IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 19, 2020, accepted April 6, 2020, date of publication April 10, 2020, date of current version April 29, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2987071

Disease-Pathway Association Prediction Based on Random Walks With Restart and PageRank

ALI GHULAM, XIUJUAN LEI, (Member, IEEE), MIN GUO, AND CHEN BIAN
School of Computer Science, Shaanxi Normal University, Xi'an 710119, China

Corresponding author: Xiujuan Lei (xjlei@snnu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61972451, Grant 61672334, and Grant 61902230. and in part by the Fundamental Research Funds for the Central Universities under Grant GK201901010.

ABSTRACT The study of disease–pathway association in human diseases is a perennial focus of the biomedical field. The association of diseases and pathways can help in the discovery of the mechanisms or relationships of human diseases. The accuracy of disease identification has been less than satisfactory despite decades of research in this area. Therefore, this study proposes a computational model for the prediction of disease–pathway associations. The proposed computational model is based on Random Walk with Restart on heterogeneous network (RWRH) and PageRank. The RWRH disease–pathway association model is a novel computational model that can predict potential disease–pathway associations. Furthermore, the model can help pathologists understand the correlations among disease–pathway associations, treatments, and reactions. We performed a pathway-based study to expand disease variation relationships and to find new molecular correlations between genetic mutations. We constructed a biological network on the basis of shared gene interactions of disease–pathways and attempted to investigate the pathogenesis of a disease by analyzing the constructed network. The network construction was based on two parts. First, the similarity between pathway–pathway networks was calculated. Second, a disease–disease (DD) similarity network was constructed, and the correlation between disease and disease similarity was calculated. We also investigated the pathway seed node and disease seed node with high PageRank. Moreover, we focused on mining the complexity of disease–pathway associations. We used the bipartite network of disease–pathway associations to combine the obtained biological information, which was based on the pair similarity of sequence expression weights. These weights, which were obtained by using the multilayer resource-allocation algorithm, were used to calculate the prediction scores of each disease–pathway pair. Here, through leave-one-out cross-validation, we examined a 210 × 1855 matrix, with the 210 rows representing diseases and 1855 columns indicating pathways. The disease–pathway adjacency matrix contained 13,838 known disease–pathway associations. The best predictive results achieved an area-under-the-curve value of 0.8218 and a two-class precision–recall curve. These results indicate that our method has higher scientific performance than previously proposed methods. We predicted pathogen, DD, and disease–pathway relationships by comparing them with known associations and through publication search. We then proposed the possible reasons for our predictions.

INDEX TERMS Pathway similarity network, disease similarity network, disease pathway association, PageRank algorithm.

I. INTRODUCTION

In recent years, researchers have become interested in disease–pathway associations. The existing literature on the comparative toxicogenomic database (CTD) [1] suggests that genes, diseases, pathways, and chemical datasets can be used to construct disease–pathway association networks. Sets of proteins that are associated with a given disease enable the construction of pathway–pathway networks with the use of Gene Ontology (GO) datasets [2]. This approach determines which genes in the GO datasets overlap via two or more different or similar pathways. The present

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

research focused on disease–pathways associations, which can be defined as similarities between diseases and pathways. Disease diagnosis, prognosis, and treatment play a vital role in clinical studies [3]. To date, however, scant attention has been paid to network-based approaches based on the Random Walk with Restart (RWR) algorithm. Additionally, previous research has largely overlooked the importance of disease–pathway relationships in identifying new cancer disease-related approaches [4]. New approaches, including network-based approaches, have been created to integrate different ''genomic'' information, such as gene expression, genome-wide association studies (GWAS), and disease–pathway association networks, to describe these difficulties. The present study surveyed subsequent methodological improvements to identify disordered connections, disease–pathway associations, priority, and the enrichment of candidate disease genes in pathways [5]. Complex diseases are mediated by multiple natural factors rather than by a single gene. Recognizable evidence for the affected pathway reveals insights into infection improvement systems. Additionally, GWAS has been aggregated for pathways [6]. The genes of the same biological pathways are assumed to promote diseases. Moreover, every normal variety of these genes contributes to the risk of disease [7]. The combination of heredity and molecular biology results in drastic simplification [8], [9]. In recent years, researchers have become increasingly interested in human genome sequencing, GWAS, transcriptomics, and proteomics; disease gene tracking has also received considerable research attention [10]. Current research indicates that understanding the unique methods for identifying infections enables us to comprehend human diseases [11], [12]. Several computing strategies for integrating complex and heterogeneous information datasets, which include information expression, sequential data, pathway function annotations, and biomedical scientific literature, help in prioritizing future research approaches in a straightforward manner [13]. *Li et al.* set up a unique gene network that can anticipate various information sources and consolidate networks into a single network that is described by multigraphs [14]. The existing literature emphasizes that disease genes, which are associated with genetic diseases, have enhanced homeopathic care and improved the understanding of human gene associations and pathways. A method for prioritizing candidate disease genes based on using global network distance measure and random walk analysis was also presented [15]. The prediction of the relationship between genetic diseases and pathogenic genes is a key challenge related to human health. The latest computing methods help scientists in improved inference of these correlations by understanding genetic interactions in human pathways in GWAS [16]. Numerous scholars have acknowledged that recognition of disease genes through differentiation is necessary to explain the relationship between genes and diseases.

The present study aimed to investigate how pathogenic genes are predicted with the help of the integrated phenotypic and genomic data. Several genetic diseases are similar in terms of either heredity or phenotype. Researchers have extended RWR algorithms to heterogeneous networks [17]. The cross-validation method has also been applied to assess the capability of these algorithms to detect phenotypic relationships. García-Campos Miguel A *et al.* (2017) established a link between pathways in a biological network containing natural interactions with biomolecules, such as proteins, nucleotides, or genes. The available pathway nodes, also known as pathway elements, might be linked to various pathways, and their functions may differ in varied cellular contexts [18]. We emphasize the scarcity of research articles in computer biology and bioinformational research showing substantial performance gains compared with established algorithms through the use of competitions. These algorithms include those used to create multiple-edge graphics models for several biological networks, such as gene-cutting associations, disease–gene (DG) associations, disease–pathway associations, and gene ontological annotations. Boundaries are weighted by form, reliability, and computerization [19]. Complex improvements in cellular machinery must consider diseases. Cell gene expression profiles exhibit characteristic patterns related to diseases. Using these profiles, we can obtain new biological knowledge of a disease, improve diagnostic capacity, and determine disease risks. Researchers also seek to reveal disease–disease (DD) correlations, disease–pathway associations, and DG associations based on data and multiscale cellular organizations. Thomas Gaudelet *et al.* [20] proposed neural networks with structures based on multiple protein organizations in protein complexes and pathways in a cell. Such analysis can accurately predict the diagnosis of the majority of patients. By studying trained models, we predicted the relationships of pathogens, diseases, and diseases–pathways by comparing them with known associations and publication search. Moreover, we proposed possible reasons for our predictions. The discovery of disease pathways in the form of protein sets related to a particular disorder is an important issue that can potentially provide clinically useful insights into the diagnosis, prognosis, and treatment of illnesses. Computer methods support discovery through the use of DD, DG, and disease–pathway networks. We then analyzed cutting-edge methods for disease discovery and showed poor performance in diseases with disconnected pathways. Thus, the connectivity structure of a network alone may be insufficient for disease detection. We also demonstrated a promising pathway for the development of new methods, such as small subgraphs, for high-order network structures [2]. Finding real-life disease markers or signatures is critical for the successful diagnosis, treatment, and prognosis of complex disorders, such as pathway cancers. Thus far, various experiments have been conducted to classify markers for diseases with a number of biological sources such as pathway databases for genetic expression profiles [21]. In the last few years, computational methods have been increasingly used for the prediction of possible disease–pathway associations based on their hidden relationships, guidance, or effectiveness.

The present work adopted three biological networks that contain pathway–pathway similarity networks and measured and applied the RWR algorithm. In this study, a DD network was constructed, and the RWR algorithm was implemented to calculate the RWR score vector with respect to (*w.r.t.*) disease seed node. A bipartite RWR heterogenous network for disease–pathway association (RWRHPDA) was developed for inferring potential disease–pathway interactions. The review of the literature shows that biological networks play a crucial role in finding genes and genomic modules that lead to diseases. The database provides manually curated data that were extracted from scientific literature. The present study proposed a novel approach for disease–pathway association prediction that is based on RWR (RWRHPDA) and PageRank. This work helps in understanding the design of experiments and the utilization of human disease and human pathway data. In this study, we analyzed and used data in different ways to establish a novel approach for disease–pathway association prediction.

Additionally, this study reviewed critical challenges related to the RWR and the usage of Random Walk on heterogeneous networks in bioinformatics. The present study aimed to clarify the relationship among human diseases from a pathway perspective. Moreover, this study attempted to predict anomalous pathways. Empirical evidence confirms the notion that genes in cells cannot function alone. Inspired by currently famous the particular nature of network-based approach in biology. A novel method for identifying potential disease-pathway association via heterogeneous network. The analysis was based on RWR algorithm, which is very famous for the links analysis. This Random Walk with Restart on Heterogeneous network (RWRHPDA) algorithm ranks the disease and pathways simultaneously. The analysis was based on RWR algorithm, which is very famous for the links analysis. RWRHPDA algorithm prioritizes the disease and pathways instantaneously. The RWRHPDA model is a novel computational model, and it can help predict the potential disease-pathway association. Our technique achieves a zone under the receiver operating characteristic curve (AUC) of 0.82, which is roughly higher than that of the state-of - the-art methods. Finally, we use various data sets for our algorithm to further prove the validity.

II. MATERIALS AND METHODS

A. DATA SOURCE

The subjects of this study consisted of pathway datasets that are included in the Kyoto Encyclopedia of Genes and Genomes (KEGG) database and the Reactome pathway database. These databases are open-source and provide a platform for computation across biological responses in networks selected to obtain predefined human biological pathways [22]. The data collected for the present study included 25,399 pathways. Moreover, the present study involved 25,392 unknown association datasets, which contained pathways, diseases, and related genes. The datasets were then normalized, and 1855 pathways and 210 diseases were determined. The data collected were mostly qualitative/quantitative in nature, and their detailed information is given in Table 1. The disease datasets were obtained from the Online Mendelian Inheritance in Man (OMIM) database and measured on the basis of disease similarity [23] and CTD [24] (Table 2). Additional data were gathered from manually curated databases of disease-related genes, and disease association among pathways was mapped in terms of pathways and genes.

TABLE 1. Data source.

Data Source	Website
CTD	http://ctdbase.org/
KEGG	https://www.genome.jp/kegg/pathway.html
Reactome	https://reactome.org/
GeneRIF	http://www.ncbi.nlm.nih.gov/gene/about-generif
GAD	https://geneticassociationdb.nih.gov/
OMIM	https://www.omim.org/downloads/
MESH	https://www.ncbi.nlm.nih.gov/mesh

TABLE 2. Types of download datasets.

