domain trust management mechanism for service–centric Internet of Things," *IEEE Access*, vol. 7, pp. 52191–52201, 2019.

[4] I. U. Din, M. Guizani, B.-S. Kim, S. Hassan, and M. K. Khan, "Trust management techniques for the Internet of Things: A survey," *IEEE Access*, vol. 7, pp. 29763–29787, 2019.

[5] I. U. Din, M. Guizani, J. J. Rodrigues, S. Hassan, and V. V. Korotaev, "Machine learning in the Internet of Things: Designed techniques for smart cities," *Future Gener. Comput. Syst.*, vol. 100, pp. 826–843, Nov. 2019.

[6] S. Khan, N. Islam, Z. Jan, I. Ud Din, and J. J. P. C. Rodrigues, "A novel deep learning based framework for the detection and classification of breast cancer using transfer learning," *Pattern Recognit. Lett.*, vol. 125, pp. 1–6, Jul. 2019.

[7] S. U. Khan, N. Islam, Z. Jan, I. U. Din, A. Khan, and Y. Faheem, "An e-Health care services framework for the detection and classification of breast cancer in breast cytology images as an IoMT application," *Future Gener. Comput. Syst.*, vol. 98, pp. 286–296, Sep. 2019.

[8] N. Islam, Y. Faheem, I. U. Din, M. Talha, M. Guizani, and M. Khalil, "A blockchain-based fog computing framework for activity recognition as an application to e-Healthcare services," *Future Gener. Comput. Syst.*, vol. 100, pp. 569–578, Nov. 2019.

[9] K. A. Awan, I. U. Din, A. Almogren, M. Guizani, A. Altameem, and S. U. Jadoon, "RobustTrust—A pro–privacy robust distributed trust management mechanism for Internet of Things," *IEEE Access*, vol. 7, pp. 62095–62106, 2019.

[10] M. Banerjee, J. Lee, and K.-K.-R. Choo, "A blockchain future for Internet of Things security: A position paper," *Digit. Commun. Netw.*, vol. 4, no. 3, pp. 149–160, Aug. 2018.

[11] Z.-C. Li, Y.-H. Lai, L.-L. Chen, X. Zhou, Z. Dai, and X.-Y. Zou, "Identification of human protein complexes from local sub-graphs of protein–protein interaction network based on random forest with topological structure features," *Analytica Chim. Acta*, vol. 718, pp. 32–41, Mar. 2012.

[12] X. Li, M. Wu, C.-K. Kwoh, and S.-K. Ng, "Computational approaches for detecting protein complexes from protein interaction networks: A survey," *BMC Genomics*, vol. 11, no. 1, p. S3, 2010.

[13] S. Pitre, M. Alamgir, J. R. Green, M. Dumontier, F. Dehne, and A. Golshani, "Computational methods for predicting protein–protein interactions," in *Protein–Protein Interaction*. Berlin, Germany: Springer, 2008, pp. 247–267.

[14] G. Rigaut, A. Shevchenko, B. Rutz, M. Wilm, M. Mann, and B. Séraphin, "A generic protein purification method for protein complex characterization and proteome exploration," *Nature Biotechnol.*, vol. 17, no. 10, pp. 1030–1032, Oct. 1999.

[15] K. Tarassov, V. Messier, C. R. Landry, S. Radinovic, M. M. S. Molina, I. Shames, Y. Malitskaya, J. Vogel, H. Bussey, and S. W. Michnick, "An *in vivo* map of the yeast protein interactome," *Science*, vol. 320, no. 5882, pp. 1465–1470, Jun. 2008.

[16] F. Y. Yu, Z. H. Yang, N. Tang, H. F. Lin, J. Wang, and Z. W. Yang, "Predicting protein complex in protein interaction network-a supervised learning based method," *BMC Syst. Biol.*, vol. 8, no. 3, p. S4, 2014.

[17] G. D. Bader and C. W. V. Hogue, "An automated method for finding molecular complexes in large protein interaction networks," *BMC Bioinf.*, vol. 4, no. 1, p. 2, 2003.

[18] M. Altaf-Ul-Amin, Y. Shinbo, K. Mihara, K. Kurokawa, and S. Kanaya, "Development and implementation of an algorithm for detection of protein complexes in large interaction networks," *BMC Bioinf.*, vol. 7, no. 1, p. 207, 2006.

