IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 29 September 2023, accepted 8 October 2023, date of publication 16 October 2023, date of current version 25 October 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3324985

TOPICAL REVIEW

A Review of Random Walk-Based Method for the Identification of Disease Genes and Disease Modules

TAY XIN HUI[1], **SHAHREEN KASIM**[1], **MOHD FARHAN MD. FUDZEE**[1], **(Senior Member, IEEE)**, **TOLE SUTIKNO**[2], **(Member, IEEE)**, **ROHAYANTI HASSAN**[3], **IZZATDIN ABDUL AZIZ**[4], **MOHD HILMI HASAN**[4], **JAFREEZAL JAAFAR**[4], **(Senior Member, IEEE)**, **METAB ALHARBI**[5], **AND SEAH CHOON SEN**[6]

[1]Faculty of Computer Sciences and Information Technology, Universiti Tun Hussein Onn Malaysia (UTHM), Parit Raja 86400, Malaysia
[2]Department of Electrical Engineering, Universitas Ahmad Dahlan (UAD), Yogyakarta 55166, Indonesia
[3]Faculty of Electrical Engineering, Universiti Teknologi Malaysia (UTM), Skudai 81310, Malaysia
[4]Computer and Information Sciences Department (CISD), Universiti Teknologi PETRONAS (UTP), Seri Iskandar, Perak 32610, Malaysia
[5]Department of Pharmacology and Toxicology, College of Pharmacy, King Saud University, Riyadh 11451, Saudi Arabia
[6]Faculty of Accounting and Management, Universiti Tunku Abdul Rahman (UTAR), Kajang 43000, Malaysia

Corresponding author: Shahreen Kasim (shahreen@uthm.edu.my)

This work was supported in part by the Fundamental Research Grant Scheme from the Ministry of Higher Education, Tun Hussein Onn University of Malaysia (UTHM), under Grant REGG FASA 1/2021 (VOT NO. H888).

ABSTRACT Traditional techniques for identifying disease genes and disease modules involve high-cost clinical experiments and unpredictable time consumption for analysis. Network-based computational approaches usually focus on the systematic study of molecular networks to predict the associations between diseases and genes. The random walk-based method is a network-based approach that utilises biological networks for analysis. As the random walk models efficiently capture the complex interplay among molecules in diseases, it is extensively applied in biological problem-solving based on networks. Despite their comprehensive employment, the fundamentals of random walk and overall background may not be fully understood, leading to misinterpretation of results. This review aims to cover the fundamental knowledge of random walk models for biological network analysis. This study reviewed diffusion-based random walk methods for disease gene prediction and disease module identification. The random walk-based disease gene prediction methods are categorised into node classification and link prediction tasks. This study details the advantages and limitations of each method. Finally, the potential challenges and research directions for future studies on random walk models are highlighted.

INDEX TERMS Random walk, disease gene prediction, disease module identification, disease-gene prioritisation, biological network.

I. INTRODUCTION

Genetic diseases are caused by gene mutations in combination with epigenetic factors or by a chromosomal abnormality [1], [2]. Genetic disorders are a result of improper protein production. The disorders can be divided into three categories, namely single-gene disorders (mutations in a single gene), complex disorders (mutations in two or more genes), and chromosomal disorders (changes in the number or structure of the chromosomes). Among the factors that affect the diagnosis of genetic disorders include variability in the phenotypic characteristics, overlapping symptoms with other disorders etc. [2]. Elucidating the relationship between human genetic diseases and their causal genes (or proteins) remains a major public issue [3].

Although traditional techniques for disease gene prediction and disease module identification provide predictive biomarkers and protein complexes through genetic variation

The associate editor coordinating the review of this manuscript and approving it for publication was Vincenzo Conti.

studies, these methods are expensive and time- and resource-consuming, as many false positives need to be analysed further [4], [5]. Moreover, traditional techniques focus on direct association between genes and diseases as well as associations between diseases and protein complexes are not cost-effective. Based on extant literature, biological molecules (genes or proteins) collaboratively perform their functions [6], [7], [8], [9]. Therefore, computational modelling techniques could be more efficient in understanding system-level diseases.

Computational modelling of biological systems uses networks to understand their structure and dynamics [10]. More helpful information may be revealed and systematic aspects can be gained by designing and defining their specific roles and collaboration to a wired network graph structure [11]. A network-based environment enables efficient tracking of disease-causing factors by trailing network perturbations (e.g. edge or node removals) in the molecular networks [11]. Thus, molecular networks like protein-protein interaction (PPI) networks, gene co-expression (GCE) networks, gene regulatory networks (GRN), and Bayesian networks are efficient and effective for complex data visualisation and interpretation. Such complex modelling interplay is represented by nodes as molecules (e.g. genes, RNA, proteins and metabolites) and edges as relationships between the nodes (e.g. regulatory relationship) [12].

The random walk model is a network-based approach that employs graph-theoretical algorithms to solve biological problems, including disease gene prediction, protein function annotation, and disease module detection. This diffusion-based method uses information encoded in the complete network topology and the placement of all known disease genes for influence propagation in different networks through symmetric diffusion. Whereby information flow diffuses through each edge to other nodes in the network. The node weights following the diffusion represent their affinity or closeness to other highly weighted nodes [13]. The random walk model is a useful tool to study the structure of graphs and the relationship between nodes [14]. The underlying assumption of random walk-based methods is that phenotypically similar diseases are caused by functionally related genes that are located close to each other in the molecular networks [15], [16], [17].

Diffusion-based random walk methods have been increasingly enhanced by considering prior information from omics data sources or topological information to calculate the network's node weights or adjacency matrix. For instance, Prioritisation with a Warped Network (PWN) [13] was designed as an enhanced random walk-based method that incorporates both network properties and prior knowledge to quantify the proximity between genes in the network. Hence, it is extensively used to complement and enrich existing statistical analyses to solve biological problems.

There are several reviews [2], [15], [18], [19], [20], [21] and benchmark [11], [17], [22] articles published on network-based methods. Most of these published works covered an overview of the existing network embedding methods, ranging from machine learning to graph representation learning methods. However, only a limited number of studies focused on presenting a wide range of diffusion-based random walk methods for disease gene prediction and disease module identification. Moreover, some articles [23], [24], [25] that extensively discussed the random walk models mainly focused on their theoretical definitions and underlying mathematical concepts. Some former surveys [14], [26] reviewed the application of random walk models in solving different biological problems. These articles either covered limited tools or did not assess various available state-of-the-art network diffusion random walk methods.

To address the abovementioned issues, this study aims to present a comprehensive review of diffusion-based random walk methods leveraging network or graph data for disease gene prediction and disease module identification. A high-level illustration of the pipeline for applying diffusion-based random walk methods to different biomedical tasks is provided. The general concepts and principles of the random walk approaches are introduced, and a classification scheme of computational approaches based on problem definition (i.e. node classification and link prediction) for disease gene prediction is discussed. A list of available random walk-based methods for disease module identification is also provided. The capabilities and limitations of the random walk-based methods were acknowledged to deliver a fast and clear initiation in using these promising research tools. Finally, challenges and future directions for the methodological development and applications of random walk-based methods were described.

II. FUNDAMENTALS OF RANDOM WALK

The basic concept of a random walk on a graph begins with a single or group of nodes that visits each node by taking serial random walking steps. For every moving step, the nodes move to a random neighbour, where a distribution value is calculated for every node in the graph, indicating the likelihood that a walker is present at that node at that particular step. The random walk process is repeated until all nodes in the graphare covered or converged. Finally, the distribution value of each node remains constant and is proportional to the time a random walker travels to that node and the distance from the starting nodes [14].

A biological network can be represented as an undirected (e.g. PPI networks, GCE networks) or directed (e.g. GRN, metabolic networks) graph. Given a graph G = (V, E), V is a set of nodes and E is a set of edges. For any node u∈V and (u, v)∈E, $B(u)$ is the set of all nodes that links to node u and $|L(v)|$ is the number of neighbours (outgoing links) of node v. $PR^{t+1}(u)$ is the probability (rank scores) of a particular node u at time step $t+1$, and $PR^{t}(v)$ is the probability of node v at time step t. The node proximity on a graph can be calculated

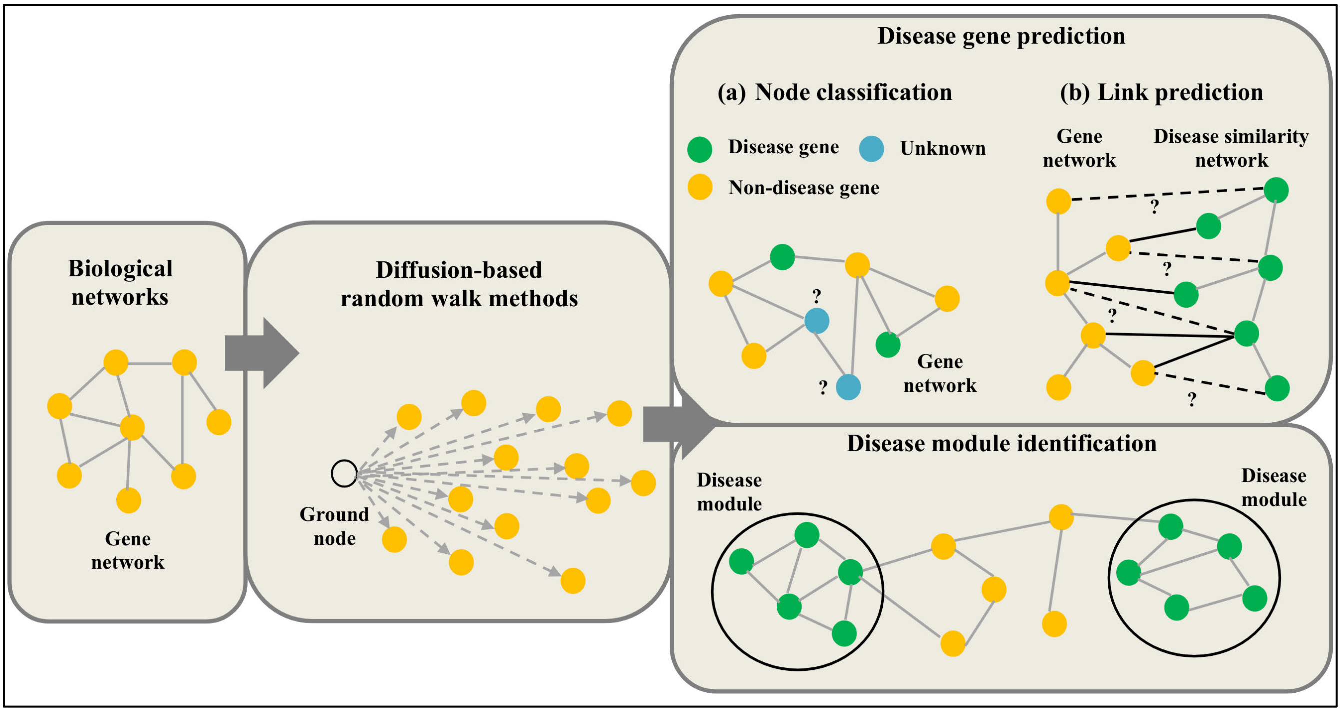

FIGURE 1. Pipeline for applying diffusion-based random walk methods to biomedical tasks.

based on a simple random walk model and some extended random walk models defined as follows.

A. PageRank ALGORITHM

PageRank is an algorithm developed to rank the importance of webpages by employing the link structure of the web [27]. A Markov chain with a primitive transition probability matrix can be built using the hyperlink structure of the web. The stationary vector or PageRank vector is obtained based on the irreducibility of the Markov chain. The values corresponding to each page in the PageRank vector are known as the PageRank scores of the page [28]. This algorithm indicates that a page with important incoming links will produce outgoing links to other pages that are also essential [28]. Thus, PageRank considers the backlinks and propagates the ranking through links: a page ranks high if the sum of the ranks of its backlinks is high [27]. A simplified version of PageRank is defined as [27] and [29]:

$$PR^{t+1}(u) = c\sum\nolimits_{v\in B(u)} \frac{PR^{t}(v)}{|L(v)|} \tag{1}$$

where u represents a node (web page). $B(u)$ is the set of nodes (pages) that point to node u and c is a factor used for normalisation. The overall PageRank score is calculated based on ranking all network nodes. However, PageRank has a rank sink problem whereby not all users follow the direct link within a website [27]. In this case, the original PageRank is modified as follows [27], [29]:

$$PR^{t+1}(u) = \frac{(1-\alpha)}{N} + \alpha\sum\nolimits_{v\in B(u)} \frac{PR^{t}(v)}{|L(v)|} \tag{2}$$

where N is the number of nodes in the network, α is a constant in [1, 0] called the damping factor or teleport probability. α can be referred to as the probability of users following the links and $1-\alpha$ as the PageRank distribution from non-directly linked pages [29].