Data Source	Datasets Name	Total Quality	Pre-Process	Used Quality
Download from CTD	Disease pathway association.	557406	25399	
KEGG	Pathways			382
Reactome	Pathways			1473
Used in experiments pathway				1855
OMIM	Disease			25
MESH	Disease			185
Used in experiments disease				210
Genes	Genes	557406	25399	25399
			Duplicate	24543
Used in experiments genes				856

B. METHOD

Figure 1 shows the flowchart of our framework, which is based on different phases and contains disease–pathway associations. The datasets were first selected. Subsequently, the data were preprocessed and standardized for the experiments. The collected datasets contained pathways related to diseases and 25,399 human gene disease and pathway aliases. Additionally, we collected 185 mesh samples and human disease related to 25 OMIM sample terms. The link prediction problem was prioritized in the disease–pathway association. In this study, heterogeneous networks were utilized; these networks consisted of three main parts: pathway–pathway interaction network [25], DD similarity. network, and disease–pathway association network. The potential disease pathway association can be regarded as the missing link in the disease pathway association network. Therefore,

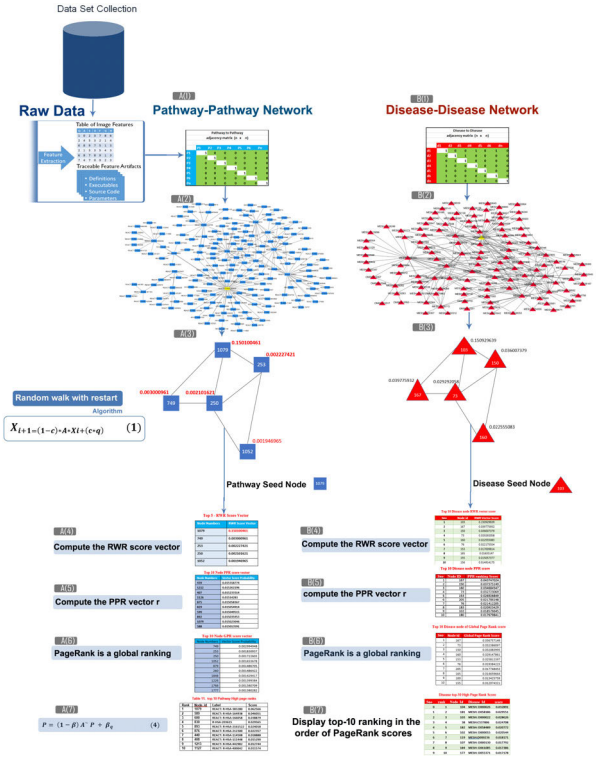

FIGURE 1. Flowchart of framework model.A(1) Creation of the pathway–pathway adjacency matrix, which contains the information on a given network with n nodes denoted by Pi(n*n) [26]. A(2) Construction of the pathway–pathway similarity network. Pathway–pathway functional similarity scores were calculated on the basis of asymmetric matrix similarity scores. A(3) Specified edge connections and detected seed pathway nodes. The single seed in RWR vector scores was computed, and the weight of each node was measured. The top 10 seed pathways with adjacent neighbor pathway nodes were detected. A(4) Top 5 RWR vector score probabilities. A(5) Personalized PageRank (PPR) vector scores of the top 10 seed pathways. A(6) Global PageRank (GPR) vector scores of the top 10 pathway nodes. A(7) Top 10 pathway nodes with high PageRank scores. B(1) Creation of the DD adjacency matrix, which contains the information of a given network with n nodes denoted by di (n*n) [26]. B(2) DD similarity network. DD functional similarity scores were calculated by using the asymmetric matrix similarity score. B(3) Specified edge connections and detected seed disease nodes. The single seed in RWR vector was computed, and the weight of each node was measured. The top 10 seed diseases with adjacent neighbor disease nodes were also identified. B(4) Top 5 RWR vector score probabilities. B(5) Top 10 disease nodes and PPR vector scores. B(6) GPR vector scores of the top 10 disease nodes. B (7) Top 10 disease nodes with high PageRank.

the present study aimed to predict such links and provide other raw representations of heterogeneous networks and nodes (diseases and pathways). The theories proposed in the present study considered disease and pathway relationships on the basis of shared genes and pathway–pathway interactions. Moreover, this study attempted to investigate the authentic datasets for the disease–pathway network because according to the existing research, pathway–pathway networks are constructed by utilizing GO datasets. This work aimed to create a disease pathway network that relies on diseases and pathway datasets only. The review of the literature shows that each disease has one pathway or occasionally more than one pathway. Here, d denotes disease, $\boldsymbol{d = \{p_1, p_2, p_3, \ldots\ldots p_n\}}$; $\boldsymbol{p_1, p_2}$ denotes pathways; $\boldsymbol{p_n}$ represents the total number of pathway sets.

1) IMPLEMENTATION OF THE RWR ALGORITHM

The present study utilized the RWR algorithm, which is widely used for link investigation and offers node-to-node proximities in arbitrary types of graphs (networks). The typical applications of the RWR algorithm include various real-world graph mining tasks, such as personalized node ranking, recommendations in graphs (***e.g.,*** "whom you may know"), and anomaly detection. Pyrwr aims to implement algorithms for computing RWR scores that are based on **power iteration**, which is achieved by using numpy and scipy in Python. Pyrwr focuses on computing a single-source RWR score vector *w.r.t.* a given query (seed) node, which is used for personalized node ranking *w.r.t.* the querying node. In addition to RWR, Pyrwr supports that computation of PPR and PageRank, which are well-known variants of RWR.

The features of Pyrwr are given as follows:

Query-type features

- **RWR**: Personalized ranking; only a single seed is allowed
- **PPR**: Personalized ranking; multiple seeds are allowed
- **PageRank**: Global ranking; all nodes are used as seeds

Graph-type features

- Unweighted/weighted graphs
- Directed graphs

In this study, we have implemented Random Walk-based ranking models, such as **PageRank** and **RWR**, and attempted to investigate real-world networks using these ranking models. The installation of Python modules required by this package is named as Pyrwr.

2) QUERY-TYPE FEATURE PARAMETERS OF RWR

The description of the query-type features of a RWR is used for a single seed. This description computes a RWR score vector *w.r.t.* the seed node given by seeds in the given graph that is specified by the ***input-path*** and written as the vector into the target file in the ***output-path***. The query-type specifies the type of query, ***i.e.,*** a **RWR** query. The ***seed*** should be ***int.*** Furthermore, the format of the file at the ***input-graph*** should follow one of the input formats described above, whereas ***r*** is a column vector (***ndarray***) with the RWR score vector *w.r.t.* ***seed*** node. The shape of ***r*** is (***n, 1***), where ***n*** denotes the number of nodes. The subjects of the present study are based on the following formula, which contains the input graph for the RWR score vector:

$$X = (1 - C)^* (A.dot\ (oldx)) + \left(c^* q\right) \tag{1}$$

This can be written as follows:

$$X_{i+1} = (1 - c)^* A^* X_i + \left(c^* q\right) \tag{2}$$

Equations **1** and **2** measure the **RWR score vector** *w.r.t.* the **seed** node. This study computed the RWR score vector *w.r.t.*

the seed node; **input seed** as **int. c = 0.15** was denoted as the float value for restart probability; **epsilon = 1e-6** was used for error tolerance for power iteration. Then, **max_iters = 100** was used as int type for the maximum number of iterations for power iteration, that is, **r ndarray RWR score vector.** We performed the row-normalization of the given matrix and computed the row-normalized adjacency matrix by **nA = invD** * **A.** An adjacency matrix was used as an input matrix (it should be row-normalized for RWR and its variants), with **q** as node array shape into **n*n** and **X** as the input graph.

3) QUERY-TYPE FEATURE PARAMETERS OF PPR

In this case, **seeds** are the list of **seeds,** and **r** is the **PPR** score vector *w.r.t.* seeds. Notably, the **PPR** vector **r** is used to obtain the personalized node ranking list, which is related to all seeds in the seed list. In this study, **PPR** was divided into personalized rankings, and multiple seeds were allowed. The **PPR** was considered the personalized node ranking given as follows:

$$r = ppr.compute\ (seeds, c, epsilon, max\_iters) \tag{3}$$

The PPR score vector w.r.t. the seed node was computed by using the above equation. We identified key seeds as the (**ppr**) file path of **seeds(str)** or list of **seeds(list),** where **c** is used as a common query type of restart probability (**rwr**) or jumping probability (otherwise). We computed **epsilon,** which is also used as common query type as counted error tolerance for power iteration. **max-iters** is the common query type used for the maximum number of iterations for power iteration.

4) QUERY-TYPE FEATURE PARAMETERS OF pAGERANK

Notably, specification of **seeds** for PageRank was unnecessary because it is a global-ranking algorithm. Therefore, this algorithm automatically sets the required seeds (***i.e.,*** all nodes are used as **seeds**).

$$r = pagerank.compute\ (c, epsilon, max\_iters) \tag{4}$$

The above equation helps in the computation of the PageRank score vector (**global ranking**). The present study used data, which included several variables, to compute the pathway rank or disease PageRank query. Hence, the specification of seeds was not required. Dead-end nodes with an out-degree of zero might exist in directed graphs. In this case, the original power iteration, would incur leaked out scores. Thus, handle dead-end exists for such an issue to handle dead-end nodes. Handle dead-end can ensure that the sum of score vectors is **1.** Otherwise, the sum would be less than **1** in directed graphs. The strategy exploited by **Pyrwr** is that whenever a random surfer visits a dead-end node, it returns to a seed node (or one of the seed nodes) and restarts.

C. LINK ANALYSIS AND PAGERANK METHOD

This study used the most common methods to analyze qualitative data and several techniques for graph link analysis. All analyses were conducted by utilizing several link analyses (**or ranking**) models, such as **PageRank,** topic-specific **PageRank**, and hyperlink-induced topic search (**HITS**). The present study mainly focused on the implementation of the algorithms of these ranking models and the ranking of nodes in real-world graphs using these models. In this section, different techniques for graph link analysis are presented. This section contains different subsections, with each providing a brief discussion of **PageRank** and **HITS.** The primary purpose is to briefly explain the implementation of PageRank based on the dense matrices obtained by using numpy in Python. Then, **PageRank** implementation is verified.

1) REVIEW OF PAGERANK

The current section provides information on the present hypothesis of PageRank and its equation, which is required for the implementation of PageRank in Python. The mathematical notations used for PageRank are given as follows:

Scalar: lower care and regular face (α, β *and* γ) **Set**: upper case and regular face (A, B *and* C) and **Vector** and **matrix**: smaller case for vectors and upper case for matrices and bold face (χ, A *and* γ).

Aij(i, j)*- th element of matrix A*; **Number domain:** blackboard bold; $\mathbb{R}$: the set of real numbers, $\mathbb{R}^{n}$; **n-dimensional** space in real numbers: set of n-dimensional vectors, ***i.e.,*** $\chi \in \mathbb{R}^{\mathbf{n}}$**;** $\mathbb{R}^{\mathbf{n\times m}}$**: n ×m-dimensional** space in real numbers; set of **n×m-** dimensional vectors, ***i.e.,*** $A \in \mathbb{R}^{\mathbf{n\times m}}$, after presenting the problem definition of PageRank.

2) MATHEMATICAL DEFINITION OF PAGERANK

Problem definition of PageRank:

Input: Adjacency matrix $A = \mathbb{R}^{n\times m}$ of a graph, **G= (V, E)** and teleport probability β.

V is the set of nodes.

E is the set of edges.

n is the number of nodes in the graph G, *i.e.*, n = **|V|.**

m is the number of edges in the graph G, *i.e.*, n = **|E|.**

Output: The PageRank score vector p $\in \mathbb{R}^{n}$ such that: The Problem of PageRank is explored extensively in the literature.

$$P = (1 - \beta)\ A^{\sim} P + \beta_q \tag{5}$$

The above equation is called the PageRank equation, and the notations used in the PageRank equation are described below:

- $\mathbf{A}^{\sim}$ is the matrix row-normalization adjacency matrix of the graph **G,** ***i.e.,*** the sum of each row of $\mathbf{A}^{\sim}$ should be 1.

$$q = \left[\frac{1}{n}\right] n$$

- n-dimensional vector whose entry is **1/n.** This vector is usually called as the query vector.....The PageRank score of nodes **u** is denoted by **Pu**, which indicates the importance of node **u** in graph **G**. The definitions of the adjacency matrix **A** and row-normalized adjacency matrix $\mathbf{A}^{\sim}$ are given as follows:
- Definition of the adjacency **matrix A** for each edge $\mathbf{u} \longrightarrow \mathbf{v}$ **in E,** $\mathbf{A_{uv}}$**= 1.** Otherwise, $\mathbf{A_{uv}}$**= 0** (*i.e.,* if no

TABLE 3. Pseuudo-code of pagerank algorithm.