[19] B. Adamcsek, G. Palla, I. J. Farkas, I. Derényi, and T. Vicsek, "CFinder: Locating cliques and overlapping modules in biological networks," *Bioinformatics*, vol. 22, no. 8, pp. 1021–1023, Apr. 2006.

[20] G. Liu, L. Wong, and H. N. Chua, "Complex discovery from weighted PPI networks," *Bioinformatics*, vol. 25, no. 15, pp. 1891–1897, Aug. 2009.

[21] T. Nepusz, H. Yu, and A. Paccanaro, "Detecting overlapping protein complexes in protein-protein interaction networks," *Nature Methods*, vol. 9, no. 5, pp. 471–472, May 2012.

[22] J. B. Pereira-Leal, A. J. Enright, and C. A. Ouzounis, "Detection of functional modules from protein interaction networks," *Proteins*, vol. 54, no. 1, pp. 49–57, Dec. 2003.

[23] M. Wu, X. Li, C.-K. Kwoh, and S.-K. Ng, "A core-attachment based method to detect protein complexes in PPI networks," *BMC Bioinf.*, vol. 10, no. 1, p. 169, 2009.

[24] Y. Qi, F. Balem, C. Faloutsos, J. Klein-Seetharaman, and Z. Bar-Joseph, "Protein complex identification by supervised graph local clustering," *Bioinformatics*, vol. 24, no. 13, pp. i250–i268, Jul. 2008.

[25] L. Shi, X. Lei, and A. Zhang, "Protein complex detection with semi-supervised learning in protein interaction networks," *Proteome SciProteome Sci.*, vol. 9, no. 1, p. S5, 2011.

[26] E. M. Hanna, N. Zaki, and A. Amin, "Detecting protein complexes in protein interaction networks modeled as gene expression biclusters," *PLoS ONE*, vol. 10, no. 12, Dec. 2015, Art. no. e0144163.

[27] L. Chen, X. Shi, X. Kong, Z. Zeng, and Y.-D. Cai, "Identifying protein complexes using hybrid properties," *J. Proteome Res.*, vol. 8, no. 11, pp. 5212–5218, Nov. 2009.

[28] 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.

[29] 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.

[30] J. Bao and R. Yuan, "A wavelet-based feature vector model for DNA clustering," *Genet. Mol. Res.*, vol. 14, no. 4, pp. 19163–19172, Jan. 2016.

[31] J.-D. Qiu, J.-H. Huang, R.-P. Liang, and X.-Q. Lu, "Prediction of G-protein-coupled receptor classes based on the concept of Chou's pseudo amino acid composition: An approach from discrete wavelet transform," *Anal. Biochem.*, vol. 390, no. 1, pp. 68–73, Jul. 2009.

[32] S. L. Salzberg, "C4.5: Programs for machine learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993," *Mach. Learn.*, vol. 16, no. 3, pp. 235–240, Sep. 1994.

[33] W. W. Cohen, "Fast effective rule induction," in *Machine Learning Proceedings 1995*. Amsterdam, The Netherlands: Elsevier, 1995, pp. 115–123.

[34] E. Frank and I. H. Witten, "Generating accurate rule sets without global optimization," in *Proc. 15th Int. Conf. Mach. Learn., Morgan Kaufmann Publishers*, San Francisco, CA, USA, 1998, pp. 144–151.

[35] B. Martin, "Instance-based learning: Nearest neighbour with generalization," Dept. Comput. Sci., Univ. Waikato, Hamilton, New Zealand, Tech. Rep. No.95/18, 1995.

[36] CORUM. *Comprehensive Resource of Mammalians Protein Complex*. Accessed: Feb. 21, 2015. [Online]. Available: http://mips.helmholtzmuenchen.de/CORUM/

[37] HPRD. (2015). *Human Protein Reference Database*. Accessed: Feb. 21, 2015. [Online]. Available: http://www.hprd.org/

[38] (2015). *UniProt*. Accessed: Feb. 21, 2015. [Online]. Available: https://www.uniprot.org/

[39] W. Peng, J. Wang, B. Zhao, and L. Wang, "Identification of protein complexes using weighted pagerank–nibble algorithm and core–attachment structure," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 12, no. 1, pp. 179–192, Jan. 2015.