B. PERSONALISED PAGERANK (PPR) ALGORITHM

PPR is a variant of PageRank algorithm that focuses on the relative significance of a target node concerning the source node in a graph [28]. In the original PageRank, the rank score of a web page is divided evenly over the pages to which it is linked. Some links may be more critical than others on an actual web-based on the users' preferences. Therefore, PPR was developed to estimate the relevance of nodes according to users' preferences, aiming for personalised search [30]. It simulates a random walker that begins simultaneously at source node u (or a set of source nodes). At each step, the random surfer either jumps to a random out-neighbour node, v, with probability, α, or returns to the current node e_u according to the probability distribution (user preference distribution) with probability $1-\alpha$. The PPR can be defined as [31]:

$$PR^{t+1}(u) = (1-\alpha)\, e_u + \alpha\sum\nolimits_{v\in B(u)} \frac{PR^{t}(v)}{|L(v)|} \tag{3}$$

where e_u is the identity vector of node u whose u^{th} entry is equal to 1, or similarly referred to as the probability vector that contains all other nodes jumping to the node. The difference between PageRank and PPR is that PageRank assumes the random walker returns to any node with uniform probability, while PPR considers the random walker to randomly return to specific states (i.e., query states) [32].

C. PAGERANK WITH PRIORS ALGORITHM

PageRank with Priors is an extension of the PPR algorithm to estimate the relative importance of nodes in a graph based on a set of root nodes. The root set can be represented as the data analyst's prior knowledge or bias based on the nodes in the graph that are deemed essential. Both the PageRank with Priors and PPR algorithms share the similar goal of ranking nodes in a graph, except for the particular context of PageRank and Web pages, which is to bias the standard PageRank rankings in favour of a set of prior topics (root set) [33]. PageRank with Priors defines a prior bias vector used to assign a probability distribution to a set of root nodes, where the root nodes have probabilities of more than zero and all probabilities add up to one [34]. The back probability parameter, α governs the probability that a walk on the graph will restart at a root node. In this context, the root node denotes a known disease gene. Meanwhile, the random surfer lands on any node during this set of walks with a probability of α or ends stochastically at the prior bias nodes (p_u) with a probability of $1-\alpha$. Mathematically, PageRank with Priors is defined as [33]:

$$PR^{t+1}(u) = (1-\alpha)\, p_u + \alpha \sum\nolimits_{v\in B(u)} \frac{PR^t(v)}{|L(v)|} \tag{4}$$

where p_u refers to the prior bias of node u. In general, the difference between PageRank with Priors and PPR is that PageRank with Priors allows any weight distribution of nodes associated with a set of defined root nodes (root nodes consist of a prior bias vector). Contrarily, PPR assumes a uniform distribution for all the nodes related to a set of topic-specific seed nodes (seed nodes consist of topic-specific identity vectors).

D. RANDOM WALK RESTART (RWR) ALGORITHM

RWR is an improved PageRank algorithm that measures each node's relevance with respect to a given single seed node in a graph [35]. RWR executes a random walker that begins simultaneously at source node s. At each state of a certain step, there is a possibility to move to a neighbouring node along an edge (based on edge weights) with probability α or to restart from the source node s with probability $1-\alpha$. RWR can be formally defined as [36] and [37]:

$$PR^{t+1}(u) = (1-\alpha)\, s + \alpha Q PR^t(v) \tag{5}$$

where s is the vector that contains N entries vector elements. All its entries are set to 0 except for the single seed node [14]. Q is the normalised adjacency (transition) matrix. Compared to PageRank with Priors, the initial probability vector of RWR was constructed to assign equal probability to each seed node (seed nodes consist of disease genes). PageRank with Priors initialises prior information vectors (e.g., seed nodes incorporating disease similarity information) to a set of defined root nodes.

E. WEIGHTED PAGERANK ALGORITHM

Weighted PageRank is an extension of PageRank combined with the RWR algorithm to compute the closeness between any two nodes in a graph. In the original PageRank, the transition of a random walker from a node to its neighbours relies upon the corresponding quantity of its neighbours. However, in Weighted PageRank, the computation is performed by iteratively visiting the neighbours with which the edges connecting the node have higher weights [14]. Thus, by reinforcing the weight of interactions, Weighted PageRank can be defined as [38]:

$$PR^{t+1}(u) = (1-\alpha)\, PR^0 + \alpha W PR^t(v) \tag{6}$$

where PR^0 is the initial probability vector, generated by assigning the set of root nodes (known disease genes) with an equal probability of being a start node, which sums to 1. All other nodes are designated a value of 0 [34]. W is the normalised adjacency (transition) matrix, whose values depend on the weight of the edges represented by $\sum_{v\in B(u)} w(u,v)$. A weight-adjustment scheme is introduced to adjust the degree of modularity in a biological network. The difference between Weighted PageRank and RWR lies in constructing the transition matrix. Weighted PageRank intensifies the weights of interactions using an efficient parameter (e.g. weight-reinforcement rate parameter) to modularise the network [39], [40]. Meanwhile, RWR naively considers the original interaction weights based on reliability scores in the PPI network [41], [42], [43], gene ontology based on the similarity of genes [44] or the relationship between heterogeneous biomedical concepts [45] for network construction.

III. RANDOM WALK ON BIOLOGICAL NETWORKS

Biological networks can be represented as graphs that serve as models of biological systems, where each node is a unit (gene or protein) and each edge indicates the interaction between two units [2]. Biological networks are categorised into homogeneous, heterogeneous, and multiple networks. Random walk models can be implemented in single (homogeneous network) or multi-networks (heterogeneous or two-separated networks) based on the classified category. A brief description of random walking on homogeneous, heterogeneous, and two-separated networks is provided as follows.

A. RANDOM WALK ON HOMOGENEOUS NETWORKS

A homogeneous network is a graph with a single type of nodes and a single type of edges. Gene-gene, PPI, phenotype, and gene expression networks are homogeneous networks [2]. Given a graph G = (V, E), V is the set of nodes and E is the set of edges. Let *A(NxN)* denote the adjacency matrix of the homogeneous graph, where it has an entry of 1, if two vertices i and j are connected and 0 otherwise. The equation can be represented as:

$$A(i,j) = \begin{cases} 1, & (i,j)\in E \\ 0, & otherwise. \end{cases} \tag{7}$$

The normalised adjacency matrix is obtained by dividing each row by the degree of the corresponding node. Formally, the normalised adjacency matrix is defined as [46]:

$$W(i,j) = \frac{1}{degree\ (i)} A(i,j) \tag{8}$$

where each row of A is normalised, summing up to 1. The computed normalised adjacency matrix is applied in equation (6) to obtain the steady-state probability vector.

B. RANDOM WALK ON HETEROGENEOUS NETWORKS

A heterogeneous network refers to a graph consisting of different types of nodes and edges. It is constructed by integrating two or more homogeneous networks with known associations. Some of the common heterogeneous networks include gene-to-phenotype networks, gene-disease networks, phenotype-disease networks, and transcription regulatory networks [2]. For instance, let A_G *(NXN)* and A_P*(MxM)* be the adjacency matrixes of two input networks. The mapping of these two networks is stored in matrix *B(NxM)*. The integration of the two input networks and their association network forms a heterogeneous network, which is denoted as follows:

$$A = \begin{bmatrix} A_G & B \\ B^T & A_P \end{bmatrix} \tag{9}$$

where B^T is a transpose of matrix B. A random walker iteratively transitions from its current node to a randomly selected neighbour, starting at a given set of seed nodes in subnetworks A_G and A_P. During a random walk on a heterogeneous network, the walker is likely to stay in a subnetwork while jumping from one subnetwork to another through their interrelationships at a certain probability [14]. The following equation illustrates the process:

$$PR^{t+1} = (1-\alpha)\, M^T PR^0 + \alpha PR^t \tag{10}$$

where M is the transition matrix of the heterogeneous network consisting of four subnetwork transition networks and is denoted as follows [47]:

$$M = \begin{bmatrix} M_G & M_{GP} \\ M_{PG} & M_P \end{bmatrix} \tag{11}$$

where M_G and M_P are intra-subnetwork transition matrices of networks G and P. M_{GP} and M_{PG} represent the inter-subnetwork transition matrices between networks G and P. Let α be the jumping probability between the two subnetworks. When the random walker is in network G, it can jump to network P or stay in network G. If a node is directly linked to network P, the random walker will jump to network P with a probability of α, or move to other nodes in network G with a probability of $1-\alpha$. Otherwise, it will not be able to jump to network P and can only move to other nodes in network G. Thus, the inter-subnetwork transition probabilities between networks G and P are described as:

$$(M_{GP})_{i,j} = \begin{cases} \frac{\alpha B_{i,j}}{\sum_j B_{i,j}}, & if \sum_j B_{i,j} \neq 0 \\ 0, & otherwise. \end{cases} \tag{12}$$

$$(M_{PG})_{i,j} = \begin{cases} \frac{\alpha B_{j,i}}{\sum_j B_{j,i}}, & if \sum_j B_{j,i} \neq 0 \\ 0, & otherwise. \end{cases} \tag{13}$$

Meanwhile, the intra-subnetwork transition matrices of networks G and P can be defined as:

$$(M_G)_{i,j} = \begin{cases} \frac{(A_G)_{i,j}}{\sum_j (A_G)_{i,j}}, & if \sum_j B_{i,j} = 0 \\ \frac{(1-\alpha)(A_G)_{i,j}}{\sum_j (A_G)_{i,j}}, & otherwise. \end{cases} \tag{14}$$

$$(M_P)_{i,j} = \begin{cases} \frac{(A_P)_{i,j}}{\sum_j (A_P)_{i,j}}, & if \sum_j B_{i,j} = 0 \\ \frac{(1-\alpha)(A_P)_{i,j}}{\sum_j (A_P)_{i,j}}, & otherwise. \end{cases} \tag{15}$$

The initial probability that begins with the seed nodes in networks G and P is denoted by u_0 and v_0, respectively. The initial probability vector of the heterogeneous network is denoted as:

$$PR^0 = \begin{bmatrix} (1-\eta)u_0 \\ \eta v_0 \end{bmatrix} \tag{16}$$

where parameter $\eta \in (0, 1)$ balances the level of importance of each subnetwork. When $\eta = 0.5$, the importance of networks G and P are equal. If $\eta > 0.5$, the importance of network G becomes greater than network P, and vice versa. A steady-state probability $PR\infty$ is achieved after several steps and is denoted as:

$$PR^\infty = \begin{bmatrix} (1-\eta)u_\infty \\ \eta v_\infty \end{bmatrix} \tag{17}$$

The nodes in networks G and P are ranked based on steady probabilities of u_∞ and v_∞, respectively.

C. RANDOM WALK ON TWO SEPARATED NETWORKS

Random walking on two separated networks can be performed based on a balanced or unbalanced bi-random walks algorithm. A balanced bi-random walk algorithm begins simultaneously with seed nodes in two input networks and walks separately across each network. The potential interrelationships between the nodes in the two networks are explored while walking following some known and recently updated connections [14]. Mathematically, the process is illustrated by the following equation [48]:

$$PR^{t+1} = (1-\alpha)\, GPR^t P + \alpha A \tag{18}$$

G and P represent the affinity matrices of networks G and P, respectively. A is the known association matrix that acts as prior knowledge to regulate the iteration process. PR is iteratively updated by extending the path in the two networks (achieved by multiplying G on the left and P on the right in each step) [14]. The parameter α regulates the weight of known associations in matrix A.

On the other hand, the process can also be taken sequentially on the two networks based on an unbalanced bi-random walk algorithm instead of random walking on two separate networks simultaneously [48]. Theoretically, the random walker employs a different number of steps for the two input networks, eventually converging to a stationary distribution by taking a series of random walking steps separately. Since the two input networks contain different topologies and structures, the optimal number of random walk steps might differ for the two networks. The two parameters introduced into the two networks include l and r, representing the numbers of maximal iterations, for which l is for network G and r for network P. The mathematical definitions are as follows:

$$\text{Network G: } PR_G^{t+1} = (1-\alpha)\, GPR^t + \alpha A \tag{19}$$

$$\text{Network P: } PR_P^{t+1} = (1-\alpha)\, PPR^t + \alpha A \tag{20}$$

$$\text{Merged result: } PR^{t+1} = \frac{(\lambda_G PR_G^{t+1} + \lambda_P PR_P^{t+1})}{(\lambda_G + \lambda_P)} \tag{21}$$

where λ_G and λ_P ensures the maximal walking steps taken on network G and network P does not exceed the threshold l and r, respectively.