Algorithm 1:	An iterative algorithm for PageRank
Input:	Row-normalized adjacency matrix $A^{\sim}$, teleport Probability β and error tolerance e
Output:	PageRank score vector P
1:	Set query vector q *to* $\frac{1}{n}-1$
2:	Initialize PageRank vector P' at the previous step to 0
3:	**Repeat**
	$p \leftarrow (1-\beta)\,\mathrm{A}^{\sim} tP' + \beta q$
	Compute residual δ //p-p//1
	Update residual $\delta < \epsilon$
	Return PageRank score vector P

edge exists between two nodes, then, the value of the adjacency matrix will be zero).

- Definition of the row-normalized adjacency matrix $\mathbf{A}^{\sim}$. For each edge:

$$u \longrightarrow vinE, \mathrm{A_{uv}} = A_{uv} = \frac{1}{|ou|}, \quad where, Auv = ou$$

is the set of out-neighbor of node **u**. Let **D** be a diagonal matrix where the $\mathbf{u^{th}}$ diagonal entry is $|\mathbf{O_u}|$ (***i.e.,*** the out-degree of node **u**). Then, the row-normalized adjacency matrix $\mathbf{A}^{\sim}$ can be obtained by using the following equation, where D^{-1} is the inverse of **D** (***i.e.*, the $\mathbf{u^{th}}$** diagonal entry of $\mathbf{D^{-1}}$ is $|\mathbf{O_u}|$).

$$\mathbf{A}^{\sim}\mathbf{D}^{-1}\mathbf{A} \tag{6}$$

3) ITERATIVE ALGORITHM FOR PAGERANK

The PageRank score vector **P** is obtained by iteratively computing the PageRank equation. The pseudocode shown in Table 3 represents the iterative algorithm for PageRank.

D. PATHWAY-PATHWAY NETWORK CONSTRUCTION

The correlation between two pathways was calculated to evaluate shared genes [27]. We reviewed findings from recent studies that included pathway data and the pathways associated with genes. Then, as shown in Figure 2 (Supplementary Note 4), we constructed a pathway–pathway network. In the present study, each node of the constructed pathway network represents a pathway, whereas edges represent the genes. A positive correlation was obtained between the pathway and genes. Two pathways are connected if they involve the same genes. Figure 2 shows a part of the known pathway–pathway association network (Supplementary Note 4). The blue rectangular nodes represent a pathway, whereas black lines represent edges that are connected to pathway-to-pathway nodes. We have used the input data in CSV format. Then, we calculated the similarity from the pathway–pathway network in accordance with the RWR score vector *w.r.t* the pathway node relationship. The input path specifies the given input graph, and the seed node or seed path node (specified by the seeds) is used to write the vector in the target file in the output path. The query type specifies the type of query, for example, a **RWR** query. The specific formats of the input and output files were further explained. In accordance with the **RWR** method, the calculation started from the **0-initial node**. The pathway node sequence was then obtained through the iterative calculation of the transfer matrix. The similarity among node-to-node pathways was examined using mutual information with the significant pair in the pathway–pathway network. In the present study, the input was the similarity matrix of nodes (between two networks). This technique considers the pathway and disease data and the semantic similarity between pathways. Thus, strategic avoidance from the node was well maintained. We set **P= {p1, p2, …., pn},** where the **n** node represents the pathway network. $\mathbf{N^*N}$ represents the rows and column at the same dimension. As per the similarity relationship between two pathways nodes, the adjacency matrix **A (i, j)** can be defined as similar matrix $\mathbf{sim_{pi,pj}}$ corresponding to a threshold θ, which is given as follows:

$$A\,(i,j) = \begin{cases} 1 \\ 0 \end{cases} ifsim\,(pi,pj) \geq^{\theta}_{others} \tag{7}$$

The analysis was based on the pathway set **P,** which indicates the mutual information of pathways and diseases. The row rank of **p** was used as the starting point. The description states that **A** is a symmetric matrix, and graph theory states that the degree of the $\mathbf{i^{th}}$ node is equal to the sum of the elements of row **i** of the matrix:

$$A_{i,j} = a_{(i,j)} = a_{(j,i)} = A_{i,j} \tag{8}$$

The pathway–pathway network information comprises nodes and edges, which are utilized to deal with data analysis effectively.

E. DD NETWORK CONSTRUCTION

The present study attempted to examine the relationship between diseases and their associated genes, which were obtained from CTD database benchmark datasets. A DD similarity network based on shared genes was developed to identify every pair of diseases (Figure 3; Supplementary Note 4). The similarity value was computed in accordance with the relevance of the shared gene. Then, we implemented the RWR algorithm, calculated the disease seed node, and counted the other approaches mentioned in the Results section. We used a combination of qualitative and quantitative analysis tools for graph network. Additionally, we used a new strategy with a low calculation of association complexity. Hence, considerable time and vitality are required to manage the large-scale real-world application of the proposed method. We investigated the accessibility of this approach. We calculated the features of known disease nodes and measured the shortest path between two nodes.

We conducted an in-depth network graph-related literature review. Furthermore, we analyzed the disease network

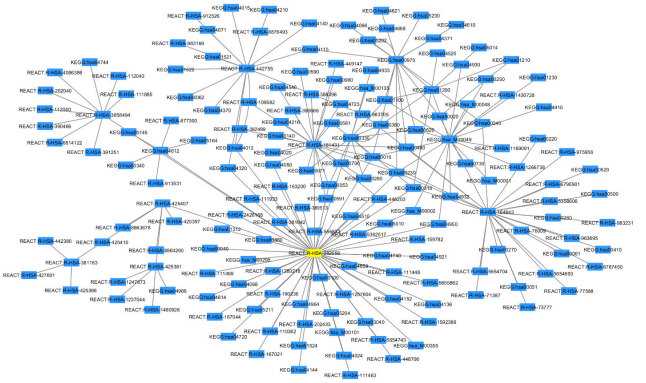

FIGURE 2. Disease-pathway association network model.

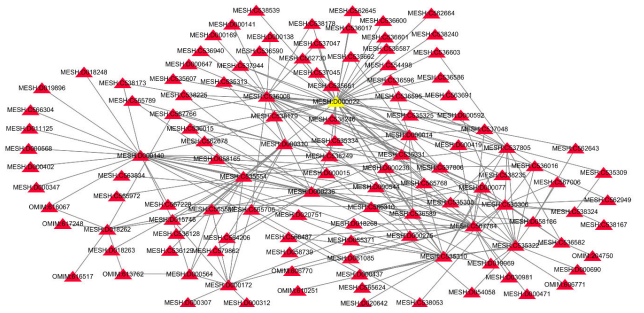

FIGURE 3. Disease-pathway heterogeneous network.

relationship and the connection between disease nodes and edges in a DD network [28]. The undirected graph, where **G = (d, d),** was considered. Here, d denotes the sets of disease (nodes), and their connections or edges are based on their similarity. The importance of closeness varies in accordance with the information used to build the network, which contains shared genes. The networks may be organic (gene and proteins), phenotypic, or have similar symptoms. Figure 3 shows a part of the known DD association network (Supplementary Note 4). The red triangular nodes represent diseases, whereas black lines represent the edges, which are connected to the disease-to-disease nodes.

The DD similarity network contained 210 nodes representing diseases. The value of the edge between two nodes indicates the similarity of two diseases. By performing the ''group projection'' of the DG association of the disease pathway adjacency matrix, we obtained the DD similarity network in the form of a DD graph. The equation behind the group projection is shown below. The DD adjacency matrix entries in the **di (i, k**) represent two kinds of disease and the number of total gene between dk. If two diseases di and dk share at least one gene, then we linked them to the DD graph. The edge of the weight is the total number of genes.

$$DD_{ik} = \sum_{j=1}^{|g|} DG_{ij}\ DG_{kj} => DD_{ik} \tag{9}$$

F. DISEASE-PATHWAY ASSOCIATION NETWORK MODEL

A general consensus in the literature [20] supports the true association scores for disease–pathways, as defined in their proposed methods. Thomas Gaudelet at el. (2019) revealed literature support for **7** out of the top **10** expected disease–pathway associations. This research-intensive disease–pathway association network, which is based on the RWRH network, has led to renewed interest in RWRHPDA-based heterogeneous networks. The RWRHPDA network is developed with the help of a pathway–pathway similarity network and DD similarity network by utilizing the disease–pathway relationship and constructing a heterogeneous network (Figure 4). The heterogeneous network is considered the depth of the disease–pathway association network. In addition, the RWRHPDA simultaneously prioritizes the pathway and disease candidates; this approach is highly encouraged by the high PageRank model. The seed needs (disease and pathway) are associated with the disease given a query disease. Moreover, the top-ranked disease is considered the most likely query disease. Figure 4 shows the disease–pathway association network.

Random Walk clarifies the role of the transition of an iterative walker from its present node to a randomly selected neighbor. This transition starts at a given source node in the network. The RWRH, on the other hand, allows the restart of the walk at each time step at node **v** with likelihood **r. P_0** is the likelihood vector at **step 0,** demonstrating that it is the initial likelihood vector with the sum of probabilities equivalent to **1. P** is the likelihood vector at step s, in which the ith component holds the likelihood of finding the random walker of **node i** at step **s.** The likelihood vector at step **s+1** is given as follows:

$$p_{s+1} = (1-\gamma)\, M^T p_s + \gamma p_0 \tag{10}$$

where M is the transformation matrix of the heterogeneous network, M_{ijis} the transformation likelihood from node **i** to node **j,** and γ **∈(0,1)** is the restart likelihood in each time step. After several iterations, $\mathbf{P}_\infty$ reaches a steady-state that is obtained by performing the emphasis until the change between P_s and P_{s+1} falls beneath $\mathbf{10^{-10}}$. Here, $\mathbf{P}_\infty$ is the measure of seed node closure. In vector, the node is more likely to be a seed node rather than the **j** node when (i)> $\mathbf{P}_\infty$ (j). Here, M is the transition matrix of the heterogeneous system, and it comprises four subnetworks, as shown below:

$$M = \begin{bmatrix} M_P & M_{Pd} \\ M_{Dp} & M_D \end{bmatrix} \tag{11}$$

where $\mathbf{M_p}$ is the transition matrix of the pathway–pathway interaction network. This matrix determines the effect of the subnetwork of the heterogeneous network. The transition matrix of the DD interaction network is denoted by $\mathbf{M_d}$, which determines the effect of the subnetworks of the heterogeneous network. $\mathbf{M_{Dp}}$ and $\mathbf{M_{Pd}}$, on the other hand, denote the transition matrixes' subnetworks. The inverse remains the same when γ is assumed as the likelihood of jumping from the pathway–pathway network to the DD network.