[40] M. M. Abbas, M. M. Mohie-Eldin, and Y. El-Manzalawy, "Assessing the effects of data selection and representation on the development of reliable E. Coli sigma 70 promoter region predictors," *PLoS ONE*, vol. 10, no. 3, Mar. 2015, Art. no. e0119721.

[41] K.-C. Chou, "Some remarks on protein attribute prediction and pseudo amino acid composition," *J. Theor. Biol.*, vol. 273, no. 1, pp. 236–247, Mar. 2011.

[42] W. Chen, P. Feng, H. Ding, H. Lin, and K.-C. Chou, "IRNA-Methyl: Identifying N6-methyladenosine sites using pseudo nucleotide composition," *Anal. Biochem.*, vol. 490, pp. 26–33, Dec. 2015.

[43] W. Chen, H. Ding, P. Feng, H. Lin, and K.-C. Chou, "iACP: A sequence-based tool for identifying anticancer peptides," *Oncotarget*, vol. 7, no. 13, p. 16895, 2016.

[44] W. Chen, H. Yang, P. Feng, H. Ding, and H. Lin, "IDNA4mC: Identifying DNA N^4-methylcytosine sites based on nucleotide chemical properties," *Bioinformatics*, vol. 33, no. 22, pp. 3518–3523, Nov. 2017.

[45] P.-M. Feng, H. Ding, W. Chen, and H. Lin, "Naïve Bayes classifier with feature selection to identify phage virion proteins," *Comput. Math. Methods Med.*, vol. 2013, pp. 1–6, Apr. 2013.

MISBA SIKARNDAR received the B.S. degree in computer science from International Islamic University at Islamabad, Pakistan, in 2009, and the M.S. degree in computer science from COMSATS, in 2016. She is currently pursuing the Ph.D. degree from the Department of Information Technology, University of Haripur, Pakistan. She has been a Lecturer of computer science with Haripur University, Khyber Pakhtunkhwa, Pakistan, since 2015. Her research interests include natural language processing, bioinformatics, and the Internet of Things.

WAQAS ANWAR received the master's degree in computer science from Hamdard University, Pakistan, in 2001, and the Ph.D. degree in computer application technology from the Harbin Institute of Technology, China, in 2008. He has been an Associate Professor with the COMSATS Institute of Information Technology, Pakistan, since 2008. He is currently an Active Researcher. His areas of interest are natural language processing, computational intelligence, and bioinformatics.

AHMAD ALMOGREN (Senior Member, IEEE) received the Ph.D. degree in computer science from Southern Methodist University, Dallas, TX, USA, in 2002. Previously, he was an Assistant Professor of computer science and a member of the scientific council, Riyadh College of Technology. He also served as the Dean of the College of Computer and Information Sciences, and the Head of the Council of Academic, Al Yamamah University. He is currently a Professor and the Vice Dean of the development and quality with the College of Computer and Information Sciences, King Saud University. His research areas of interest include mobile and pervasive computing, cyber security, and computer networks. He has served as a Guest Editor for several computer journals.

IKRAM UD DIN (Senior Member, IEEE) received the M.Sc. degree in computer science and the M.S. degree in computer networking from the Department of Computer Science, University of Peshawar, Pakistan, and the Ph.D. degree in computer science from the School of Computing, Universiti Utara Malaysia (UUM). He has served as the IEEE UUM Student Branch Professional Chair. He is currently working as a Lecturer with the Department of Information Technology, University of Haripur. He has 12 years of teaching and research experience in different universities/organizations. His current research interests include resource management and traffic control in wired and wireless networks, vehicular communications, mobility and cache management in information-centric networking, and the Internet of Things.

NADRA GUIZANI is currently pursuing the Ph.D. degree Purdue University, where completing a thesis in prediction and access control of disease spread data on dynamic network topologies. She is also a Graduate Lecturer with Purdue University Her research interests include machine learning, mobile networking, large data analysis, and prediction techniques. She is an Active Member with the Women in Engineering Program and the Computing Research Association for Women.