IV. DISEASE GENE PREDICTION

Identifying disease-associated genes is a task of predicting the most plausible candidate genes involved in a disease [49]. With the development of high-throughput technologies, genetic mapping approaches emerged to generate candidate disease genes. The traditional genetic mapping methods include linkage analysis and genome-wide association studies (GWAS), which provide chromosomal regions containing up to ten or even hundreds of candidate genes possibly associated with genetic diseases [50]. However, it may not be possible to experimentally validate the candidate disease genes that lie on the specified genomic intervals. Thus, computational disease gene prioritisation may be an optimal strategy for identifying the most promising candidates among the long list of genes to reduce experimental costs and efforts.

Random walk-based methods are network-based computational approaches that represent biological data as a network and apply graph mining techniques to predict disease candidate genes [21]. The ability to amplify association between genes that lie in network proximity facilitates the analysis of biological pathways for disease gene prediction. Random walk-based methods can be categorised into node classification and link prediction tasks. Node classification uses the known disease genes to infer the disease label of the unlabeled genes, whereas link prediction uses gene-disease associations to identify the potential disease-causing genes [11]. The following subsections describe the formal definitions of node classification and link prediction tasks. The relevant biological applications of random walk methods based on node classification and link prediction are also elaborated. A comprehensive review of random walk-based methods for disease gene prediction based on the two tasks is described below.

A. NODE CLASSIFICATION

Node classification aims to predict or classify unlabeled nodes (genes with unknown disease associations) in the biological network, with known labels on some nodes (genes with known disease associations). In a homogeneous graph G = (V, E), V refers to the set of nodes/genes and E is the relationships between nodes. Let a subset of genes labelled as disease-causing genes, $V_{\text{labeled}} \in V$, and another set of genes with unlabeled disease associations, $V_{\text{unknown}} = V \backslash V_{\text{labeled}}$ [11]. Node classification on network G predicts the labels of nodes in V_{unknown} [11]. Similar criteria can also be used to define this node classification task in heterogeneous graphs and multi-view graphs.

Disease gene prioritisation, also known as disease gene association prediction, is one of the popular biological applications of random walk methods for disease gene prediction based on node classification tasks. The prioritisation (or the selection of a smaller subset) of candidate genes is the process of assigning similarity or confidence scores to genes before ranking them based on the probability of being causally related to a disease of interest [51], [52]. Disease gene prioritisation primarily comprises three steps. First, some known disease genes are chosen as seed genes. Then, the positions of the seed genes on their chromosomes are determined based on gene expression profiling, linkage regions, and other chromosomal abnormalities [14]. However, these approaches have identified thousands of candidate genes, most of which are irrelevant to the disease of interest, indicating the need to rank candidate genes using a prioritisation method to identify the most likely disease genes from these candidates.

1) RANDOM WALK METHODS BASED ON NODE CLASSIFICATION

Network-based candidate gene prioritization (ToppNet) [53] adopts PageRank with a priors algorithm to prioritise disease candidate genes based on their relative importance in PPI network. A list of known disease-related genes is used as the prior bias vectors to run the PageRank algorithm with different parameter values. Besides that, PRIoritizatioN and Complex Elucidation (PRINCE) [54] is proposed to prioritise genes and protein complexes associated with a disease of interest. It computes the disease similarity measures of known causal genes as prior probability vectors to run PageRank with the priors algorithm on a weighted PPI network. While Network Propagation with Dual Flow (NPDE) [55] employs a dual-flow PageRank with a priors approach to prioritise candidate disease genes. It aims to analyse the topology associations between disease and essential proteins by assigning positive flow to known disease proteins and negative flow to essential proteins within the PPI network. The empirical results demonstrated that disease genes are not well connected with essential genes to conclude further that disease proteins are topologically more important than other proteins in the network.

On the other hand, VAVIEN [56] aims to measure the topological similarity among the protein pairs in the PPI network. It uses the RWR algorithm to construct topological similarity between the seed and candidate proteins based on the proximity of the protein of interest to every other protein in the network. The computed topological profile scores are then used for candidate disease gene prioritisation. Next, Diffusion Profile based on Linear Correlation Coefficient (DP-LCC) [57] is proposed as a diffusion-based method to prioritise candidate disease genes in PPI network. It constructs separate diffusion profiles for disease genes and candidate genes to compare both profile vectors with the query disease based on a linear correlation coefficient. Whereas PRioritization bY protein NeTwork (PRYNT) [58] employs two closeness-based algorithms, shortest-path and random walk, to prioritise the kidney disease genes. The PPI network is contextualised by grouping the proteins within cliques. The multiplication of rank scores computed from both strategies proved that the results were better than direct ranking implemented in previous studies.

RWR [41], [59] is proposed to prioritise candidate disease genes based on random walk methods. The calculated rank score reflects the global similarity of candidate genes to known members of a disease-gene family in the PPI network. While Degree-Aware Disease Gene Prioritization (DADA) [60] introduces a disease gene prioritisation method based on statistical adjustment to correct degree bias in the conventional RWR algorithm. DADA suggests three reference models: the degree of disease genes (seed nodes), the degree of candidate genes, and the likelihood ratio using eigenvector centrality to adjust the degree distribution of the PPI network. Although these methods successfully identified the loosely connected disease genes, they also created more false negatives for highly connected genes. Neighbour-favoring weight reinforcement (ORIENT) [38] proposed a Weighted PageRank algorithm to prioritise candidate disease genes in the PPI network improving the conventional RWR algorithm by introducing an efficient parameter to reinforce the weights of interactions close to the known disease genes. The proposed method thoroughly considered the modularity principle through proper neighbour-favoring weight reinforcement.

Directed Random Walk (DRW) [61] applies the RWR algorithm to infer robust pathway biomarkers at functional categories level than of the individual genes. It introduces an efficient gene-weighting strategy according to topological importance, effectively enhancing the reproducibility of pathway activities for cancer classification. Whereas significant Directed Random Walk (sDRW) [62] aims to assess the optimal restart probability parameter according to different genomic datasets by introducing an additional weight to enhance the conventional RWR algorithm. Enhanced Directed Random Walk (eDRW+) [63] adopted the RWR algorithm to identify breast cancer prognostic markers from multiclass expression data. This method utilises an analysis of variance (ANOVA) F-test statistic and pathway selection to improve the weight of genes in the directed pathway network. Integrative Directed Random Walk (iDRW) [64] proposed a multi-omics data integration method based on DRW algorithm for disease gene prediction. It constructs a directed gene-gene interaction graph based on gene expression and copy number alteration. It further defines an effective weight initialisation and genes scoring method to identify topologically important genes and pathways. PWN [17] is a variant RWR algorithm that prioritises disease targets based on a combination of internal and external features of network warping: graph curvature and prior knowledge. It generates a weighted asymmetric network from unweighted and undirected networks by computing the edges' Ricci curvature and assigning higher weights to prior knowledge-related edges based on RWR. The final gene scores are obtained via diffusion through the warped network. Figure 2 illustrates the graphical overview of PWN.

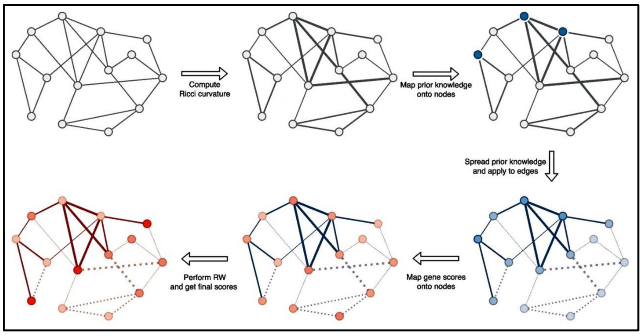
Compute Ricci curvature
Map prior knowledge onto nodes
Spread prior knowledge and apply to edges
Map gene scores onto nodes
Perform RW and get final scores

FIGURE 2. Graphical overview of PWN [17].

On the other hand, BioGraph [45] utilises a stochastic RWR model on an integrated network containing 21 publicly available curated databases for disease gene prioritisation. This data integration and mining platform computes the posterior probability for a given candidate gene prioritisation query to identify genes for hereditary diseases. Simplified Laplacian Normalization-Supervised Random Walk (SLN-SRW) [65] integrates biomedical data from heterogeneous sources to predict disease genes. It proposed a Laplacian normalisation-based supervised random walk algorithm to model an integrated network's edge weights for the prediction of gene-disease relationships. Meanwhile, Driver genes discovery with Improved Random Walk method (Driver_IRW) [66] is a novel method based on the RWR algorithm to identify cancer driver genes by integrating transcriptomic data and interaction networks. This method incorporates transition probabilities and global centrality measures to compute the probability vectors for random walking to seed nodes.

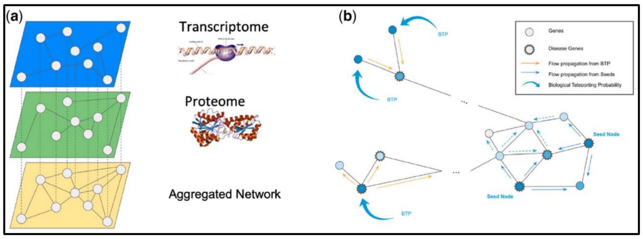

FIGURE 3. Framework of BRW [68].

Weighted PageRank [67] aims to prioritise type 2 diabetes genes by leveraging the modified PageRank algorithm on bilayer biomolecular networks. It constructs the network based on differential mutual information and ranks the diabetes genes using RWR on the heterogeneous networks. Biological Random Walks (BRW) [68] employs RWR to leverage the integration of multiple biological sources for disease gene prioritisation. This method computes personalisation vectors and aggregated transition matrix using a convex combination before applying a random walk model to rank genes. Figure 3 illustrates the framework of BRW.

Improved sDRW [69] is an enhanced sDRW algorithm that implements sequential random walks on two biological networks. It aims to enhance the sensitivity of cancer prediction in conventional sDRW algorithm by introducing a walker network to identify significant genes in both networks. Meanwhile, entropy-based Directed Random Walk (e-DRW) [70] performs RWR on two separated networks to prioritise disease genes. It constructs separate biological networks from different pathway databases to improve the coverage of pathway information for random walking. A robust gene-weighting and pathway activity inference method incorporating an entropy probability parameter is proposed to infer pathway biomarkers at the functional categories level. Table 1 summarises a collection of random walk-based methods for disease gene prediction based on node classification tasks (refer to Appendix Table S1 for more details).

B. LINK PREDICTION

Link prediction aims to predict unknown links between two sets of nodes (i.e. genes and diseases) based on known associations between the nodes (known disease-gene associations). In a heterogeneous graph G, denoted as G(U, V, E), U and V represent the sets of genes and diseases, respectively. At the same time, E indicates the edges in U, V and those between U and V (i.e. known disease-gene associations) [11]. Link prediction on network G predicts disease–gene associations by measuring the proximity or similarity between the nodes/genes for the disease of interest.

Random walk methods have several biological applications for disease gene prediction based on link prediction tasks. These applications include protein function prediction, drug target interaction prediction, microRNA-disease association prediction, and lncRNA-disease association prediction. Protein function prediction predicts the function of a protein by exploring the protein-function relationship from PPI networks and Functional Interrelationship Networks (FIN). PPI network refers to a complex network of associated proteins, whereas the FIN network is constructed based on Gene Ontology (GO) term functional similarity. Based on the assumption that proteins that are located close to each other in a PPI network tend to share similar functional annotations, and two similar functions usually co-annotate a common protein [14], random walk models effectively diffuse information to the whole networks by discovering the interrelationships between nodes of different biological networks based on converged probability distribution.

On the other hand, drug target interaction prediction is a typical link prediction problem that aims to facilitate drug repositioning. Random walk methods are a drug repositioning tool used to predict unknown drug targets or drug-disease interactions. Since similar drugs often target similar proteins, several biological networks like drug-drug interaction networks and PPI networks are employed to explore the nodes' associations and solve the prediction problem. Suppose random walk is considered for heterogeneous network. In that case, a drug-drug interaction network can be constructed based on drug chemical structure similarity, and PPI network can be constructed based on amino acid sequences of target proteins [14]. Random walk models perceive heterogeneous network as input and compute the likelihood of an edge between pairs of proteins and drugs through network diffusion.

Apart from that, microRNAs are single-stranded noncoding RNAs that play an important role in the pathogenesis of human diseases [71]. Random walk models represent a promising tool to uncover potential miRNA-disease associations using the constant accumulation of miRNA, disease, and miRNA-disease association data. Suppose functionally related miRNAs are frequently associated with phenotypically similar diseases [71]. In that case, microRNA-disease associations can be predicted by constructing two subnetworks, miRNA functional similarity networks and disease phenotype similarity networks bridged by known miRNA-disease associations. As such, random walk models jump from one subnetwork to another due to their interrelationships at a certain probability to detect miRNA candidates that could potentially be associated with diseases.