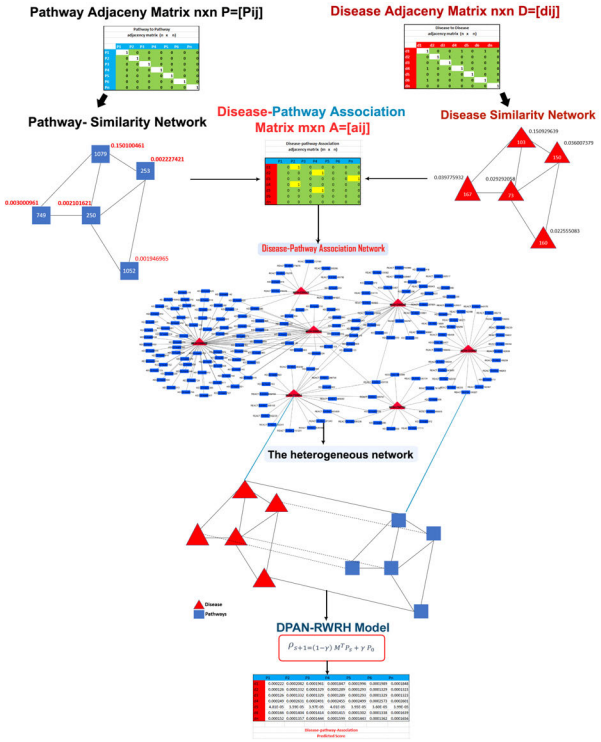

FIGURE 4. ROC–AUC score and PR curve. (a) AUC score of RWRHPDA and PR checking performance metric with computational methods for DPAN as shown in (b).

In the pathway–pathway network, $\gamma = \mathbf{0}$ if a node is not associated with the disease. Moreover, a node will jump to the disease network with the RWR vector score probability γ when a node is directly connected to the disease phenotype network. The node will jump to the other nodes present in the disease–pathway association network with $\mathbf{1}-\gamma$ as the vector score probability. Therefore, the likelihood of transition from d_i to p_j can be denoted as follows:

$$(M_{DP})\, ij = p\left(P_j|D_i\right) D = \begin{cases} \dfrac{\lambda\beta ij}{\sum_j Bji_0} & if \sum_j Bij \neq 0,\ otherwise \end{cases} \tag{12}$$

In the same manner, the probability of transition from $\mathbf{D_i}$ to $\mathbf{g_i}$ can be denoted as follows:

$$(M_{PD})\, i,j = p\left(D_j|P_i\right) D = \begin{cases} \dfrac{\lambda\beta ji}{\sum_j Bji_0} & if \sum_j Bji \neq 0,\ otherwise \end{cases} \tag{13}$$

where M is the transition matrix of the heterogeneous system. A part of the known disease–pathway association network is shown in Figure 5, where red and blue nodes represent diseases and pathways, respectively.

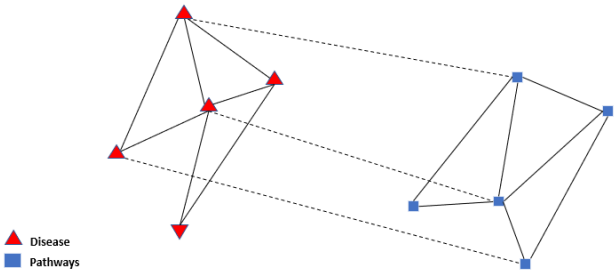

FIGURE 5. (a) Our proposed model RWRHPDA prediction score (ROC–AUC = 0.82) and checking of performance metric compared with other computational methods (b), including MERWPDA, GRMF, and RWRHPDA.

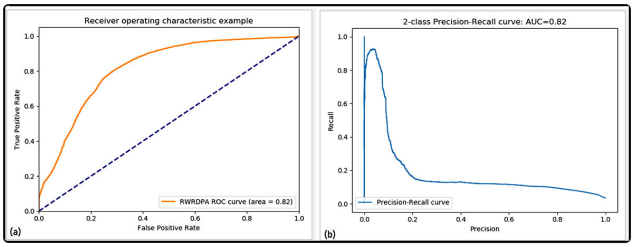

FIGURE 6. ROC–AUC score and PR curve. (a) AUC score of RWRHPDA and PR checking performance metric with computational methods for DPAN as shown in (b).

III. RESULTS

A. PERFORMANCE EVALUATION OF RWRHPDA

The predictions, execution, and evaluation of the algorithm are discussed in this section. As mentioned, similar to the estimation of the performance of other computational validation models, we estimated the performance of our strategy by using a receiver operating characteristic (ROC) and area under curve (AUC). ROC–AUC scores and precision–recall (PR) curves are shown in Figures 6(a) and 6(b), respectively.

Our results might suggest a connection between cancer pathway and disease cancer pathway. The article demonstrates potential disease-pathways association predictions. Our research uses quantitative techniques to analyze understanding of disease associations. We conducted all analyses of our proposed method on three major works, RWR and PageRank and RWRH to predict the disease pathways and identify potentially new disease-pathway associations. We have predicted 13,838 know disease–pathway associations.

The estimation algorithm is described as follows. First, we measured the association between disease and pathway (N) pairs from the bipartite network. Subsequently, these sets were predicted by the algorithm in the score matrix. We conducted all analyses using leave-one-out cross-validation (LOOCV) by selecting one sample as test data and other remaining data as training data. The number of data points included 1855 pathways associated with 211 diseases and 13,838 disease–pathway known associations. Then, we ran the LOOCV experiment 13,838 times. In every round, we defined disease/pathway in accordance with known experimentally verified disease–pathway associations. Unknown

association data were used as the test sample. However, the association of the remaining known diseases with a pathway should be used as a training sample. Predictions could be obtained when executing RWR with LOOCV. Content analysis was performed to determine when a LOOCV experiment was implemented. Each experimentally verified disease–pathway association was removed from our gold standard dataset. In the ROC curves, also known as sensitivity curves, false and actual positive rates are used as the horizontal and vertical axes, respectively. The predicted data, which were based on several variables, were used in the ROC metric to assess the AUC score. The ROC curve has the actual true positive rate *(TPR)* on the Y-axis and the false-positive rate *(FPR)* on the X-axis. This study used true data, which were predicted by using all of the data points of these curves. We implemented the sklearn package to measure the average precision score. Finally, the actual *TPR* and *FPR* of each threshold can be calculated as follows:

$$TPR = \frac{TP}{TP + FN} \tag{14}$$

$$FPR = \frac{FP}{FP + TN} \tag{15}$$

where *TP* and *TN* demonstrate the numbers of true positive case and negative samples that can be correctly recognized, respectively. *FP* and *FN* represent the numbers of positive and negative samples that cannot be correctly identified, respectively. Perfect performance occurs at $\mathbf{AUC = 1}$ and random performance at $\mathbf{AUC = 0.5}$. Additional variables were derived through another increasingly popular method, the PR curve PR. The data provided convincing evidence, which was also considered a broad conceptual basis for evaluating classification performance. The PR curve relates the positive predicted cases of the classifier to the true positive ratio and is particularly useful in applications with a small total number of positive cases. Different PR pairs can be found by setting different thresholds. The following pipelines can be used to calculate precision and recall rate:

$$Precision = \frac{TP}{TP + FP} \tag{16}$$

$$Recall = \frac{TP}{TP + FN} \tag{17}$$

where *TP* represents the number of true positive samples identified. *FP* and *FN* demonstrate the number of negative samples that were incorrectly labeled as negative samples. According to our prediction results, the ROC (AUC = 0.818) confirmed the superior performance of our method shown in Figure 7(a). Our findings suggest the need for a performance that is better than that of other methods, such as maximum entropy random walk on the heterogeneous network for the prediction of disease–pathway association (MERWPDA). The maximum entropy theory was applied to a random walk, and the potential disease–pathway association on the heterogeneous network was revealed. Graph-regularized matrix factorization (GRMF) for disease–pathway association prediction and MERWPDA were used as comparison methods

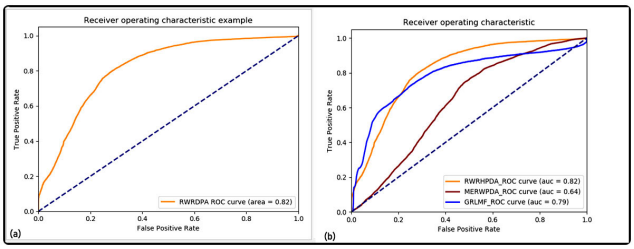

FIGURE 7. (a) Our proposed model RWRHPDA prediction score (ROC−AUC=0.82) and checking of performance metric compared with other computational methods (b), including MERWPDA, GRMF, and RWRHPDA.

TABLE 4. Calculation of metrics R^2.

Leave One Out Cross Validation	Leave One Out Cross Validation
R^2: 10.71767%, MSE: 0.04830	*R^2: 6.31235%, MSE: 0.03210*

to verify the effectiveness of our approach. Figure 7(b) shows the comparison results of RWRHPDA, MERWMDA [29], and GRMF [30] in terms of ROC (AUC). ROC and PR curves were drawn to evaluate the effectiveness of our approach.

B. CROSS-VALIDATION CALCULATION R2 FOR (LOOCV)

We conducted all analyses using true and predicted data for the calculation of metrics ($\boldsymbol{R^2}$ in particular). LOOCV was performed on the basis of the sklearn model selection cross-validation score (**model, *X, y,* scoring** = ***'r²'***). This approach aims to utilize cross-validation to obtain the generalized score of the model to improve its effective prediction from future data inputs. The percentages of cross-validation accessed new data, and a part of them can be used for testing on ''test'' data while utilizing the remaining data to assemble the model's ''training'' data. The $\boldsymbol{R}^2$, mean square error (**MSE**), accuracy score, and other metrics of the model can be calculated as shown in Table 4.

C. PERFORMANCE EVALUATION OF RWR, PPR, PGR AND PAGE RANK

Result analysis was based on three stages. The data were analyzed using RWR method on the pathway–pathway network. The RWR score vector w.r.t. pathway seed node, top 10 seed pathway nodes, and the list of near-neighbor pathway nodes were calculated. As shown in Figure 8 (Supplementary Note 5), the data were analyzed using different approaches, such as RWR, PPR, GPR, and PageRank. Figure 8(a) (Supplementary Note 5) shows a graphic summary of the pathway RWR vector score. The horizontal axis describes the pathway number nodes, whereas the vertical axis highlights the RWR vector score probability (0.150100461). Figure 8(b) (Supplementary Note S5) presents the description of pathway PPR vector score. The horizontal axis describes the pathway number nodes, whereas the vertical axis highlights the pathway, with the PPR vector score probability of 0.150100461. Figure 8(c)

(Supplementary Note 5) shows the description of pathway GPR vector score. The horizontal axis describes the pathway number nodes, whereas the vertical axis highlights the pathway with the GPR vector score probability of 0.002994948. Figure 8(d) (Supplementary Note 5) presents the graphic summary of pathway change in residuals from PageRank. The horizontal axis describes the iterations, whereas the vertical axis highlights the residual ratio. We stored iterate_PageRank, which indicates the difference between the current PageRank score vector, in residuals. We plotted the residuals and checked the tendency of the residuals over the iterations.

D. RESULT ANALYSIS BASED ON RWR, PPR, PGR AND PAGERANK

The analysis was conducted based on RWR, PPR, and PageRank. The RWR method outperformed the score vector on the basis of the power iteration of the walker transition from a given source node input graph. The results were used to calculate the query (seed) pathway nodes provided by the single-source RWR score vector. We first calculated the RWR seed pathway nodes, which had additional connections. The input similarity network was constructed using the RWR score vector, similar to the variance used in the pathway–pathway network. Table S1 shows the results of the seed pathway node, that is, the pathway seed node. Then, we set up the neighborhood node representing the pathway closest to the seed pathway. Table S2 lists the nearest neighbors of the seed pathway nodes. Table S3 provides the list of top 10 seed pathway nodes. Additional results on the pathway of the top 10 nodes showed high-probability pathway node RWR score vector toward the scale (Table S4; Supplementary Note 1). Next, we calculated the PPR seed vector fraction w.r.t the seed pathway. The results obtained by PPR were consistent with our findings in Table S5 (Supplementary Note 1). The results showed that personalized pathway node ranking vector score was associated with all seeds in the pathways (Table S5) and the top 10 seed pathway nodes (Supplementary Note 1). Then, we calculated the rank node pathway, which is a well-known variation of RWR. We ranked candidate pathways based on the near neighbor nodes. PageRank is an essential mathematical formulation algorithm for biological networks and applied to random walk formula. We used PageRank in the graph of the pathway–pathway network. We analyzed the relationship between the pathway nodes and edges of the graph. As shown in Table S6, a considerable diversity was present in the pathway–pathway network nodes with significant variation. The result of the global rank score vector presented the top 10 pathway nodes (Supplementary Note 1).

The data were analyzed using different methods of the link analysis (or ranking) model, such as PageRank, subject-specific PageRank, and HITS. Given that PageRank is a global ranking, seed specification is unnecessary. PageRank automatically sets the desired seed pathway (that is, all nodes will be used as seeds). GPR vector r was used to obtain personalized pathway node ranking. As shown in Table S6, PageRank was used to obtain the results of the top 10 nodes and labels from the graph, whereas other variables were obtained from the numbers of nodes (n: 1 855) and edges (m: 20 527). Then, the pathway–pathway network of the row-normalized pathway PageRank of the given adjacency matrix was obtained (Table S7). PageRank score vector P was obtained by calculating the PageRank algorithm. Table S8 shows the top 10 PageRank scores from each node of the pathway–pathway network. The graphs suggest that the PageRank score was calculated directly without iteration. We inversed the matrix to obtain the exact solution. Table S9 shows the PageRank scores of the precise and iterative operations of the pathway–pathway network. In contrast to our expectations, PageRank scored the ranking results. We implemented the function rank node to rank the nodes in the order of a given rank score. Table S10 shows the top 10 highest PageRank pathway node scores (Supplementary Note 1).

Considerable research attention has been directed toward the prediction of novel disease-associated genes by the following main directions: (1) according to biological network, (2) functional annotation, and (3) machine learning. A computational method has been proposed to predict such associations by ranking candidate diseases based on their association with diseases. Network-based approaches are dominating because they use ''disease module'' principle in DD similarity networks. Results were obtained using RWR, PPR, and PageRank implemented in the DD network. The application of the PageRank model to DD networks was consistent with our findings. Table S11 shows the results of the DD networks used to compute the variance of a single-source RWR score vector w.r.t. with a given query (seed) disease node (Supplementary Note 2). Second, we set up near neighbor nodes to indicate the disease closest to seed disease (Table S12). Then, we calculated the top 10 seed disease nodes (Table S13). Table S14 presents the RWR vector scores for the DD network for the top 10 disease nodes (Supplementary Note 2).

Meanwhile, our PPR score was reasonable in terms of statistical and probabilistic measures on the basis of technology that prioritized related diseases in the network (Table S15). Additional findings support these conclusions (personalized ranking). Multiple seed disease was allowed. Therefore, we used other network-based ranking technologies, such as global ranking, to calculate the scores rather than GPR. PageRank is a global ranking. Thus, this algorithm automatically sets the desired seed (that is, all nodes were used as seeds), as shown in Table S16 (Supplementary Note 2). The global ranking methods that model information flow to assess the proximity and connectivity between disease nodes, e.g., with a priori PageRank, were used in our calculations. Our method with these ranking approaches performed better than the approach based on localized methods. We implemented the dense matrix version of PageRank, with the input graph network number (n) of nodes set to 210 and the number (m) of edges equal to 4398. Table S17 shows the number of the listed disease nodes and edges of the DD network. Then,

we normalized the given adjacency matrix from DD networks (Table S18). A further complication for the presented result was that it implemented the iterative algorithm for PageRank (Table S19). The analysis of results consisted of the exact PageRank score vector and its stages. We implemented a function for ranking nodes in the order of PageRank scores and results of exact and iterative disease PageRank score vector (Table S20; Supplementary Note 2). Table S21 shows the presented results of disease rank node which were used to rank nodes in the order of given ranking scores and top 10 disease ranking score in the order of PageRank scores (Supplementary Note 2).

We mainly focused on the algorithms of the models and how they are used to rank nodes in real-world graphs. We implemented PageRank in Python. First, we calculated PageRank through a dense matrix. In this study, input graph adjacency matrix was adopted, and data rows were normalized by row normalization adjacency matrix A. Table S22 lists the normalized data used in the formula for the row-normalized matrix A (Supplementary Note 3). We obtained the out-degree vector d by row-wise summation and obtained the inverse of the out-degree matrix score (Table S23). The results were considered iterative algorithm on the basis of PageRank. We used the function iterative PageRank and obtained the PageRank score vectors of the top 20 disease pathways (Table S24; Supplementary Note 3). The results of computed PageRank on the basis of the iterative solution were equal to the exact solution of PageRank (Table S25). Table S25 shows the top 20 disease pathway exact scores and iterative score vector (Supplementary Note 3). Finally, we achieved the PageRank scores from the function. Table S26 presents the top 10 disease pathways and top 10 highest page rank scores in the graph (Supplementary Note 3).

TABLE 5. 25 Potential disease-pathway association predicted result and their evidence.

Disease name	Pathway ID's	Evidence	Predicted score
22q11 Deletion Syndrome	KEGG:hsa05200	KeggDB	0.00019308
Abortion, Spontaneous	KEGG:hsa05202	KeggDB	0.000186
Achondroplasia	KEGG:hsa05206	KeggDB	0.00016535
Acrocapitofemoral Dysplasia	KEGG:hsa05205	KeggDB	0.00018171
Acute Coronary Syndrome	KEGG:hsa05204	KeggDB	0.0001958
Acute Coronary Syndrome	KEGG:hsa05203	KeggDB	0.00019083
Acute Kidney Injury	KEGG:hsa05230	KeggDB	0.00019537
Adenocarcinoma, Clear Cell	KEGG:hsa05214	KeggDB	0.00012831
Adenocarcinoma of Lung	KEGG:hsa05216	KeggDB	0.00018908
Adenoma	KEGG:hsa05221	KeggDB	0.00019009
Adenoma, Liver Cell	KEGG:hsa05220	KeggDB	0.00017635
Adenomatous Polyposis Coli	KEGG:hsa05217	KeggDB	0.00018508
Alzheimer Disease	KEGG:hsa05219	KeggDB	0.00019772
Alzheimer Disease	KEGG:hsa05215	KeggDB	0.00018913
Ameloblastoma	KEGG:hsa05213	KeggDB	0.00019872
Acute Kidney Injury	KEGG:hsa05223	KeggDB	0.00019886
22q11 Deletion Syndrome	KEGG:hsa04110	KeggDB	0.221243109
Adjustment Disorders	KEGG:hsa04150	KeggDB	0.381654024
Adjustment Disorders	KEGG:hsa04330	KeggDB	0.217863366
Adrenal Gland Neoplasms	KEGG:hsa04510	KeggDB	0.217863366
Adult-onset citrullinemia type 2	KEGG:hsa03320	KeggDB	0.015214291
Adult-onset citrullinemia type 2	KEGG:hsa03320	KeggDB	0.571620621
Adjustment Disorders	KEGG:hsa04330	KeggDB	0.217863366
Adrenal Gland Neoplasms	KEGG:hsa04510	KeggDB	0.015214291
Adult-onset citrullinemia type 2	KEGG:hsa03320	KeggDB	0.571620621

IV. CASE STUDIES

The KEGG [31] and Reactome [32], [33] pathway data were used to explain the known molecules associated with the network. These data, combined with the diseases, pathways, chemicals, and genes in the CTD database [34], provide insights into the molecular systems that may be affected by the underlying mechanisms of chemical and environmental diseases. CTD database and associated gene can be used to cure the disease, in accordance with the KEGG and Reactome pathways, on the basis of the total shared genes. The reports described the diseases that are associated with the human pathway theme and genes that are related to these pathways.

Human pathways are located in cancer-related regions of the genome, many of which are involved in the development of various human malignancies [35]. For example, MESH:C535334 is associated with ABCD syndrome gene and causes diseases, including albinism, black locks, cell migration disorders of the gut nerve cells, and deafness. KEGG:hsa05219 pathway is a genetic signature with disease prediction in bladder cancer and other cancer pathways observed in KEGG:hsa05200 [36]. Signaling pathways associated with hepatocellular carcinoma cancer disease pathway KEGG:hsa05225 are a central carbon metabolism in hepatocellular carcinoma and human hepatoma HepG2 cells, which are depleted in macroH2A1. KEGG pathway:hsa05210 (i.e., colorectal cancer (CRC)). In the present study, the pathway involved in bladder cancer hsa05219 [37] from the KEGG pathway database was selected. To evaluate the efficacy of RWRMDA on independent datasets, we conducted a case study of three cancers, that is, bladder cancer, CRC, and lung cancer. Various databases and literature verified the predicted results.

Cancer pathway is one of the most common cancers in humans, accounting for 22% of all cancers in women. In this study, which used benchmark datasets, 98 pathways were associated with human disease. The priority of candidate pathways on the basis of RWRHPDA was determined. Data points were used in 386 KEGG pathways, 1469 Reactome pathways, and 210 diseases. A total of 389550 confirmed predictions were obtained using MeSH or OMIM, whereas CTD has been updated several times since the link between

disease and pathway was downloaded in December 2019. Table 5 lists the evidence for 25 diseases and their associations with human pathways in cancer. In total, 98% of those predictions were correct. As shown in Table 5, we implemented RWRPDA for disease–pathway cancers containing a specific human pathway. RWRHPDA method can be utilized to predict dysregulated interactions and disease-associated pathways. Our method was used to inspect the variables associated with various diseases and predict the potential related pathways. We also conducted a case study of the first pattern, with cancer and reaction pathways related to a specific type of pathways used as the training sample. Then, unknown associations between cancer were observed as a test sample. For RWRHPDA prediction, records in the KEGG and CTD databases confirmed and identified potentially specific types of pathways in 24 cancer-related human pathways. Table 5 shows the pathway–pathway network analysis results of 24 human disease pathways, which are primary cancers pathways. Considering space limitations, we focused on the results of specific cancer types. The cancer enrichment pathway aims to investigate the function of particular types of known cancer-related pathways in cancer. The prediction of potential pathways confirmed the rationality of RWRHPDA because the functions of the possible cancer-related pathways are related to cancer development. We used quantitative techniques to analyze the disease associations. We also analyzed the relationship between diseases and pathways, and the results can be used to distinguish the dysregulated connections and disease-related pathways. We prioritized candidate disease genes and categorized pathways.

A new bidirectional network of diseases, pathways, and genetic associations was revealed. Our results provide strong evidence for the association of these variables in a number of ways using RWR and RWRH calculations (see Methods for additional information).

Computational methods have been used to examine the mechanisms of dysregulation in complex pathways, identify disease associations, and improve treatment. However, given the heterogeneity of the samples and patients, obtaining biological insights from conventional, single-gene-based analyses of the ''omics'' data from high-throughput trials is challenging. Overall, these studies support the effectiveness of the challenges and have developed new ways and network-based approaches to analyze the comprehensive data ''based on the biological pathways,'' such as KEGG human pathways, Reactome human pathways, WikiPathways, classification of diseases, gene expression, pathway enrichment analysis of GWAS, and biological networks. Our findings were based on the development of pathway-based methodologies for the prediction of novel interactions and heterogeneous networks of disease-related pathways. We encourage further research to examine whether these models can provide reasonable predictions for the majority of patients. We predicted novel associations, verified the comparison of the pathway–pathway, DD, disease–pathway, and known interactions and proposed explanations for the novel predictions from our study. We have described the disease–pathway association score in accordance with our method and tested the validity of our prediction by comparing the result with the CTD database. The first association and pathway in cancer, that is, in *Homo sapiens*, may also be correct. Lactose synthesis pathway (hsa05200) contains the following genes: BAX, BDKRB2, EGFR, GSK3B, GSTP1, IGF1, IL6, NFKB1, NOS2, PPARG, TGFB1, TP53, and VEGFA. All these genes may be involved in cancer disease-related pathway.