Long-non-coding RNAs (lncRNAs) are long chains of nucleotides with various biological mechanisms closely related to human diseases, including cancers and degenerative neurological diseases [72]. Based on the hypothesis that functionally similar lncRNAs are possibly related to diseases with similar phenotypes [73], lncRNA-disease association prediction has rapidly gained attention among researchers in understanding the pathogenesis of diseases at a molecular level. By integrating multiple biological data sources, random walk models can effectively integrate disease semantic similarity networks and lncRNA function similarity networks with known lncRNA-disease associations to predict lncRNA-disease associations.

1) RANDOM WALK METHODS BASED ON LINK PREDICTION

Random Walk with Restart on Heterogeneous Network (RWRH) [42] is an extended RWR algorithm that prioritises genes and phenotypes simultaneously using known gene-phenotype relationships. Gene-phenotype associations connect the gene and phenotype networks to construct a heterogeneous network. Random Walker on Protein Complex Network (RWPCN) [74] is proposed to predict and prioritise disease genes on a heterogeneous network comprising of phenotype similarity network, protein complex network, and protein interaction network. It uses protein complexes to aid in their inference of gene-phenotype associations for disease gene prioritisation. Figure 4 presents the overall network structure of RWPCN. Meanwhile, Random Walk with Restart on Multiplex-Heterogeneous network (RWR-MH) [75] extended the RWR algorithm to multiplex and heterogeneous networks to prioritise disease genes. A multiplex network is formed by integrating PPI, pathway, and co-expressed networks. This multiplex network is further connected to a disease-disease similarity network through gene-diseases associations to predict disease-associated genes.

Laplacian normalisation based Random Walk with Restart on Heterogeneous network (LapRWRH1 and LapRWRH2) [76], a Laplacian normalisation-based RWR on heterogeneous network algorithm, prioritises disease genes and identifies potential gene-phenotype relationships. Laplacian normalisation is utilised to normalise the weight of edges in heterogeneous networks and transition probability matrices. Besides that, Network-based Random Walk with Restart on the Heterogeneous network (NRWRH) [77] was developed to infer potential drug–target interactions based on RWR in a heterogeneous network. Drug similarity and protein (target) similarity networks are connected via drug-target interactions for drug–target prediction. Random Walk with Restart on Heterogeneous Network with Multiple Data Sources (RWRH-MDA) [78] operates RWR on a heterogeneous network to predict miRNA-disease associations. The heterogeneous network is constructed based on disease similarity, while the miRNA similarity network is connected by known miRNA-disease interaction networks. This heterogeneous network overcomes the limitations of previous methods (i.e. use of only a single dataset, inadequate disease semantic similarity, and overestimation of the predictive accuracy) to identify potential disease-related miRNAs.

Random Walk with Restart on Multigraphs (RWRM) [79] adopts the RWR algorithm to prioritise disease genes based on the proposed Complex Heterogeneous Network (CHN). Whereas the CHN model is constructed based on PPI network and multigraph gene network (i.e. integration of Biological Process (BP), Cellular Component (CC), and Molecular Function (MF) network). A phenotype network is then connected to the model as a subgraph to guide the random walk.

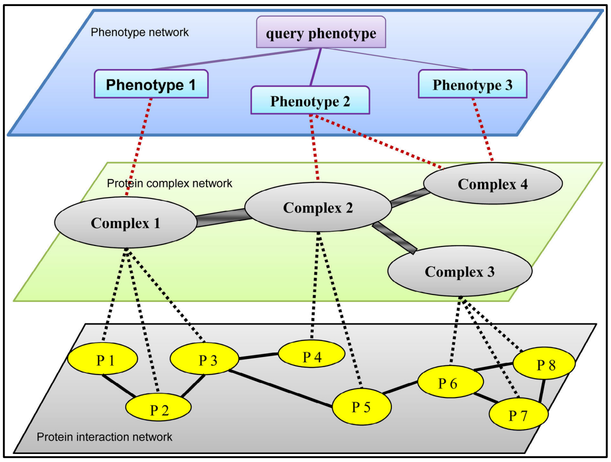

FIGURE 4. Overall network structure of RWPCN [77].

Two-Rounds Random Walk with Restart based on Multiple Biological Data (TRWR-MB) [80] is an extension of the RWRH algorithm to explore cancer genes based on a quadruple-layer heterogeneous network. The network integrates multiple biological data consisting of PPI network, pathway network, microRNA similarity network, lncRNA similarity network, cancer similarity network, and protein complexes. A two-round RWR is then executed on the network to obtain the final ranking score. RWRH-Malaria [81] was proposed to predict malaria-associated genes based on RWR on cross-species PPI networks for humans and parasites. The network integrates human-human, parasite-parasite (*Plasmodium falciparum*), and human-parasite protein interactions using known malaria genes as the seeds to identify candidate malaria genes.

Furthermore, the prediction of potential miRNA-disease associations based on degree-based RWR on a heterogeneous network was discovered called Biased Random Walk with Restart on Multilayer Heterogeneous networks for MiRNA–Disease Association prediction (BRWRMHMDA), an enhanced Biased Random Walk with Restart (BRWR) method [82]. This method designed a multilayer heterogeneous network based on known miRNA–disease associations, disease semantic similarity, miRNA functional similarity, and Gaussian interaction profile kernel similarity for diseases and miRNAs. Biased RWR was then implemented on the degree-based heterogeneous network to obtain potential miRNA–disease associations. Bi-Random Walk (BiRW) [48], on the other hand, employs a bi-random walk algorithm to prioritise disease genes based on paired phenotype-gene associations. It aims to capture the patterns of the phenome-genome association network based on a regularisation framework for graph matching. RWR is performed on the Kronecker product graph between PPI and phenotype similarity networks based on balanced and unbalanced steps. Meanwhile, Unbalanced Bi-Random Walk (UBiRW) [83] applies an unbalanced bi-random walk on PPI network and functional interrelationship network to predict

TABLE 1. Random walk-based methods based on node classification.

Methods	Approaches	Network Format	Types of Input Data	Applications	Platform	Availability	References
ToppNet	PageRank with Priors	Homogeneous	PPI	Disease gene prioritisation	-	-	[53]
PRINCE	PageRank with Priors	Homogeneous	PPI, PHE	Disease gene prioritisation, protein complex prediction	-	-	[54]
NPDE	PageRank with Priors	Homogeneous	PPI, PHE	Disease gene prioritisation	-	-	[55]
VAVIEN	Random Walk Restart	Homogeneous	PPI, PHE	Disease gene prioritisation	No longer available	No longer available	[56]
DP-LCC	Random Walk Restart	Homogeneous	PPI, PHE, TEM	Disease gene prioritisation	-	-	[57]
PRYNT	Random Walk Restart	Homogeneous	PPI, GPD	Disease gene prioritisation	R package	https://github.com/Boizard/PRYNT	[58]
RWR	Random Walk Restart	Homogeneous	PPI, PHE, SEQ	Disease gene prioritisation	No longer available	No longer available	[41, 59]
DADA	Random Walk Restart	Homogeneous	PPI, PHE, TEM	Disease gene prioritisation	Matlab code	http://compbio.case.edu/omics/software/dada/	[60]
ORIENT	Weighted PageRank	Homogeneous	PPI, Functional Linkage Network (FLN)	Disease gene prioritisation	-	-	[38]
DRW	Random Walk Restart	Homogeneous	GPD, FAP	Disease gene prioritisation	No longer available	No longer available	[61]
sDRW	Random Walk Restart	Homogeneous	GPD, FAP	Disease gene prioritisation	-	-	[62]
eDRW+	Random Walk Restart	Homogeneous	GPD, FAP	Disease gene prioritisation	-	-	[63]
iDRW	Random Walk Restart	Homogeneous	GPD, FAP	Disease gene prioritisation	-	-	[64]
PWN	Random Walk Restart	Homogeneous	GPD, PPI, TEM	Disease gene prioritisation	Python code	https://github.com/Standigm/PWN	[17]
BioGraph	PageRank	Heterogeneous	PPI, GPD, ONT, FAP, PHE, TEM	Disease gene prioritisation	Web tool	https://biograph.be/	[45]
SLN-SRW	Weighted PageRank	Heterogeneous	PPI, PHE, ONT, TEM	Disease gene prioritisation	-	-	[65]
Driver_IRW	Random Walk Restart	Heterogeneous	GPD, FAP, PHE	Disease gene prioritisation	-	-	[66]
Weighted PageRank	Random Walk Restart	Heterogeneous	GRN, PPI, GPD	Disease gene prioritisation	-	-	[67]
BRW	Weighted PageRank	Heterogeneous	PPI, ONT, FAP, GPD, PHE	Disease gene prioritisation	Python code	https://github.com/LeoM93/BiologicalRandomWalks	[68]
Improved sDRW	Random Walk Restart	Two Separated	PPI, FAP, GPD	Disease gene prioritisation	-	-	[69]
e-DRW	Random Walk Restart	Two Separated	GPD, FAP	Disease gene prioritisation	R package	eDRW R package	[70]

PPI, Protein-Protein Interaction; ONT, Ontology; ORT, Orthology; PHE, Phenotype Relationship; GPD, Genomic or Proteomic Data; SEQ, Sequence Data; Gene Co-expression Network; GRN, Gene Regulatory Network; DDI, Drug-drug Interaction; DTI, Drug-Target Interaction; FAP, Functional Annotation and Pathways; TEM, Text Mining.

protein functions. It adopts a different number of walking steps on the two networks to infer protein-gene ontology term associations.

Unbalanced Random Walk on Three Biological Networks (ThrRW) [84] implements RWR by considering several steps of random walking on three biological networks: protein interaction network, domain co-occurrence network, and functional interrelationship network to predict functions for unknown proteins. Functional protein information is propagated among the three networks through associations between the nodes in different networks. Three-layer heterogeneous network Combined with unbalanced Random Walk for MiRNA-Disease Association prediction (TCRWMDA) [85] aims to predict the potential miRNA-disease associations based on an unbalanced random walk on a three-layer heterogeneous network. To compute the potential association scores between disease and its associated miRNAs, it takes three different random walking steps on lncRNA similarity network, disease similarity network, and miRNA similarity network for miRNA-disease association prediction. Multiple Similarities Fusion based on Unbalanced Bi-Random Walk (MSF-UBRW) [73] is based on a multiple similarities fusion of an unbalanced bi-random walk used to identify lncRNA-disease associations. This method fuses multiple

TABLE 2. Random walk-based methods based on link prediction.

Methods	Approaches	Network Format	Types of Input Data	Applications	Platform	Availability	References
RWRH	Random Walk Restart	Heterogeneous	PPI, PHE	Disease gene association prediction	No longer available	No longer available	[42]
RWPCN	Random Walk Restart	Heterogeneous	PPI, GPD, PHE	Disease gene association prediction	-	-	[74]
RWR-MH	Random Walk Restart	Heterogeneous	PPI, GCN, FAP, PHE	Disease gene association prediction	R package	https://github.com/alberto-valdeolivas/RWR-MH	[75]
LapRWRH1 and LapRWRH2	Random Walk Restart	Heterogeneous	PHE	Disease gene association prediction	-	Available upon request	[76]
NRWRH	Random Walk Restart	Heterogeneous	Drug and protein domain information, DDI, DTI	Drug–target interaction prediction	-	-	[77]
RWRH-MDA	Random Walk Restart	Heterogeneous	PHE, ONT, GPD, PPI, FAP	microRNA-disease association prediction	No longer available	No longer available	[78]
RWRM	Random Walk Restart	Heterogeneous	PHE, PPI, GO, FAP	Disease gene association prediction	-	-	[79]
TRWR-MB	Random Walk Restart	Heterogeneous	PPI, FAP, PHE, GPD	Disease gene association prediction	-	-	[80]
RWRH-Malaria	Random Walk Restart	Heterogeneous	PPI, TEM	Disease gene association prediction	-	-	[81]
BRWRMH MDA	Random Walk Restart	Heterogeneous	PPI, PHE, FAP	miRNA-disease association prediction	-	-	[82]
BiRW	Random Walk Restart	Two Separated	PHE, PPI	Disease gene association prediction	-	-	[48]
UBiRW	Random Walk Restart	Two Separated	PPI, PHE, ONT, FAP	Disease gene association prediction, protein function association prediction	-	-	[83]
ThrRW	Random Walk Restart	Two Separated	PPI, ONT, GPD	Protein function association prediction	-	-	[84]
TCRWMDA	Random Walk Restart	Two Separated	PPI, PHE	miRNA-disease association prediction	Matlab code	https://github.com/ylm0505/TCRWMDA	[85]
MSF-UBRW	Random Walk Restart	Two Separated	GPD, PHE	lncRNA- disease association prediction	-	-	[73]
BRWMC	Random Walk Restart	Two Separated	GPD, PHE	lncRNA- disease association prediction	-	-	[86]

PPI, Protein-Protein Interaction; ONT, Ontology; ORT, Orthology; PHE, Phenotype Relationship; GPD, Genomic or Proteomic Data; SEQ, Sequence Data; Gene Co-expression Network; GRN, Gene Regulatory Network; DDI, Drug-drug Interaction; DTI, Drug-Target Interaction; FAP, Functional Annotation and Pathways; TEM, Text Mining.

similarities (including functional, Gaussian Interaction Profile Kernel, and linear neighbour similarities) of lncRNAs and diseases to assist different random walking steps for the lncRNA and disease similarity networks, respectively. While Bi-Random Walk and Matrix Completion (BRWMC) [86] is a network-based approach used to predict lncRNA disease association based on a bi-random walk and matrix completion method. It employs RWR to preprocess the known lncRNA-disease association matrix and combines the matrix completion method to predict the association of lncRNA and disease. Table 2 presents a collection of random walk-based methods for disease gene prediction based on link prediction tasks (refer to Appendix Table S2 for more information).