Cancer pathways include genes, proteins, and their complex interactions. The findings suggest that cancer drugs work on only a part of a particular pathway, and the direction where the entire pathway will evolve, or whether the treatment will be beneficial is unknown. The possible reason for this discrepancy may explain why cancer pathways have not met our expectations. The possible explanation for this finding was the effectivity of anticancer therapy. Thus, the effect of drugs on the entire cancer pathway should be considered. Over the past decade, imperative genes responsible for the development of various cancers have been revealed, their mutations have been accurately identified, and their pathways have been described. To test our hypothesis, we built a computational model of the cancer pathway and ran it on a supercomputer. RWRHPDA transformed the predictive association scores of the mammalian target of rapamycin signaling pathway, Notch signaling pathway, peroxisome proliferator-activated receptor signaling pathway, focal adhesion, and cAMP signaling pathway, which have been confirmed by KEGG. Finally, we compared the calculated results with the experimental findings of the existing and proposed methods. The calculated efficacy of the cancer disease-related pathway was 82.18%.

The results analysis was based on three stages. The data were analyzed using the RWR method used on the pathway–pathway network, and the RWR score vector w.r.t. pathway seed node, top 10 seed pathway nodes, and the list of near neighbors (181 mm/43 picas). The maximum depth can reach 8.5 in (216 mm/54 picas). In the depth selection for a graphic, a space for a caption should be allowed. Figure size can be between column and page widths if the author intends to. However, as recommended, the figures should not be less than column width unless necessary. Cancer pathway was used in the first case study with a similar pipeline to CRC. For the first 20 different types of cancer, that is, cancer-related disease pathways for specific pathways predicted by RWRHPDA, the evidence recorded in the KEGG pathway database also confirmed the 20 different types of pathways (Table 6). Research on the pathway (KEGG:hsa05210) is limited. The majority of work in this field focused on the mechanism of genomic instability, which has been identified in sporadic CRC advances. CRC is the third most common cancer worldwide, and more than half of CRC deaths occur in developed countries. Studies showing that epithelial cells in the large intestine result from the accumulation of genetic changes in specific oncogenes and the information on tumor suppressor genes are limited. Limited studies have

TABLE 6. 20-Pathway and their evidence from KEGG database.

Pathway Names	Evidence
Glycolysis (Embden-Meyerhof pathway), glucose => pyruvate	KEGG:hsa_M00001
Glycolysis, core module involving three-carbon compounds	KEGG:hsa_M00002
Gluconeogenesis, oxaloacetate => fructose-6P	KEGG:hsa_M00003
Pentose phosphate pathway (Pentose phosphate cycle)	KEGG:hsa_M00004
Pentose phosphate pathway, oxidative phase, glucose 6P	KEGG:hsa_M00006
Citrate cycle (TCA cycle, Krebs cycle)	KEGG:hsa_M00009
Citrate cycle, first carbon oxidation, oxaloacetate => 2-oxoglutarate	KEGG:hsa_M00010
Citrate cycle, second carbon oxidation, 2-oxoglutarate	KEGG:hsa_M00011
Malonate semialdehyde pathway, propanoyl-CoA => acetyl-CoA	KEGG:hsa_M00013
Glucuronate pathway (uronate pathway)	KEGG:hsa_M00014
Proline biosynthesis, glutamate => proline	KEGG:hsa_M00015
Serine biosynthesis, glycerate-3P => serine	KEGG:hsa_M00020
Lysine degradation, lysine => saccharopine => acetoacetyl-CoA	KEGG:hsa_M00032
Methionine salvage pathway	KEGG:hsa_M00034
Methionine degradation	KEGG:hsa_M00035
Leucine degradation, leucine => acetoacetate + acetyl-CoA	KEGG:hsa_M00036
Catecholamine biosynthesis, tyrosine => dopamine =>	KEGG:hsa_M00042
Tyrosine degradation, tyrosine => homogentisate	KEGG:hsa_M00044
Creatine pathway	KEGG:hsa_M00047
Inosine monophosphate biosynthesis, PRPP + glutamine => IMP	KEGG:hsa_M00048
Adenine ribonucleotide biosynthesis, IMP => ADP,ATP	KEGG:hsa_M00049
Guanine ribonucleotide biosynthesis IMP => GDP,GTP	KEGG:hsa_M00050

TABLE 7. 12- Pathways and their evidence from the reactome database.

Pathway Names	Reactome ID	Evidence
Interleukin-6 signaling	REACT:R-HSA-1059683	ctd/ Reactome
Apoptosis	REACT:R-HSA-109581	ctd/ Reactome
Hemostasis	REACT:R-HSA-109582	ctd/ Reactome
Intrinsic Pathway for Apoptosis	REACT:R-HSA-109606	ctd/ Reactome
PKB-mediated events	REACT:R-HSA-109703	ctd/ Reactome
PI3K Cascade	REACT:R-HSA-109704	ctd/ Reactome
MAPK3 (ERK1) activation	REACT:R-HSA-110056	ctd/ Reactome
Translesion synthesis by REV1	REACT:R-HSA-110312	ctd/ Reactome
Translesion Synthesis by POLH	REACT:R-HSA-110320	ctd/ Reactome
Cleavage of the damaged	REACT:R-HSA-110329	ctd/ Reactome
Cleavage of the damaged purine	REACT:R-HSA-110331	ctd/ Reactome
Displacement of DNA	REACT:R-HSA-110357	ctdb/ Reactome

also been conducted on (KEGG: hsa_M00001) glycolysis–Embden/Meyerhof/Parnas pathways. Previous studies primarily overlooked glycolysis as a metabolic pathway that is commonly found in biological systems. Most studies focused on glucose metabolic energy from the entry of pyruvate into the citric acid cycle and oxidative phosphorylation. The research on glycolysis is limited. Core module is involved in three-carbon compounds. In the present study, we attempted to establish a link between Embden–Meyerhof pathway and KEGG module:hsa_M00002.

This paper provides the key factors that influence glycolysis and a metabolic pathway that is generally present in biological networks. Many researchers converted glucose into two pyruvate molecules in a series of reactions. These pathways occur under aerobic conditions. Under anaerobic conditions, pyruvate can be converted to lactic acid in muscles or ethanol in yeast. Empirical evidence appeared to support the following view: (REACT: R-HSA-1059683) IL-6 signaling is a pleiotropic cytokine that plays a role in immune regulation, hematopoiesis, inflammation, tumorigenesis, metabolic control, and sleep. Trans-signaling pathway is responsible for the proinflammatory activity of Il-6, whereas membrane-bound receptors control the regeneration and anti-inflammatory activity of IL-6, and IL-6R signal transduction is mediated by the Janus family tyrosine kinase signaling and transcription-activated pathway of transcription and Ras-mitogen-activated protein kinase (MAPK) pathway. In the present study, we checked whether Reactome has a human pathway and conducted two experiments, namely, RWR and RWRHPDA.

These pathways have been established through decades of molecular biology research and have been validated in the following URL in various regular pathway repositories (KEGG and REACTIOME pathway repositories): https://reactome.org/download/current/ReactomePathways.txt. These pathways have been also been established from ctdbase.org (Table 7).

Many scholars provided empirical evidence to support (R-HSA-109581) the claim of apoptotic *Homo sapiens*. Apoptosis is a unique form of cell death that differs from necrosis in terms of function and morphology. Apoptosis is commonly characterized by nuclear chromatin concentration, cytoplasmic contraction, endoplasmic reticulum, and membrane blebbing. In various nonimmune cells, death signals initiated by extrinsic pathways are amplified by connections to internal pathways. The widely accepted hypothesis (R-HSA-109582) is that hemostasis is a physiological response that eventually stops the bleeding of injured vessels. Under normal circumstances, the vascular endothelium supports vasodilation, inhibits platelet adhesion and activation, inhibits coagulation, enhances fibrin lysis, and performs anti-inflammatory action. In the present study, we used qualitative/quantitative techniques to analyze the disease dataset, which was validated through the Mesh and OMIM websites (Table 8).

The similarity between two disease-paired entities was calculated based on the DD similarity network. Thus, disease association analysis is essential for our understanding of the human disease pathways. We constructed a cellular

TABLE 8. 20- Disease datasets confirm verified from MESH & OMIM.

Disease Name	Disease ID	Evidence
17-Hydroxysteroid Dehydrogenase Deficiency	MESH:C537805	MeSH/ Ctd
18-Hydroxylase deficiency	MESH:C537806	MeSH/ Ctd
22q11 Deletion Syndrome	MESH: D058165	MeSH/ Ctd
2,4-Dienoyl-CoA Reductase Deficiency	MESH:C565624	MeSH/ Ctd
2-AMINOADIPIC 2-OXOADIPIC ACIDURIA	OMIM:204750	OMIM/ Ctd
2-Hydroxyglutaricaciduria	MESH:C535306	MeSH/ Ctd
2-Methylbutyryl-CoA Dehydrogenase Deficiency	MESH:C566487	MeSH/ Ctd
3b-Hydroxysteroid Dehydrogenase Deficiency	MESH:C579862	MeSH/ Ctd
3C syndrome	MESH:C535313	MeSH/ Ctd
3-Hydroxyacyl-CoA Dehydrogenase Deficiency	MESH:C535310	MeSH/ Ctd
3-methylcrotonyl CoA carboxylase 1 deficiency	MESH:C535308	MeSH/ Ctd
3-methylcrotonyl CoA carboxylase 2 deficiency	MESH:C535309	MeSH/ Ctd
3-Methylglutaconic Aciduria, Type I	MESH:C562801	MeSH/ Ctd
3-Methylglutaconic Aciduria, Type V	MESH:C565706	MeSH/ Ctd
5-oxoprolinase deficiency	MESH:C535322	MeSH/ Ctd
6-pyruvoyl-tetrahydropterin synthase deficiency	MESH:C535325	MeSH/ Ctd
ABCD syndrome	MESH:C535334	MeSH/ Ctd
Abdominal Pain	MESH: D015746	MeSH/ Ctd
Aberrant Crypt Foci	MESH: D058739	MeSH/ Ctd
Abetalipoproteinemia	MESH: D000012	MeSH/ Ctd

TABLE 9. Disease associated with a large number of pathways.

No.	Disease Name	Disease ID	Number of associated pathways
1	Adenocarcinoma	MESH:D000230	951
2	Adenocarcinoma of Lung	MESH:D000077	941
3	Alzheimer Disease	MESH:D000544	656
4	Abortion, Spontaneous	MESH:D000022	458
5	Abnormalities, Multiple	MESH:D000015	380
6	Adenocarcinoma Of Esophagus	MESH:C562730	371
7	Adrenocortical Carcinoma, Hereditary	MESH:C565972	134
8	Amyloidosis, familial visceral	MESH:C538249	117
9	Abnormalities, Drug-Induced	MESH:D000014	116
10	Acromicric dysplasia	MESH:C535662	94

network on the basis of genetics, using common physiology and pathophysiology to analyze the relationship between the disease and pathway. Infections are often linked to the web search for common causes common to similar diseases. Several disease nodes join diseases together on the basis of their genetic overlap. Meanwhile, DD similarities that are easily detected at the molecular level rather than at the phenotypic level will be missed. Other researchers have attempted to find genetic overlap between diseases. If the disease overlaps with disease genes, then they will link together [38] or link with metabolites or biological pathways [39].

After verifying the accuracy of RWRHPDA with ROC-AUC and case studies on specific types of human cancer pathways, we further predicted novel diseases associated with multiple pathways.

Here, all known disease-related pathways in the baseline data were used as seed pathways. For all 210 diseases, the top 20 potential pathways were published to facilitate the discovery of human disease and pathway associations. This article outlined the first 24 major pathways that are associated with cancer, as described above. The potential diseases predicted by RWRHPDA concerning other pathways will also be confirmed through further experiments and attention to the disease.

This study combined computational and quantitative tools. We checked the limitations to RWRHPDA. First, although RWRHPDA can predict the data at hand in seconds, the size of the heterogeneous network can affect the model speed. Therefore, if the number of diseases and pathways investigated is large, then the efficiency of the model will be affected to an extent. Second, RWRHPDA cannot be conducted on a weighted biological network. This shortage is due to the original construction of RWRH. Third, although nodes are assigned to different markers as diseases or pathways in the constructed heterogeneous network, edges corresponding to varying relationships between nodes are not distinguished. Therefore, further research should focus on how to apply RWRH to multisource biological datasets in a feasible pipeline, regardless of the use of edge-coloring methods.

After confirming the efficacy of RWRHPDA against LOOCV validation, to further validate our method's ability to predict new disease-pathway indications, all known disease-pathway pairs in the standard gold data set were used as training sets, and the remaining unknown disease-pathway pairs were considered candidate associations. By applying RWRHPDA, we were able to obtain predictive scores for all pathway candidate - disease associations. For specific pathways, all candidate diseases were ranked according to their predicted scores, and we collected the association between the predicted disease and pathway in the top 24 pathways as the predicted results. For all approaches, the predicted results are listed in (see Supplementary Note 1).

We conducted case studies on the highest-ranking predicted diseases on the basis of the KEGG pathway database of public biological [40] and current methods to verify the correctness of predicted results. In the KEGG pathway database, several newly validated diseases–pathway pairs provided basis for our validation. For example, we selected several pathways and the corresponding first five candidate diseases (Supplementary Note 1). Several novel disease–pathway pairs were identified in the KEGG pathway database. RWR method has been predicted as potential disease path-associated cancer. This method has been verified in RWR or the KEGG pathway database. These successful case studies showed that our proposed approach can predict new disease–pathway associations. A possible interpretation of this finding was that the types of diseases are associated with a large number of pathways (Table 9).

V. FUTURE DIRECTION

We also explored the limitations of these models and discussed how computer models for human pathway research may be established in the future. However, we cannot perform wet experiments to validate the forecasts due to the limitations of laboratory conditions. Hence, the tests must be followed up in the future under laboratory conditions. In additional notes, we present the prediction results. The mechanisms of cancers are one of the most widely recognized organizational concepts for cancer research. miRNAs and illnesses have certain associations [41]. Several researchers have studied the nature of biological experiments, which are effectively complemented by computational methods in predicting the possible relationship between miRNA and diseases. Xing Chen Chen *et al.* established a novel inductive matrix completion model (miRNA-disease association prediction [IMCMDA]) [41]. Two databases have been reviewed for the miRNA forecast for certain diseases. The study of possible miRNA disease prediction will help in identifying disease pathogenesis and facilitate clinical treatment. Researchers also developed an inductive matrix completion model for IMCMDA [41]. Drug-target association recognition is an important research process. Although high-performance testing and other biomedical tests are possible, experimental methods for the detection of drug/target interactions are still extremely expensive and challenging at present. Therefore, various models for the calculation of potential drug–target associations have been developed on a large scale. Researchers implemented several modern computational models, including a network-based approach and machine-based learning system, to predict drug–target interactions. In the machine-based learning process, special attention was paid to supervised and semi-supervised models with significant differences in the acceptance of negative samples. Although significant improvements have been reached in the study of drug–target interactions with effective computational models, network- and machine-learning methods have their own limitations. Network is an important way to predict underlying drug interactions [42]. We are also investigating the future directions of a network-based pathway and network strategy for custom pathways on the basis of personalized medicine, genome sequence, clone tumor, and cancer markings. Drug combinations are effective approaches to resolve the resistance to fungal drugs and fight against complex diseases. The NLLSS project was developed to predict the possible combinations of synergistic drugs by incorporating well-known combinations of synergistic drugs, unknown pharmaceutical combinations, pharmaceutical target interactions, and large-scale chemical structures. This project promoted NLLSS, which often appears identical to adjoining medicines and vice versa, as a key medication with a synergistic effect. In the identification of potential synergistic combinations of drugs, NLLSS proved to be outstanding through cross-validations and experimental validations. Out of the 13 predicted antipilling synergistic drug combinations, 7 candidates were experimentally tested. NLLSS can develop a new method to identify potential synergistic drug combinations, discover new indices for existing medicines, and provide insights into the synergistic mechanisms of drug synergies underlying molecular mechanisms. Previous studies on synergistic drug combinations can be categorized into three categories. When describing synergy to decide whether the combined drug is synergistic, only synergistic combination experiments are performed, whereas only statistical estimates are conducted to provide synergistic combinations. For example, methods, such as combined index equations, Loewe additive model, HAS models, and a general approach to the universal reaction surface [43], define only the concept of synergy. In recent decades, lncRNAs have attracted the attention of researchers worldwide. In the last couple of years, thousands of lncRNAs over eukaryotic organisms lain out from humans were identified with the speedy development in experimental technology and computational algorithms. Research indicated that lncRNAs also play a significant role in several essential biological processes in nearly the entire cell cycle through various mechanisms. Therefore, lncRNAs are mutated and dysregulated to develop different complex human diseases.

In the present study, several lncRNAs that are related to human diseases have been experimentally identified. Thus, the study and prediction of the potential human interactions of lncRNA–disease and prediction of human lncRNA–diseases have become significant bioinformatic tasks that would support the mechanism for understanding complex human disease systems at the lncRNA level, biomarker–disease identification, and disease detection, treatment, prognosis, and prevention [44]. New methods and strategies have been developed to combine different data from ''omics,'' such as gene expression, alteration of number replication, GWAS, and interaction data, to address these challenges. Recent methodological advances will be discussed in this analysis for pathways to identify dysregulated interactions, correlate subnetworks with diseases, prioritize candidate genes, and classify illnesses. We will also address the related problems and possible future directions for each program. Challenges in determining disease-related pathways are the absence of complete and precise human interactions, inadequate understanding of biological processes and the role of human genome intergenic regions, and lack of complete set of epigenetic data.

VI. DISCUSSION

RWR: To achieve the transition matrix, we used two techniques, namely, the traditional method and transition matrix estimation via the Laplacian concept of normalization. RWR is an algorithm of ranking [15] used for the prioritization of candidate genes in a previous work [45]. RWR assumes a random walker that starts at a seed node or at a setting of seed nodes and jumps to their near neighbors or reverts to the seed nodes at every stage randomly. The probability of

the random walker that reaches the corresponding node can be calculated for all nodes in the graph. RWR gives a good significance score in weighted graphs between two nodes and has been used successfully in configurations, including automated image subtitling, ''connection subgraph'' generalizations, and PPR. At present, several quantitative biological methods provide new and efficient resources to recognize the associations between miRNA and disease, lncRNA and disease, disease and pathways, and microbes and diseases. Extensive RWR approach was used to obtain the potential relationship between microbes and disease. In this research, we proposed the heterogeneous linked network of the human microbe–disease associations as the latest calculation model of expanded random walking with restarting optimized by particle swarm optimization [38]. Different random walking approaches have been used to determine the likelihood of the interaction between the predicted microbes and diseases. On the basis of a classic random walk with heterogeneous Spearman correlation parameters, Shen *et al.* [46] derived a priority method for the prediction of disease–microbe associations for candidate microbes. Zou *et al.* [47] developed a calculation model to predict a birandom walk on a heterogeneous network. Different predictive models are expected to enhance the identification of novel associations between microbes and diseases.

VII. CONCLUSION

We proposed a novel method that can be used to determine whether targeted disease–pathway associations can be determined through gene-related pathways, integrating genetic and natural relationships to characterize disease-related goals. The effective screening of environmental factors was also considered because screening can help in setting important pathways, which will further provide research guidance on human diseases and health issues. Several findings of this study warrant further discussion on several areas, such as diseases, pathways, genes, and chemical associations within graph networks. Overall, the findings of this study will support the validity of disease–pathway associations. Disease–pathway association network is an important area of biomedical research. However, the information about pathways is relatively limited. In bioinformatics, crucial questions remain unanswered for human disease pathway associations. In this study, a positive association was obtained between human disease pathways by the algorithm of our proposed method (RWRHPDA) to determine the relationship between cancers and pathways. We connected the pathway–pathway and DD networks by disease–pathway association and constructed a heterogeneous network. This method has limitations that should be improved in future research. We conducted all analyses of our proposed method using three major networks, namely, RWR, PageRank, and RWRH, to predict disease pathways and identify potentially new disease–pathway associations. The algorithms included RWR algorithm which works on pathway–pathway network, RWR which was applied on DD network, and heterogeneous network for disease–pathway association network. A number of studies have focused on the interactions between disease–pathway and DG associations to modify diagnosis. Several findings of the present study, such as the data of KEGG and Reactome pathways describing known molecular interactions and reactions, warrant further discussion. These data were integrated into CTD database of chemicals, genes, and diseases to provide visibility into biological networks that may be influenced by chemicals and potential processes influencing environmental diseases. These links were determined with the notion that despite the independent relationships of pathways and diseases with the same gene or genetic groups, they were concluded by the study of research publications, the development of networks, and analysis of statistics. However, out of all diseases with at least one pathway, 82% of the total set of the possible pairs are extremely high in pair sharing with at least one pathway. All these diseases were unlikely to have a connection. Hence, the more pathways the diseases share, the more likely for them to interact with each other. Although the current knowledge of these diseases is strongly supportive of our molecular mechanisms, a further move is to identify appropriate biomarkers and drug targets among the expected genes and pathways that can be used to enhance diagnosis, prognosis, and care. In our future works, we will focus on integrating additional human disease–pathway data in various dimensions to expand drug chemical pathway associations and achieve significant results.

SUPPORTING INFORMATION

S1. Supplementary Information

S2 File. The data file of disease-pathway associations.

S3. File. Predicted results of potential pathway-disease associations in descending order

LIST OF ABBERIVATIONS

CTD:	Comparative toxicogenomics database
NCI:	National cancer institute
GWAS:	Genome-wide association study
RWR:	Random Walk with Restart
PPR:	Personalized PageRank
GPR:	PageRank: global ranking

AUTHOR CONTRIBUTIONS

Xiujuan Lei, Ali Ghulam, and Chen Bian jointly contributed to the design of the study. Xiujuan Lei conceptualized the review and finalized the manuscript. Ali Ghulam wrote the initial manuscript. Min Guo helped to draft the manuscript. Chen Bian revised the manuscript and polished the expression of English. All of the authors have read and approved the final manuscript.

REFERENCES

[1] A. P. Davis, C. J. Grondin, K. Lennon-Hopkins, C. Saraceni-Richards, D. Sciaky, B. L. King, T. C. Wiegers, and C. J. Mattingly, ''The comparative toxicogenomics Database's 10th year anniversary: Update 2015,'' *Nucleic Acids Res.*, vol. 43, no. D1, pp. D914–D920, Jan. 2015, doi: 10.1093/nar/gku935.

[2] M. Agrawal, M. Zitnik, and J. Leskovec, "Large-scale analysis of disease pathways in the human interactome," in *Proc. Biocomputing*, Jan. 2018, pp. 111–122, doi: 10.1142/9789813235533_0011.

[3] L. Yu and L. Gao, "Human pathway-based disease network," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 16, no. 4, pp. 1240–1249, Jul. 2019, doi: 10.1109/TCBB.2017.2774802.

[4] L. Li, Y. Wang, L. An, X. Kong, and T. Huang, "A network-based method using a random walk with restart algorithm and screening tests to identify novel genes associated with Meniáre's disease," *PLoS ONE*, vol. 12, no. 8, Aug. 2017, Art. no. e0182592, doi: 10.1371/journal.pone.0182592.