V. DISEASE MODULE IDENTIFICATION

Identification of a disease module is also called module inference or graph clustering. It detects a group of genes related to a disease phenotype [87]. These groups of genes are involved in similar biological functions are called communities, modules, or clusters. It is driven by the underlying assumption that disease-related proteins tend to interact closely in biological networks [88]. Meanwhile, traditional techniques for disease module identification focus on a particular protein or biological pathway and are neither economical nor, by definition, able to study the entire system [89]. For this reason, network-based approaches that model the structure and dynamics of biological systems can aid in identifying disease modules in the human interactome. It offers a comprehensive understanding of the disease mechanisms and pathophenotypes at a system level and directs the search for therapeutic targets [90].

Functional module or protein complex detections, is the main biological application of random walk methods for disease module identification. A functional module can be defined as a group of genes or products connected by one or more genetic or cellular interactions [91]. Since the interactions of gene products in PPI drive the biological process, functional module detection has become a significant biological problem for predicting densely clustered essential proteins and disease genes in biological networks. As clusters of genes or proteins are typically highly and loosely connected with the rest of the nodes in the network, random walk models are more likely to stay within a cluster of connected nodes than travel between them. Based on this concept, various random walk clustering methods were developed to identify the functional modules from the PPI networks.

Markov Clustering algorithm (MCL) is a network-based computational approach based on the simulation of stochastic flow in graphs [92]. The main idea of MCL algorithm is that if a random walker starts from a node and randomly travels to a connected node, it is more likely to stay within a cluster than to cross clusters. In general, the MCL algorithm involves six steps to cluster a network. Firstly, an association matrix given an undirected graph as an input is created. Then, self-loops are added to each node and a normalised adjacent matrix is constructed for the network. Next, repeated multiplication of the adjacent matrix occurs to expand the information flow to other network regions. Followed by the rescaling of the resulting matrix using inflation to strengthen strong currents and weaken weak currents. As the fifth step, the expansion and inflation operations are repeated until they reach a steady state (convergence). Finally, the resulting matrix is interpreted to discover clusters.

Markov Clustering based on Core Attachment on weighted networks (MCL-CAw) [93] is developed as a core-attachment-based refinement method coupled with MCL to identify yeast complexes using weighted PPI networks. It refines the clusters produced by MCL using the core-attachment structure and utilises the affinity-scoring PPI network to derive meaningful yeast complexes. Meanwhile, Soft Regularised Markov Clustering (SR-MCL) [94] adopted MCL as a base algorithm by iteratively re-executing the clustering operation to identify functional modules in PPI network. To ensure different clustering results in each execution, it introduces a penalised ratio to control the stochastic flow of each node. Following the clustering algorithm, a post-processing algorithm is applied to remove redundant and low-quality clusters. Another study proposed Markov Clustering [95] as a graph clustering method to identify protein complexes within highly interconnected PPI networks. It optimises MCL parameter to further compute network modularity and density from the MCL cluster granules to generate protein complexes with high protein interaction.

Next, the Repeated Random Walks (RRW) [37] was proposed as an extended RWR algorithm based on repeated random walks on graphs to discover molecular complex and functional modules within protein interaction networks. The edges in network are weighted by the strength of functional associations and the random walk process is repeated to identify overlapping clusters of yeast genes. In another literature, Node-Weighted Expansion of clusters of proteins (NWE) [96] was used as an enhanced RRW algorithm to detect protein complexes on the PPI network. This method weighted the clusters of nodes by the total sum of the weights of all the adjacent edges in the network. Whereas Local Protein Community Finder [97] applied two local clustering algorithms called Nibble [98] and PageRank-Nibble [99] to discover high-quality communities near a queried protein in a PPI network. This method locally partitions a protein network to identify quality clusters with high conductance and functional coherence.

Weighted PageRank-Nibble and Core Attachment structure (WPNCA) [100] aims to detect protein complexes from PPI networks using weighted PageRank-Nibble algorithm and core-attachment structure. The method assumes that neighbours that tend to construct clusters with the node should assign higher values. It treats adjacent nodes equally by assigning weights with different probabilities based on an edge-clustering coefficient. Walktrap [101] is another algorithm based on RWR to detect modules that are significantly enriched with cancer genes. It develops an integrated network weighted by an average weighting scheme and utilises distances to derive transition probability vectors. An efficient scoring method is proposed to partition the clusters and is further customised based on its modularity, module size, and maximum module score to guide clustering. Figure 5 presents the flow diagram of the Walktrap. A network-based approach [102] is proposed to identify genes and gene modules in breast invasive carcinoma (BRCA) based on RWR on the PPI network. DNA methylation and gene expression data are integrated to calculate the weights of the PPI network using Principal Component Analysis (PCA) and Canonical Correlation Analysis (CCA). The detected significant genes are then used for sub-network

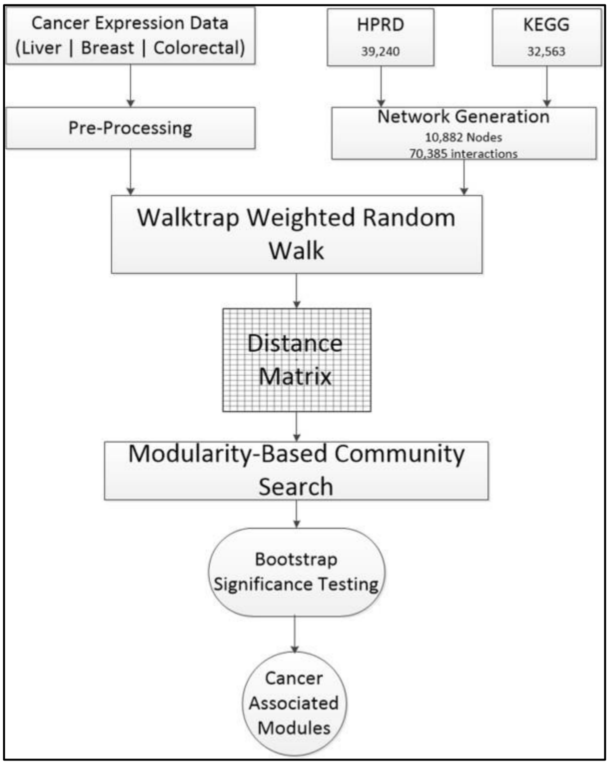
Cancer Expression Data
(Liver | Breast | Colorectal)
HPRD
39,240
KEGG
32,563
Pre-Processing
Network Generation
10,882 Nodes
70,385 interactions
Walktrap Weighted Random Walk
Distance Matrix
Modularity-Based Community Search
Bootstrap Significance Testing
Cancer Associated Modules

FIGURE 5. Flow diagram of Walktrap [105].

construction, while a random walk algorithm is applied to discover candidate disease-related modules.

Isolation [91] is a multiplex approach based on random walks for functional module identification. It integrates mRNA expression information and biomedical knowledge to reveal the functional relations of genes. This algorithm transforms the PPI network based on a k-step random walk that enumerates each node to identify clusters with locally optimal isolation. Mutual EXclusion and Coverage based random walk (MEXCOwalk) [103] is a vertex-weighted, edge-weighted random walk-based approach to extract TCGA pan-cancer driver modules in the PPI network. The weight of edges incorporates coverage information and the degree of mutual exclusivity between pairs of gene neighbourhoods in the network for module detection. TOP-down Attachment of Seeds (TOPAS) [89] implements a top-down approach to detect disease modules based on the RWR of the functional association networks. It seeks to connect the largest number of seed nodes while adding the fewest connectors in the final module. Meanwhile, Active Module Identification using Experimental data and Network Diffusion (AMEND) [104] is a novel active module identification method that uses network diffusion with Equivalent Change Index (ECI) to identify a connected subset of genes regulated similarly or opposingly between the two experimental conditions. It employs RWR to select genes and create gene weights before applying Heinz (heaviest induced subgraph) to determine the maximum-weight connected subgraph using the node weights derived from RWR. The process is iterated until the highest-scoring network is derived as the final module. Table 3 lists a collection of MCL and random walk-based methods for disease module identification (refer to Appendix Table S3 for more details).

VI. CHALLENGES AND FUTURE DIRECTIONS

The random walk model plays a significant role in solving biological problems, such as ranking nodes in biological networks, measuring similarity or distance between nodes in biological networks, detecting modules from biological networks, and determining interrelationships between nodes from different biological networks [14]. It is a highly efficient algorithm as it is fast to implement and applies to large biological networks for analysis. A random walk model can be used to compute the proximity of a node to a set of source nodes and not just a single source node. This property is beneficial when a core set of members of a pathway or complex is known, and queries (or the initial node) for candidate members are being conducted on this network [36]. However, some challenges are observed in applying random walk models to solve biological problems based on networks. Therefore, improvements are necessary to increase the computational efficiency and scalability when such models are used extensively for genome-scale biological networks.

Parameters in random walk models are crucial in controlling the performance of the algorithms. As mentioned before, the random walk model iteratively updates the values vector and obtains a steady-state probability vector when the Euclidean Distances between the current value vector (PRt) and the last time-step vector (PRt+1) are less than the threshold ε. Parameter ε acts as a threshold parameter that controls the precision of values vector in the algorithm. The larger the ε, the faster the convergence of the algorithm [24]. On the other hand, parameter α also known as restart probability or back probability) controls the information flow returning to the seed nodes at each iteration of the algorithm. The larger the α, the more likely it is for the nodes close to the seed nodes to be ranked forward and vice versa [24]. In brief, parameters ε and α not only regulate the number of iterations in the algorithm but also affect the performance in terms of overall accuracy and prediction results.

Besides that, the size of the biological network for random walking can significantly affect the algorithm's computing time. It takes longer for the random walk models to converge when the network size is huge. It implicitly creates a high computational complexity, ultimately limiting the networks' in-depth analyses. However, modelling a dynamic network is another challenge for random walking in a biological network. Inherently, biological networks can change with time, context, and complexity [105]. Although many networks contain such temporal information, most studies applied the random walk model on static snapshots of the graph and have largely ignored the temporal dynamics of the network [106]. Thus, biological network construction is important to yield appropriate and meaningful results for large-scale information networks.

TABLE 3. MCL and random walk-based methods for disease module identification.

Methods	Approaches	Network Format	Types of Input Data	Applications	Platform	Availability	References
MCL-CAw	Markov Clustering	Homogeneous	PPI	Yeast complex detection	-	-	[93]
SR-MCL	Markov Clustering	Homogeneous	PPI, ONT	Functional module detection	-	Available upon request	[94]
Markov Clustering	Markov Clustering	Homogeneous	PPI	Protein complex detection	-	-	[95]
RRW	Personalized PageRank	Homogeneous	PPI, ONT	Functional module detection	-	-	[37]
NWE	Random Walk Restart	Homogeneous	PPI	Protein complex detection	-	-	[96]
Local Protein Community Finder	Random Walk Restart	Homogeneous	PPI	Protein complex detection	No longer available	No longer available	[97]
WPNCA	Random Walk Restart	Homogeneous	PPI	Protein complex detection	No longer available	No longer available	[100]
Walktrap	Random Walk Restart	Heterogeneous	PPI, FAP, GPD	Functional module detection	No longer available	No longer available	[101]
Network-based approach	Random Walk Restart	Homogeneous	PPI, GPD	Disease gene module detection	-	-	[102]
Isolation	Random Walk Restart	Heterogeneous	PPI, GPD, TEM	Functional module detection	-	-	[91]
MEXCOwalk	Random Walk Restart	Homogeneous	PPI, GPD	Functional module detection	Python code	https://github.com/abu-compbio/MEXCOwalk	[103]
TOPAS	Random Walk Restart	Homogeneous	PPI, FAP, SEQ, DTI, GPD	Disease gene module detection	R package	https://bitbucket.org/sonnhammergroup/topas/src	[89]
AMEND	Random Walk Restart	Homogeneous	PPI, GPD	Functional module detection	R package	https://github.com/samboyd0/AMEND	[104]

PPI, Protein-Protein Interaction; ONT, Ontology; ORT, Orthology; PHE, Phenotype Relationship; GPD, Genomic or Proteomic Data; SEQ, Sequence Data; Gene Co-expression Network; GRN, Gene Regulatory Network; DDI, Drug-drug Interaction; DTI, Drug-Target Interaction; FAP, Functional Annotation and Pathways; TEM, Text Mining.

With the development of high-throughput techniques, random walk models can be improved to overcome these limitations. The optimal value of the parameter for applications should be determined using theoretical formulas or equations [14]. Additionally, the cost of computing probability vectors should be reduced to improve the efficiency of the algorithms for various machine learning tasks, such as node similarity measure, link prediction, classification, and clustering. There should be more focus on the construction of biological network dynamics, including dynamical network construction [107], [108], dynamical disease genes prediction [109], and dynamical functional module identification [110], [111], [112]. In the future, more efforts should be made towards designing effective random walk-based methods to work on active subnetworks as well as association networks of those subnetworks. Besides, the integration of multi-biological resources and multi-biological networks should be emphasised to improve the application of random walk model in solving biological problems based on networks.

VII. CONCLUSION

Identifying disease genes and disease modules are critical for understanding disease mechanisms and uncovering disease-gene associations. Random walk-based approaches have been widely applied in bioinformatics for solving biological problems based on biological networks. This study reviewed some diffusion-based random walk methods leveraging various networks in their problem formulation for disease gene prediction and disease module identification. The basic concepts of the random walk model, including a variation of random walk approaches, were introduced for specific applications. This review focused on underscoring the strengths and weaknesses of state-of-the-art random walk methods for disease gene prediction and disease module identification instead of their prediction performance. An organised, up-to-date overview of the computational approaches provided merit exploitation for researching the genetic causes of human diseases.

Selecting a random walk computational approach for specific biological problems is difficult because it depends on various factors. Some general principles are provided as guidance for potential users of such applications. An important consideration that needs to be addressed is whether the random walk methods can integrate multi-biological resources and networks. Multi-dimension data can reflect various biological features, while multi-biological networks serve as the framework to capture the complex hierarchical relationships among those biological molecules. Thus, undoubtedly both properties contribute to solving biological problems. However, integrating different biological networks into a heterogeneous or multiplex network may ignore the inherent differences between those networks. In conclusion, an effective random walk-based method should treat biological networks unequally by considering different numbers of walking steps on multiple networks.

APPENDIX

SUPPLEMENTARY DATA

Table S1. Random walk-based methods based on node classification. Table S2. Random walk-based methods based on link prediction. Table S3. MCL and Random walk-based methods for disease module identification.

REFERENCES

[1] C. Simon and P. Farndon, "What causes genetic disorders?" *InnovAiT, Educ. Inspiration Gen. Pract.*, vol. 1, no. 8, pp. 544–553, Aug. 2008, doi: 10.1093/innovait/inn087.

[2] D. T. Tran and M.-T. Nguyen, "Network approaches for identification of human genetic disease genes," *Vietnam J. Sci. Technol.*, vol. 60, no. 4, pp. 700–712, Aug. 2022, doi: 10.15625/2525-2518/17026.

[3] M. D. Adams, J. M. Kelley, J. D. Gocayne, M. Dubnick, M. H. Polymeropoulos, H. Xiao, C. R. Merril, A. Wu, B. Olde, R. F. Moreno, A. R. Kerlavage, W. R. Mccombie, and J. C. Venter, "Complementary DNA sequencing: Expressed sequence tags and human genome project author," *Science*, vol. 252, no. 5013, pp. 1651–1656, 1991.

[4] G. U. Ganegoda, Y. Sheng, and J. Wang, "ProSim: A method for prioritising disease genes based on protein proximity and disease similarity," *BioMed Res. Int.*, vol. 2015, no. 5, 2015, Art. no. 213750.

[5] S. Yoon, H. C. T. Nguyen, Y. J. Yoo, J. Kim, B. Baik, S. Kim, J. Kim, S. Kim, and D. Nam, "Efficient pathway enrichment and network analysis of GWAS summary data using GSA-SNP2," *Nucleic Acids Res.*, vol. 46, no. 10, p. e60, Jun. 2018, doi: 10.1093/nar/gky175.

[6] K. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, and A. L. Barabási, "The human disease network," *Proc. Nat. Acad. Sci. USA*, vol. 104, no. 21, pp. 8685–8690, 2007. [Online]. Available: https://www.pnas.org/cgi/content/full/

[7] M. Oti and H. Brunner, "The modular nature of genetic diseases," *Clin. Genet.*, vol. 71, no. 1, pp. 1–11, Oct. 2006, doi: 10.1111/j.1399-0004.2006.00708.x.

[8] T. Ideker and R. Sharan, "Protein networks in disease," *Genome Res.*, vol. 18, no. 4, pp. 644–652, Apr. 2008, doi: 10.1101/gr.071852.107.

[9] J. Menche A. Sharma, M. Kitsak, S. D. Ghiassian, M. Vidal, J. Loscalzo, and A.-L. Barabási, "Uncovering disease-disease relationships through the incomplete interactome," *Science*, vol. 347, no. 6224, p. 841, Feb. 2015, doi: 10.1126/science.1257601.

[10] U. Stelzl and E. Wanker, "The value of high quality protein–protein interaction networks for systems biology," *Current Opinion Chem. Biol.*, vol. 10, no. 6, pp. 551–558, Dec. 2006, doi: 10.1016/j.cbpa.2006.10.005.

[11] S. K. Ata, M. Wu, Y. Fang, L. Ou-Yang, C. K. Kwoh, and X.-L. Li, "Recent advances in network-based methods for disease gene prediction," *Briefings Bioinf.*, vol. 22, no. 4, pp. 1–15, Jul. 2021, doi: 10.1093/bib/bbaa303.

[12] K. S. Grennan, C. Chen, E. S. Gershon, and C. Liu, "Molecular network analysis enhances understanding of the biology of mental disorders," *BioEssays*, vol. 36, no. 6, pp. 606–616, Jun. 2014, doi: 10.1002/bies.201300147.

[13] S. Han, J. Hong, S. J. Yun, H. J. Koo, and T. Y. Kim, "PWN: Enhanced random walk on a warped network for disease target prioritization," *BMC Bioinf.*, vol. 24, no. 1, p. 105, Mar. 2023, doi: 10.1186/s12859-023-05227-x.

[14] W. Peng, J. Wang, Z. Zhang, and F.-X. Wu, "Applications of random walk model on biological networks," *Current Bioinf.*, vol. 11, no. 2, pp. 211–220, Apr. 2016, doi: 10.2174/1574893611666160223200823.

[15] A.-L. Barabási, N. Gulbahce, and J. Loscalzo, "Network medicine: A network-based approach to human disease," *Nature Rev. Genet.*, vol. 12, no. 1, pp. 56–68, Jan. 2011, doi: 10.1038/nrg2918.

[16] M. Vidal, M. E. Cusick, and A.-L. Barabási, "Interactome networks and human disease," *Cell*, vol. 144, no. 6, pp. 986–998, Mar. 2011, doi: 10.1016/j.cell.2011.02.016.

[17] Y. Kim, J.-H. Park, and Y.-R. Cho, "Network-based approaches for disease-gene association prediction using protein-protein interaction networks," *Int. J. Mol. Sci.*, vol. 23, no. 13, p. 7411, Jul. 2022, doi: 10.3390/ijms23137411.

[18] C. Su, J. Tong, Y. Zhu, P. Cui, and F. Wang, "Network embedding in biomedical data science," *Briefings Bioinf.*, vol. 21, no. 1, pp. 182–197, Dec. 2018, doi: 10.1093/bib/bby117.

[19] X. Wang, N. Gulbahce, and H. Yu, "Network-based methods for human disease gene prediction," *Briefings Funct. Genomics*, vol. 10, no. 5, pp. 280–293, Sep. 2011, doi: 10.1093/bfgp/elr024.

[20] J. E. Shim and I. Lee, "Network-assisted approaches for human disease research," *Animal Cells Syst.*, vol. 19, no. 4, pp. 231–235, Jul. 2015, doi: 10.1080/19768354.2015.1074108.

[21] M. R. Raj and A. Sreeja, "Analysis of computational gene prioritization approaches," *Proc. Comput. Sci.*, vol. 143, pp. 395–410, Jan. 2018, doi: 10.1016/j.procs.2018.10.411.

[22] X. Yue, Z. Wang, J. Huang, S. Parthasarathy, S. Moosavinasab, Y. Huang, S. M. Lin, W. Zhang, P. Zhang, and H. Sun, "Graph embedding on biomedical networks: Methods, applications and evaluations," *Bioinformatics*, vol. 36, no. 4, pp. 1241–1251, Feb. 2020, doi: 10.1093/bioinformatics/btz718.

[23] L. Lovász, "Random walks on graphs," *Combinatorics, Paul Erdos is Eighty*, vol. 2, no. 4, pp. 1–46, Oct. 1993.

[24] N. Masuda, M. A. Porter, and R. Lambiotte, "Random walks and diffusion on networks," *Phys. Rep.*, vols. 716–717, pp. 1–58, Nov. 2017, doi: 10.1016/j.physrep.2017.07.007.

[25] E. A. Codling, M. J. Plank, and S. Benhamou, "Random walk models in biology," *J. Roy. Soc. Interface*, vol. 5, no. 25, pp. 813–834, Aug. 2008, doi: 10.1098/rsif.2008.0014.

[26] J. Zhang, "Application of random walk for disease prediction," *Highlights Sci., Eng. Technol.*, vol. 16, pp. 76–85, Nov. 2022.

[27] S. Brin, "The PageRank citation ranking: Bringing order to the web," in *Proc. ASIS*, vol. 98, 1998, pp. 161–172.

[28] K. Roshni and K. Unnikrishnan, "A review on PageRank and Personalised PageRank algorithms," *Int. Res. J. Eng. Technol.*, vol. 8, no. 4, pp. 253–257, Apr. 2021. [Online]. Available: https://www.irjet.net

[29] W. Xing and A. Ghorbani, "Weighted PageRank algorithm," in *Proc. 2nd Annu. Conf. Commun. Netw. Services Res.*, 2004, pp. 305–314.

[30] K. J. Shin, "Scalable methods for random walk with restart and tensor factorisation," B.Sc. thesis, Dept. CS Eng., Seoul Nat. Univ., Seoul, South Korea, 2015.

[31] T. H. Haveliwala, "Topic-sensitive PageRank," in *Proc. 11th Int. Conf. World Wide Web*, May 2002, pp. 517–526.

[32] S. Park, W. Lee, B. Choe, and S.-G. Lee, "A survey on personalized PageRank computation algorithms," *IEEE Access*, vol. 7, pp. 163049–163062, 2019, doi: 10.1109/ACCESS.2019.2952653.

[33] S. White and P. Smyth, "Algorithms for estimating relative importance in networks," in *Proc. 9th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2003, pp. 266–275.

[34] C. Padilla, A. Younkins, C. Broadbent, O. Smith, S. Chen, and K. Finstuen-Magro, "Network analysis algorithms for disease gene prioritization," Carleton Digital Commons, Mar. 2020. [Online]. Available: https://digitalcommons.carleton.edu/comps/2603/

[35] H. Tong, C. Faloutsos, and J.-Y. Pan, "Random walk with restart: Fast solutions and applications," *Knowl. Inf. Syst.*, vol. 14, no. 3, pp. 327–346, Mar. 2008, doi: 10.1007/s10115-007-0094-2.

[36] T. Can, O. Çamoğlu, and A. K. Singh, "Analysis of protein-protein interaction networks using random walks," in *Proc. 5th Int. Workshop Bioinf.*, Aug. 2005, pp. 61–68.

[37] K. Macropol, T. Can, and A. K. Singh, "RRW: Repeated random walks on genome-scale protein networks for local cluster discovery," *BMC Bioinf.*, vol. 10, no. 1, p. 283, Sep. 2009, doi: 10.1186/1471-2105-10-283.