[5] Y. Liu and M. R. Chance, "Pathway analyses and understanding disease associations," *Current Genetic Med. Rep.*, vol. 1, no. 4, pp. 230–238, Dec. 2013, doi: 10.1007/s40142-013-0025-3.

[6] R. M. Cantor, K. Lange, and J. S. Sinsheimer, "Prioritizing GWAS results: A review of statistical methods and recommendations for their application," *Amer. J. Hum. Genet.*, vol. 86, no. 1, pp. 6–22, Jan. 2010, doi: 10.1016/j.ajhg.2009.11.017.

[7] B. Bakir-Gungor and O. U. Sezerman, "A new methodology to associate SNPs with human diseases according to their pathway related context," *PLoS ONE*, vol. 6, no. 10, Oct. 2011, Art. no. e26277, doi: 10.1371/journal.pone.0026277.

[8] B. Childs and D. Valle, "Genetics, biology, and disease," *Annu. Rev. Genomics Human Genet.*, vol. 1, no. 1, pp. 1–19, Sep. 2000, doi: 10.1146/annurev.genom.1.1.1.

[9] D. Botstein and N. Risch, "Discovering genotypes underlying human phenotypes: Past successes for mendelian disease, future approaches for complex disease," *Nature Genet.*, vol. 33, no. S3, pp. 228–237, Mar. 2003, doi: 10.1038/ng1090.

[10] J. N. Hirschhorn and M. J. Daly, "Genome-wide association studies for common diseases and complex traits," *Nature Rev. Genet.*, vol. 6, no. 2, pp. 95–108, Feb. 2005, doi: 10.1038/nrg1521.

[11] J. Loscalzo, I. Kohane, and A. Barabasi, "Human disease classification in the postgenomic era: A complex systems approach to human pathobiology," *Mol. Syst. Biol.*, vol. 3, no. 1, p. 124, Jan. 2007, doi: 10.1038/msb4100163.

[12] A.-L. Barabási, "Network medicine—From obesity to the diseasome," *New England J. Med.*, vol. 357, no. 4, pp. 404–407, Jul. 2007, doi: 10.1056/NEJMe078114.

[13] Y. Moreau and L.-C. Tranchevent, "Computational tools for prioritizing candidate genes: Boosting disease gene discovery," *Nature Rev. Genet.*, vol. 13, no. 8, pp. 523–536, Aug. 2012, doi: 10.1038/nrg3253.

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

[15] 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.

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

[17] Y. Li and J. C. Patra, "Genome-wide inferring gene–phenotype relationship by walking on the heterogeneous network," *Bioinformatics*, vol. 26, no. 9, pp. 1219–1224, May 2010, doi: 10.1093/bioinformatics/btq108.

[18] M. A. García-Campos and E.-E. Enrique, "Pathway analysis: State of the art," *Frontiers Physiol.* vol. 6, p. 383, Dec. 2015, doi: 10.3389/fphys.2015.00383.

[19] L. Eronen and H. Toivonen, "Biomine: Predicting links between biological entities using network models of heterogeneous databases," *BMC Bioinf.*, vol. 13, no. 1, p. 119, Dec. 2012.

[20] T. Gaudelet, N. Malod-Dognin, J. Sanchez-Valle, V. Pancaldi, A. Valencia, and N. Przulj, "Unveiling new disease, pathway, and gene associations via multi-scale neural networks," 2019, *arXiv:1901.10005*. [Online]. Available: http://arxiv.org/abs/1901.10005

[21] H. Lee and M. Shin, "Mining pathway associations for disease-related pathway activity analysis based on gene expression and methylation data," *BioData Mining*, vol. 10, no. 1, p. 3, Dec. 2017.

[22] F. Zhang and R. Drabier, "IPAD: The integrated pathway analysis database for systematic enrichment analysis," *BMC Bioinf.*, vol. 13, no. S15, pp. 1–21, Sep. 2012, doi: 10.1186/1471-2105-13-S15-S7.

[23] A. Hamosh, "Online mendelian inheritance in man (OMIM), a knowledgebase of human genes and genetic disorders," *Nucleic Acids Res.*, vol. 33, pp. D514–D517, Dec. 2004, doi: 10.1093/nar/gki033.

[24] C. J. Mattingly, M. C. Rosenstein, A. P. Davis, G. T. Colby, J. N. Forrest, and J. L. Boyer, "The comparative toxicogenomics database: A cross-species resource for building chemical-gene interaction networks," *Toxicol. Sci.*, vol. 92, no. 2, pp. 587–595, Aug. 2006, doi: 10.1093/toxsci/kfl008.

[25] F. Zheng, L. Wei, L. Zhao, and F. Ni, "Pathway network analysis of complex diseases based on multiple biological networks," *BioMed Res. Int.*, vol. 2018, pp. 1–12, Jul. 2018, doi: 10.1155/2018/5670210.

[26] B. Dutta, "PathNet: A tool for finding pathway enrichment and pathway cross-talk using topological information and gene expression data," Ph.D. dissertation, HPC Softw. Appl. Inst. Telemedicine Adv. Technol. Res. Center U.S. Army Med. Res. Materiel Command Ft. Detrick, MD, USA, 2018.

[27] A. G. Cirincione, K. L. Clark, and M. G. Kann, "Pathway networks generated from human disease phenome," *BMC Med. Genomics*, vol. 11, no. S3, p. 75, Sep. 2018, doi: 10.1186/s12920-018-0386-2.

[28] P. Ni, J. Wang, P. Zhong, Y. Li, F. Wu, and Y. Pan, "Constructing disease similarity networks based on disease module theory," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, early access, Mar. 21, 2018, doi: 10.1109/TCBB.2018.2817624.

[29] Y.-W. Niu, H. Liu, G.-H. Wang, and G.-Y. Yan, "Maximal entropy random walk on heterogenous network for MIRNA-disease association prediction," *Math. Biosci.*, vol. 306, pp. 1–9, Dec. 2018.

[30] Z. Gao, Y.-T. Wang, Q.-W. Wu, J.-C. Ni, and C.-H. Zheng, "Graph regularized l2,1-nonnegative matrix factorization for miRNA-disease association prediction," *BMC Bioinf.*, vol. 21, no. 1, pp. 10–16, Dec. 2020.

[31] D. Seo, M.-H. Lee, and S. Yu, "Development of network analysis and visualization system for KEGG pathways," *Symmetry*, vol. 7, no. 3, pp. 1275–1288, 2015, doi: 10.3390/sym7031275.

[32] E. Schmidt, "Reactome—A Knowledgebase of biological pathways," *Nucleic Acids Res.*, vol. 2, pp. 428–432, Jan. 2005, doi: 10.1093/nar/gki072.

[33] I. Vastrik, P. D'Eustachio, E. Schmidt, G. Gopinath, D. Croft, B. de Bono, M. Gillespie, B. Jassal, S. Lewis, L. Matthews, G. Wu, E. Birney, and L. Stein, "Reactome: A knowledge base of biologic pathways and processes," *Genome Biol.*, vol. 10, no. 2, p. 402, 2009.

[34] A. P. Davis, C. G. Murphy, C. A. Saraceni-Richards, M. C. Rosenstein, T. C. Wiegers, and C. J. Mattingly, "Comparative toxicogenomics database: A knowledgebase and discovery tool for chemical-gene-disease networks," *Nucleic Acids Res.*, vol. 37, pp. D786–D792, Jan. 2009, doi: 10.1093/nar/gkn580.

[35] W. Ding, H. Yang, S. Gong, W. Shi, J. Xiao, J. Gu, Y. Wang, and B. He, "Candidate miRNAs and pathogenesis investigation for hepatocellular carcinoma based on bioinformatics analysis," *Oncol. Lett.*, vol. 13, no. 5, pp. 3409–3414, May 2017, doi: 10.3892/ol.2017.5913.

[36] G. Hernandez-Suarez, M. Sanabria, M. Serrano, J. Zabaleta, and A. Tenesa, "Abstract 4840: TGFBR1 and TP53 SNPs interactions associated with colorectal cancer risk: Analysis of metabolic pathways using a random forest approach," in *Proc. Epidemiology*, Apr. 2013, p. 840, doi: 10.1158/1538-7445.

[37] R. Krishnappa, "Molecular expression profiling with respect to KEGG hsa05219 pathway," *Ecancermedical Sci.* vol. 5, no. 1, p. 189, 2011, doi: 10.3332./ecancer.2011.189.

[38] C. Wu, R. Gao, D. Zhang, S. Han, and Y. Zhang, "PRWHMDA: Human microbe-disease association prediction by random walk on the heterogeneous network with PSO," *Int. J. Biol. Sci.*, vol. 14, no. 8, pp. 849–857, 2018, doi: 10.7150/ijbs.24539.

[39] Y. Li and P. Agarwal, "A pathway-based view of human diseases and disease relationships," *PLoS ONE*, vol. 4, no. 2, Feb. 2009, Art. no. e4346, doi: 10.1371/journal.pone.0004346.

[40] M. Tanabe and M. Kanehisa, "Using the KEGG database resource," *Current Protocols Bioinf.*, vol. 38, no. 1, pp. 1–5, Jun. 2012, doi: 10.1002/0471250953.bi0112s38.

[41] X. Chen and L. Wang, "Predicting miRNA-disease association based on inductive matrix completion," *Bioinformatics*, vol. 34, no. 24, pp. 4256–4265, 2018, doi: 10.1093/bioinformatics/bty503.

[42] X. Chen, C. C. Yan, X. Zhang, X. Zhang, F. Dai, J. Yin, and Y. Zhang, "Drug–target interaction prediction: Databases, Web servers and computational models," *Briefings Bioinf.*, vol. 17, no. 4, pp. 696–712, Jul. 2016.

[43] X. Chen, B. Ren, M. Chen, Q. Wang, L. Zhang, and G. Yan, "NLLSS: Predicting synergistic drug combinations based on semi-supervised learning," *PLOS Comput. Biol.*, vol. 12, no. 7, 2016, Art. no. e1004975.

[44] X. Chen, C. C. Yan, X. Zhang, and Z.-H. You, "Long non-coding RNAs and complex diseases: From experimental results to computational models," *Briefings Bioinf.*, vol. 18, no. 4, pp. 558–576, Jul. 2017, doi: 10.1093/bib/bbw060.
[45] 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, p. 12, Jan. 2010.
[46] X. Shen, Y. Chen, X. Jiang, X. Hu, T. He, and J. Yang, "Predicting disease-microbe association by random walking on the heterogeneous network," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Dec. 2016, pp. 4–771.
[47] S. Zou, J. Zhang, and Z. Zhang, "A novel approach for predicting microbe-disease associations by bi-random walk on the heterogeneous network," *PLoS ONE*, vol. 12, no. 9, 2018, Art. no. e0184394.

ALI GHULAM is currently pursuing the Ph.D. degree with the School of Computer Science, Shaanxi Normal University, Xian, China. His research interests include human disease pathway network modeling and biological pathway databases discovery.

XIUJUAN LEI (Member, IEEE) received the Ph.D. degree from Northwestern Polytechnical University, in 2005. She is currently a Professor and a Ph.D. Supervisor with Shaanxi Normal University. Her research interests include bioinformatics and intelligent computing.

MIN GUO received the Ph.D. degree from Shaanxi Normal University, Shaanxi, China, in 2003. She is currently a Professor and a Ph.D. supervisor with Shaanxi Normal University. Her main research interests include image processing, pattern recognition, and intelligent information processing.

CHEN BIAN is currently pursuing the master's degree with the School of Computer Science, Shaanxi Normal University, Xian, China. Her major is bioinformatics.