[38] D.-H. Le and Y.-K. Kwon, "Neighbor-favoring weight reinforcement to improve random walk-based disease gene prioritization," *Comput. Biol. Chem.*, vol. 44, pp. 1–8, Jun. 2013, doi: 10.1016/j.compbiolchem.2013.01.001.

[39] W. Liu, X. Sun, L. Peng, L. Zhou, H. Lin, and Y. Jiang, "RWRNET: A gene regulatory network inference algorithm using random walk with restart," *Frontiers Genet.*, vol. 11, p. 591461, Sep. 2020, doi: 10.3389/fgene.2020.591461.

[40] W. Peng, J. Wang, W. Wang, Q. Liu, F.-X. Wu, and Y. Pan, "Iteration method for predicting essential proteins based on orthology and protein-protein interaction networks," *BMC Syst. Biol.*, vol. 6, no. 1, pp. 1–17, Dec. 2012. [Online]. Available: http://www.biomedcentral.com/1752-0509/6/87

[41] 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, doi: 10.1016/j.ajhg.2008.02.013.

[42] 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, Mar. 2010, doi: 10.1093/bioinformatics/btq108.

[43] X. Yao, H. Hao, Y. Li, and S. Li, "Modularity-based credible prediction of disease genes and detection of disease subtypes on the phenotype-gene heterogeneous network," *BMC Syst. Biol.*, vol. 5, no. 1, pp. 1–11, May 2011, doi: 10.1186/1752-0509-5-79.

[44] Y. Li and J. C. Patra, "Integration of multiple data sources to prioritize candidate genes using discounted rating system," *BMC Bioinf.*, vol. 11, no. S1, pp. 1–10, Jan. 2010, doi: 10.1186/1471-2105-11-S1-S20.

[45] A. M. Liekens, J. De Knijf, W. Daelemans, B. Goethals, P. De Rijk, and J. Del-Favero, "BioGraph: Unsupervised biomedical knowledge discovery via automated hypothesis generation," *Genome Biol.*, vol. 12, no. 6, p. R57, 2011, doi: 10.1186/gb-2011-12-6-r57.

[46] B. Waggoner, "Colorado CSCI 5454: Algorithms lecture 19-20—Random walks on graphs," Accessed: Oct. 20, 2023. [Online]. Available: https://www.bowaggoner.com/courses/2019/csci5454/docs/spectral.pdf

[47] Q. Zou, J. Li, C. Wang, and X. Zeng, "Approaches for recognizing disease genes based on network," *BioMed Res. Int.*, vol. 2014, pp. 1–10, 2014, doi: 10.1155/2014/416323.

[48] M. Xie, T. Hwang, and R. Kuang, "Prioritising disease genes by bi-random walk," in *Proc. 16th Pacific-Asia Conf. Adv. Knowl. Discovery Data Mining (PAKDD)*, Kuala Lumpur, Malaysia, May/Jun. 2012, pp. 292–303.

[49] D.-H. Le, "Machine learning-based approaches for disease gene prediction," *Briefings Funct. Genomics*, vol. 19, nos. 5–6, pp. 350–363, Dec. 2020, doi: 10.1093/bfgp/elaa013.

[50] S. Erten and M. Koyutürk, "Role of centrality in network based prioritisation of disease genes," in *Proc. Eur. Conf. Evol. Comput., Mach. Learn. Data Mining Bioinf.*, 2010, pp. 13–25.

[51] 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, doi: 10.1111/j.1742-4658.2012.08471.x.

[52] B. Xie, G. Agam, S. Balasubramanian, J. Xu, T. C. Gilliam, N. Maltsev, and D. Börnigen, "Disease gene prioritization using network and feature," *J. Comput. Biol.*, vol. 22, no. 4, pp. 313–323, Apr. 2015, doi: 10.1089/cmb.2015.0001.

[53] J. Chen, B. J. Aronow, and A. G. Jegga, "Disease candidate gene identification and prioritization using protein interaction networks," *BMC Bioinf.*, vol. 10, no. 1, pp. 1–14, Feb. 2009, doi: 10.1186/1471-2105-10-73.

[54] O. Vanunu, O. Magger, E. Ruppin, T. Shlomi, and R. Sharan, "Associating genes and protein complexes with disease via network propagation," *PLoS Comput. Biol.*, vol. 6, no. 1, Jan. 2010, Art. no. e1000641, doi: 10.1371/journal.pcbi.1000641.

[55] S. Wu, F. Shao, J. Ji, R. Sun, R. Dong, Y. Zhou, S. Xu, Y. Sui, and J. Hu, "Network propagation with dual flow for gene prioritization," *PLoS ONE*, vol. 10, no. 2, Feb. 2015, Art. no. e0116505, doi: 10.1371/journal.pone.0116505.

[56] S. Erten, G. Bebek, and M. Koyutürk, "Vavien: An algorithm for prioritizing candidate disease genes based on topological similarity of proteins in interaction networks," *J. Comput. Biol.*, vol. 18, no. 11, pp. 1561–1574, Nov. 2011, doi: 10.1089/cmb.2011.0154.

[57] J. Zhu, Y. Qin, T. Liu, J. Wang, and X. Zheng, "Prioritization of candidate disease genes by topological similarity between disease and protein diffusion profiles," *BMC Bioinf.*, vol. 14, no. S5, pp. 1–11, Apr. 2013, doi: 10.1186/1471-2105-14-S5-S5.

[58] F. Boizard, B. Buffin-Meyer, J. Aligon, O. Teste, J. P. Schanstra, and J. Klein, "PRYNT: A tool for prioritization of disease candidates from proteomics data using a combination of shortest-path and random walk algorithms," *Sci. Rep.*, vol. 11, no. 1, p. 5764, Mar. 2021, doi: 10.1038/s41598-021-85135-3.

[59] S. Navlakha and C. Kingsford, "The power of protein interaction networks for associating genes with diseases," *Bioinformatics*, vol. 26, no. 8, pp. 1057–1063, Feb. 2010, doi: 10.1093/bioinformatics/btq076.

[60] S. Erten, G. Bebek, R. M. Ewing, and M. Koyutürk, "DA DA: Degree-aware algorithms for network-based disease gene prioritization," *BioData Mining*, vol. 4, no. 1, pp. 1–20, Dec. 2011, doi: 10.1186/1756-0381-4-19.

[61] W. Liu, C. Li, Y. Xu, H. Yang, Q. Yao, J. Han, D. Shang, C. Zhang, F. Su, X. Li, Y. Xiao, F. Zhang, M. Dai, and X. Li, "Topologically inferring risk-active pathways toward precise cancer classification by directed random walk," *Bioinformatics*, vol. 29, no. 17, pp. 2169–2177, Sep. 2013, doi: 10.1093/bioinformatics/btt373.

[62] C. S. Seah, S. Kasim, M. F. M. Fudzee, J. M. L. T. Ping, M. S. Mohamad, R. R. Saedudin, and M. A. Ismail, "An enhanced topologically significant directed random walk in cancer classification using gene expression datasets," *Saudi J. Biol. Sci.*, vol. 24, no. 8, pp. 1828–1841, Dec. 2017, doi: 10.1016/j.sjbs.2017.11.024.

[63] H. W. Nies, M. S. Mohamad, Z. Zakaria, W. H. Chan, M. A. Remli, and Y. H. Nies, "Enhanced directed random walk for the identification of breast cancer prognostic markers from multiclass expression data," *Entropy*, vol. 23, no. 9, p. 1232, Sep. 2021, doi: 10.3390/e23091232.

[64] S. Y. Kim, H.-H. Jeong, J. Kim, J.-H. Moon, and K.-A. Sohn, "Robust pathway-based multi-omics data integration using directed random walks for survival prediction in multiple cancer studies," *Biol. Direct*, vol. 14, no. 1, pp. 1–13, Apr. 2019, doi: 10.1186/s13062-019-0239-8.

[65] J. Peng, K. Bai, X. Shang, G. Wang, H. Xue, S. Jin, L. Cheng, Y. Wang, and J. Chen, "Predicting disease-related genes using integrated biomedical networks," *BMC Genomics*, vol. 18, no. S1, pp. 1–11, Jan. 2017, doi: 10.1186/s12864-016-3263-4.

[66] P. J. Wei, F. X. Wu, J. Xia, Y. Su, J. Wang, and C. H. Zheng, "Prioritising cancer genes based on improved random walk method," *Frontiers Genet.*, vol. 11, p. 377, Apr. 2020, doi: 10.3389/fgene.2020.00377.

[67] H. Shang and Z.-P. Liu, "Prioritizing type 2 diabetes genes by weighted PageRank on bilayer heterogeneous networks," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 18, no. 1, pp. 336–346, Jan. 2021, doi: 10.1109/TCBB.2019.2917190.

[68] M. Gentili, L. Martini, M. Sponziello, and L. Becchetti, "Biological random walks: Multi-omics integration for disease gene prioritization," *Bioinformatics*, vol. 38, no. 17, pp. 4145–4152, Sep. 2022, doi: 10.1093/bioinformatics/btac446.

[69] C. S. Seah, S. Kasim, R. R. Saedudin, M. Fudzee, M. Farhan, M. S. Mohamad, R. Hassan, and M. A. Ismail, "Topologically significant directed random walk with applied Walker network in cancer environment," *Pakistan J. Pharmaceutical Sci.*, vol. 32, no. 3, pp. 1395–1408, May 2019.

[70] X. H. Tay, S. Kasim, T. Sutikno, M. F. M. Fudzee, R. Hassan, E. A. P. Akhir, N. Aziz, and C. S. Seah, "An entropy-based directed random walk for cancer classification using gene expression data based on bi-random walk on two separated networks," *Genes*, vol. 14, no. 3, p. 574, Feb. 2023, doi: 10.3390/genes14030574.

[71] A. Li, Y. Deng, Y. Tan, and M. Chen, "A novel miRNA-disease association prediction model using dual random walk with restart and space projection federated method," *PLoS ONE*, vol. 16, no. 6, Jun. 2021, Art. no. e0252971, doi: 10.1371/journal.pone0252971.

[72] L. Wang, M. Shang, Q. Dai, and P.-A. He, "Prediction of lncRNA-disease association based on a Laplace normalized random walk with restart algorithm on heterogeneous networks," *BMC Bioinf.*, vol. 23, no. 1, pp. 1–20, Jan. 2022, doi: 10.1186/s12859-021-04538-1.

[73] L. Dai, R. Zhu, J. Liu, F. Li, J. Wang, and J. Shang, "MSF-UBRW: An improved unbalanced bi-random walk method to infer human lncRNA-disease associations," *Genes*, vol. 13, no. 11, p. 2032, Nov. 2022, doi: 10.3390/genes13112032.

[74] P. Yang, X. Li, M. Wu, C.-K. Kwoh, and S.-K. Ng, "Inferring gene-phenotype associations via global protein complex network propagation," *PLoS ONE*, vol. 6, no. 7, Jul. 2011, Art. no. e21502, doi: 10.1371/journal.pone.0021502.

[75] A. Valdeolivas, L. Tichit, C. Navarro, S. Perrin, G. Odelin, N. Levy, P. Cau, E. Remy, and A. Baudot, "Random walk with restart on multiplex and heterogeneous biological networks," *Bioinformatics*, vol. 35, no. 3, pp. 497–505, Feb. 2019, doi: 10.1093/bioinformatics/bty637.

[76] Z.-Q. Zhao, G.-S. Han, Z.-G. Yu, and J. Li, "Laplacian normalization and random walk on heterogeneous networks for disease-gene prioritization," *Comput. Biol. Chem.*, vol. 57, pp. 21–28, Aug. 2015, doi: 10.1016/j.compbiolchem.2015.02.008.

[77] X. Chen, M. X. Liu, and G. Y. Yan, "Drug-target interaction prediction by random walk on the heterogeneous network," *Mol. Biosyst.*, vol. 8, no. 7, pp. 1970–1978, 2012, doi: 10.1039/c2mb00002d.

[78] Y. Liu, X. Zeng, Z. He, and Q. Zou, "Inferring MicroRNA-disease associations by random walk on a heterogeneous network with multiple data sources," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 14, no. 4, pp. 905–915, Jul. 2017, doi: 10.1109/TCBB.2016.2550432.

[79] Y. Li and J. Li, ''Disease gene identification by random walk on multigraphs merging heterogeneous genomic and phenotype data,'' *BMC Genomics*, vol. 13, no. 7, p. S27, 2012, doi: 10.1186/1471-2164-13-S7-S27.

[80] W. Zhang, X. Lei, and C. Bian, ''Identifying cancer genes by combining two-rounds RWR based on multiple biological data,'' *BMC Bioinf.*, vol. 20, no. S18, pp. 1–12, Nov. 2019, doi: 10.1186/s12859-019-3123-8.

[81] Y. Chen and R. Xu, ''Network-based gene prediction for plasmodium falciparum malaria towards genetics-based drug discovery,'' *BMC Genomics*, vol. 16, no. S7, pp. 1–9, Jun. 2015, doi: 10.1186/1471-2164-16-S7-S9.

[82] J. Qu, C.-C. Wang, S.-B. Cai, W.-D. Zhao, X.-L. Cheng, and Z. Ming, ''Biased random walk with restart on multilayer heterogeneous networks for MiRNA–disease association prediction,'' *Frontiers Genet.*, vol. 12, p. 720327, Aug. 2021, doi: 10.3389/fgene.2021.720327.

[83] W. Peng, M. Li, L. Chen, and L. Wang, ''Predicting protein functions by using unbalanced random walk algorithm on three biological networks,'' *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 14, no. 2, pp. 360–369, Mar. 2017, doi: 10.1109/TCBB.2015.2394314.

[84] W. Peng, M. Li, L. Chen, and L. Wang, ''Predicting protein functions by using unbalanced random walk algorithm on three biological networks,'' *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 14, no. 2, pp. 360–369, Mar. 2017, doi: 10.1109/TCBB.2015.2394314.

[85] X. Lei and J. Tie, ''Prediction of disease-related metabolites using bi-random walks,'' *PLoS ONE*, vol. 14, no. 11, Nov. 2019, Art. no. e0225380, doi: 10.1371/journal.pone.0225380.

[86] G.-Z. Zhang and Y.-L. Gao, ''BRWMC: Predicting lncRNA-disease associations based on bi-random walk and matrix completion on disease and lncRNA networks,'' *Comput. Biol. Chem.*, vol. 103, Apr. 2023, Art. no. 107833, doi: 10.1016/j.compbiolchem.2023.107833.

[87] B. Tripathi, S. Parthasarathy, H. Sinha, K. Raman, and B. Ravindran, ''Adapting community detection algorithms for disease module identification in heterogeneous biological networks,'' *Frontiers Genet.*, vol. 10, p. 164, Mar. 2019, doi: 10.3389/fgene.2019.00164.

[88] M. Agrawal, M. Zitnik, and J. Leskovec, ''Large-scale analysis of disease pathways in the human interactome,'' in *Proc. Biocomput.*, Jan. 2018, pp. 111–122. [Online]. Available: https://www.worldscientific.com

[89] D. Buzzao, M. Castresana-Aguirre, D. Guala, and E. L. L. Sonnhammer, ''TOPAS, a network-based approach to detect disease modules in a top-down fashion,'' *NAR Genomics Bioinf.*, vol. 4, no. 4, p. Iqac093, Oct. 2022, doi: 10.1093/nargab/lqac093.

[90] R.-S. Wang and J. Loscalzo, ''Network-based disease module discovery by a novel seed connector algorithm with pathobiological implications,'' *J. Mol. Biol.*, vol. 430, no. 18, pp. 2939–2950, Sep. 2018, doi: 10.1016/j.jmb.2018.05.016.

[91] L. Liang, V. Chen, K. Zhu, X. Fan, X. Lu, and S. Lu, ''Integrating data and knowledge to identify functional modules of genes: A multilayer approach,'' *BMC Bioinf.*, vol. 20, no. 1, pp. 1–15, May 2019, doi: 10.1186/s12859-019-2800-y.

[92] S. M. Van Dongen, ''Graph clustering by flow simulation,'' Ph.D. dissertation, Center Math. CS, Utrecht Univ., The Netherlands, 2000.

[93] S. Srihari, K. Ning, and H. W. Leong, ''MCL-CAw: A refinement of MCL for detecting yeast complexes from weighted PPI networks by incorporating core-attachment structure,'' *BMC Bioinf.*, vol. 11, no. 1, pp. 1–25, Oct. 2010, doi: 10.1186/1471-2105-11-504.

[94] Y.-K. Shih and S. Parthasarathy, ''Identifying functional modules in interaction networks through overlapping Markov clustering,'' *Bioinformatics*, vol. 28, no. 18, pp. i473–i479, Sep. 2012, doi: 10.1093/bioinformatics/bts370.

[95] P. J. Ochieng, W. A. Kusuma, and T. Haryanto, ''Detection of protein complex from protein-protein interaction network using Markov clustering,'' *J. Phys., Conf. Ser.*, vol. 835, May 2017, Art. no. 012001, doi: 10.1088/1742-6596/835/1/012001.

[96] O. Maruyama and A. Chihara, ''NWE: node-weighted expansion for protein complex prediction using random walk distances,'' *Proteome Sci.*, vol. 9, no. 1, p. S14, 2011, doi: 10.1186/1477-5956-9-S1-S14.

[97] K. Voevodski, S.-H. Teng, and Y. Xia, ''Finding local communities in protein networks,'' *BMC Bioinf.*, vol. 10, no. 1, p. 297, Sep. 2009, doi: 10.1186/1471-2105-10-297.

[98] D. A. Spielman and S.-H. Teng, ''A local clustering algorithm for massive graphs and its application to nearly linear time graph partitioning,'' *SIAM J. Comput.*, vol. 42, no. 1, pp. 1–26, Jan. 2013, doi: 10.1137/080744888.

[99] R. Andersen, F. Chung, and K. Lang, ''Local graph partitioning using PageRank vectors,'' in *Proc. 47th Annu. IEEE Symp. Found. Comput. Sci. (FOCS)*, 2006, pp. 475–486.

[100] 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, doi: 10.1109/TCBB.2014.2343954.

[101] D. Petrochilos, A. Shojaie, J. Gennari, and N. Abernethy, ''Using random walks to identify cancer-associated modules in expression data,'' *BioData Mining*, vol. 6, no. 1, pp. 1–25, Dec. 2013, doi: 10.1186/1756-0381-6-17.

[102] Y. Zhang, J. Zhang, Z. Liu, Y. Liu, and S. Tuo, ''A network-based approach to identify disease-associated gene modules through integrating DNA methylation and gene expression,'' *Biochem. Biophys. Res. Commun.*, vol. 465, no. 3, pp. 437–442, Sep. 2015, doi: 10.1016/j.bbrc.2015.08.033.

[103] R. Ahmed, I. Baali, C. Erten, E. Hoxha, and H. Kazan, ''MEXCOwalk: Mutual exclusion and coverage based random walk to identify cancer modules,'' *Bioinformatics*, vol. 36, no. 3, pp. 872–879, Feb. 2020, doi: 10.1093/bioinformatics/btz655.

[104] S. S. Boyd, C. Slawson, and J. A. Thompson, ''AMEND: Active module identification using experimental data and network diffusion,'' *BMC Bioinf.*, vol. 24, no. 1, pp. 1–21, Jul. 2023, doi: 10.1186/s12859-023-05376-z.

[105] T. Y. Berger-Wolf, T. M. Przytycka, M. Singh, and D. K. Slonim, ''Dynamics of biological networks—Session introduction,'' in *Proc. Pacific Symp. Biocomput.*, vol. 15, Dec. 2010, pp. 120–122.

[106] G. H. Nguyen, J. B. Lee, R. A. Rossi, N. Ahmed, E. Koh, and S. Kim, ''Dynamic network embeddings: From random walks to temporal random walks,'' in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2018, pp. 1085–1092.

[107] J. Wang, X. Peng, W. Peng, and F. Wu, ''Dynamic protein interaction network construction and applications,'' *PROTEOMICS*, vol. 14, nos. 4–5, pp. 338–352, Mar. 2014, doi: 10.1002/pmic.201300257.

[108] M. Li, X. Wu, J. Wang, and Y. Pan. (2012). *Towards the Identification of Protein Complexes and Functional Modules by Integrating PPI Network and Gene Expression Data*. [Online]. Available: http://www.biomedcentral.com/1471-2105/13/109http://netlab.csu.edu.cn/bioinfomatics/limin/DFM-CIN/index.html.http://www.biomedcentral.com/1471-2105/13/109

[109] S.-Y. Sun, Z.-P. Liu, T. Zeng, Y. Wang, and L. Chen, ''Spatio-temporal analysis of type 2 diabetes mellitus based on differential expression networks,'' *Sci. Rep.*, vol. 3, no. 1, p. 2268, Jul. 2013, doi: 10.1038/srep02268.

[110] X. Tang, J. Wang, B. Liu, M. Li, G. Chen, and Y. Pan, ''A comparison of the functional modules identified from time course and static PPI network data,'' *BMC Bioinf.*, vol. 12, no. 1, pp. 1–15, Aug. 2011, doi: 10.1186/1471-2105-12-339.

[111] Q. Xiao, J. Wang, X. Peng, and F.-X. Wu, ''Detecting protein complexes from active protein interaction networks constructed with dynamic gene expression profiles,'' *Proteome Sci.*, vol. 11, no. S1, pp. 1–8, Dec. 2013, doi: 10.1186/1477-5956-11-S1-S20.

[112] J. Wang, X. Peng, M. Li, and Y. Pan, ''Construction and application of dynamic protein interaction network based on time course gene expression data,'' *PROTEOMICS*, vol. 13, no. 2, pp. 301–312, Jan. 2013, doi: 10.1002/pmic.201200277.

TAY XIN HUI received the M.Sc. degree in computer science and information technology (information security) from the Tun Hussein Onn University of Malaysia. She has been a Research Assistant with the Faculty of Computer Science and Information Technology, UTHM, since 2020. Her research interests include data mining, bioinformatics, and machine learning algorithms.

SHAHREEN KASIM is currently an Associate Professor with the Department of Security Information and Web Technology, Faculty of Computer Science and Information Technology, Tun Hussein Onn University of Malaysia. Her research interests include bioinformatics, soft computing, data mining, web, and mobile application.

MOHD FARHAN MD. FUDZEE (Senior Member, IEEE) is currently a Professor with the Faculty of Computer Science and Information Technology, Tun Hussein Onn University of Malaysia, and also a Principal Researcher with the Advance Research on Multimedia and Applications (AROMA) Focus Group. His research/technical interests include information systems, multimedia computing and applications, and ICT governance.

TOLE SUTIKNO (Member, IEEE) is currently an Associate Professor with the Electrical Engineering Department, Universitas Ahmad Dahlan (UAD), Yogyakarta, Indonesia. His research interests include digital design, industrial applications, industrial electronics, industrial informatics, power electronics, motor drives, renewable energy, FPGA applications, embedded systems, robotics and automation, artificial intelligence, intelligent systems, information systems, and digital libraries.

ROHAYANTI HASSAN is currently an Associate Professor with the Faculty of Electrical Engineering, University of Technology Malaysia (UTM), where she is also the Head of the Software Engineering Research Group. Her research interests include requirements analysis process, multi stakeholder requirements, requirements ambiguities, mutation testing, and fuzzy logic.

IZZATDIN ABDUL AZIZ is currently a Researcher with the High-Performance Cloud Computing Centre (HPC3), Universiti Teknologi PETRONAS (UTP), where he focuses on solving complex upstream oil and gas (O&G) industry problems from the viewpoint of computer sciences. He also serves as the Deputy Head of the Computer and Information Sciences Department, UTP.

MOHD HILMI HASAN is currently a Senior Lecturer with Universiti Teknologi PETRONAS, Malaysia, where he is also an Active Researcher with the Centre for Research in Data Science. His research interests include artificial intelligence and data analytics. He has published research articles in various journals and conferences, two of which are in the reputable *Artificial Intelligence Review* journal.

JAFREEZAL JAAFAR (Senior Member, IEEE) is currently an Associate Professor and the Dean of the Faculty of Science and Information Technology, Universiti Teknologi PETRONAS (UTP). His research interests include AI, machine learning, and data analytics. He was a member of the Academy Science Malaysia SIG in machine learning, a Senior Member, and an Executive Committee Member of IEEE CS Malaysia Chapter, from 2016 to 2018, and MyAIS, from 2017 to 2019. He has been awarded as the Professional Technology (P.Tech.) by the Malaysia Board of Technology.

METAB ALHARBI is currently a Lecturer with the Department of Pharmacology and Toxicology, King Saud University, where he became the Deputy of the Department. He is also the Chairperson of the Founding Committee of the Executive Master in Drug Regulatory Affairs. He also involves in establishing the Regulatory Affairs simulation Laboratory, which was in partnership between KSU and Saudi Food and Drug Authority (SFDA).

SEAH CHOON SEN is currently an Associate Professor with the Faculty of Accounting and Management, Universiti Tunku Abdul Rahman. His research interests include data science, digital entrepreneurship, financial technology, and precision farming and information systems.