



Article

An Entropy-Based Directed Random Walk for Cancer Classification Using Gene Expression Data Based on Bi-Random Walk on Two Separated Networks

Xin Hui Tay ¹, Shahreen Kasim ^{1,*} , Tole Sutikno ², Mohd Farhan Md Fudzee ¹, Rohayanti Hassan ³, Emelia Akashah Patah Akhir ⁴, Norshakirah Aziz ⁴ and Choon Sen Seah ⁵ 

¹ Faculty of Computer Sciences and Information Technology, Universiti Tun Hussein Onn Malaysia, Batu Pahat 83000, Malaysia

² Department of Electrical Engineering, Universitas Ahmad Dahlan, Yogyakarta 55166, Indonesia

³ School of Computing, Faculty of Electrical Engineering, Universiti Teknologi Malaysia, Skudai 81310, Malaysia

⁴ Department of Computer and Information Sciences, Universiti Teknologi Petronas, Seri Iskandar 32610, Malaysia

⁵ Faculty of Accounting & Management, Universiti Tunku Abdul Rahman, Kajang 43000, Malaysia

* Correspondence: shahreen@uthm.edu.my

Abstract: The integration of microarray technologies and machine learning methods has become popular in predicting the pathological condition of diseases and discovering risk genes. Traditional microarray analysis considers pathways as a simple gene set, treating all genes in the pathway identically while ignoring the pathway network's structure information. This study proposed an entropy-based directed random walk (e-DRW) method to infer pathway activities. Two enhancements from the conventional DRW were conducted, which are (1) to increase the coverage of human pathway information by constructing two inputting networks for pathway activity inference, and (2) to enhance the gene-weighting method in DRW by incorporating correlation coefficient values and *t*-test statistic scores. To test the objectives, gene expression datasets were used as input datasets while the pathway datasets were used as reference datasets to build two directed graphs. The within-dataset experiments indicated that e-DRW method demonstrated robust and superior performance in terms of classification accuracy and robustness of the predicted risk-active pathways compared to the other methods. In conclusion, the results revealed that e-DRW not only improved the prediction performance, but also effectively extracted topologically important pathways and genes that were specifically related to the corresponding cancer types.

Keywords: directed random walk; pathway-based analysis; cancer classification



Citation: Tay, X.H.; Kasim, S.; Sutikno, T.; Fudzee, M.F.M.; Hassan, R.; Patah Akhir, E.A.; Aziz, N.; Seah, C.S. An Entropy-Based Directed Random Walk for Cancer Classification Using Gene Expression Data Based on Bi-Random Walk on Two Separated Networks. *Genes* **2023**, *14*, 574. <https://doi.org/10.3390/genes14030574>

Academic Editors: Yuanyan Xiong and Hui Zhang

Received: 2 February 2023

Revised: 19 February 2023

Accepted: 23 February 2023

Published: 24 February 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The accurate prediction of prognosis and the metastatic potential of cancer is a major challenge in clinical cancer research. Through the evolution of high-throughput technologies, deoxyribonucleic acid (DNA) microarray analysis can classify tumor samples overriding the traditional diagnostic methods. This technology allows for the extraction of a huge amount of molecular information, which aids in the discovery of tumor-specific biomarkers. However, the reproducibility of individual gene biomarkers has been challenging, as the identified gene markers in one dataset failed to predict the same disease phenotype obtained in other datasets [1]. This discrepancy is usually due to the cellular heterogeneity within tissues, the inherent genetic heterogeneity across patients, and the measurement error in microarray platforms [2]. Furthermore, microarray analysis of gene expression data generally produces plenty of genes from patients with the same diseases, hence, leading to a high dimension small sample size problem. All of these factors often

decrease the prediction performance and reproducibility of individual gene biomarkers in independent cohorts of patients.

To address the unreliable or inconsistent prediction of gene biomarkers in datasets, biological pathway data were introduced to identify robust pathway biomarkers in functional categories [3–6]. As gene products are known to function coordinately in functional modules, the mutual interest between the pathway data and gene expression data can extract function-related genes to produce consistent and reproducible biomarkers [7]. Such biomarkers at the functional level can reduce the impact of noise in the microarray data by allowing for a more accurate biological interpretation of the disease-canonical pathway correlations [2]. In fact, several studies [3,4,8,9] have shown that pathway markers are more reliable compared to single gene markers as they provide crucial biological insights into the underlying processes that give rise to various disease phenotypes. Furthermore, pathway-based classifiers often achieve comparable or better classification performance compared to conventional gene-based classifiers [4,5].

In cancer classification, a robust gene weight merit is vital to reflect the importance of genes from different aspects and establish significant genes with related diseases [10]. Several studies in pathway-based methods typically use the *t*-test as the gene weighting method to measure the gene expression levels for further cancer classification. Consider that those existing pathway-based methods including directed random walk (DRW), significant DRW, and pathway activity inference using condition-responsive genes (PAC method) all targeted on the *t*-test as the single statistical measurement to weigh each gene in the gene expression data. However, the lack of a comprehensive gene weighting method could affect the classification performance of pathway-based methods [10,11].

To combat the aforementioned issue, this study proposed an entropy-based directed random walk (e-DRW) on two separated biological networks that enhances the accuracy of cancer classification. Two inputting networks were proposed for the random walking of e-DRW, which comprises 328 KEGG pathways collected from the KEGG database [12] and 208 pathways gathered from the Pathway Interaction Database (PID) [13]. The representation of two biological networks (KEGG-PID) is known as the directed pathway network. An improved gene weighting strategy using point biserial correlation (PBC) coefficients and the T-test was proposed in e-DRW. The gene weighting method modeled the combined effect of the statistical measurement of the gene expression levels with the class label (normal, cancer). The weight initialization of genes and the scoring of pathways were further enhanced by the application of the entropy metric to calculate the pathway activity score. The proposed method was implemented in the R platform with version 4.2.1 in 64-bit using Windows 10 (refer the supplementary e-DRW R package for more details). Figure 1 shows the workflow of e-DRW. The steps involved in e-DRW include data pre-processing and the construction of biological networks, normalization based on z-scores, differential expression analysis, entropy-based directed random walk, entropy-based pathway activity inference, and classification.

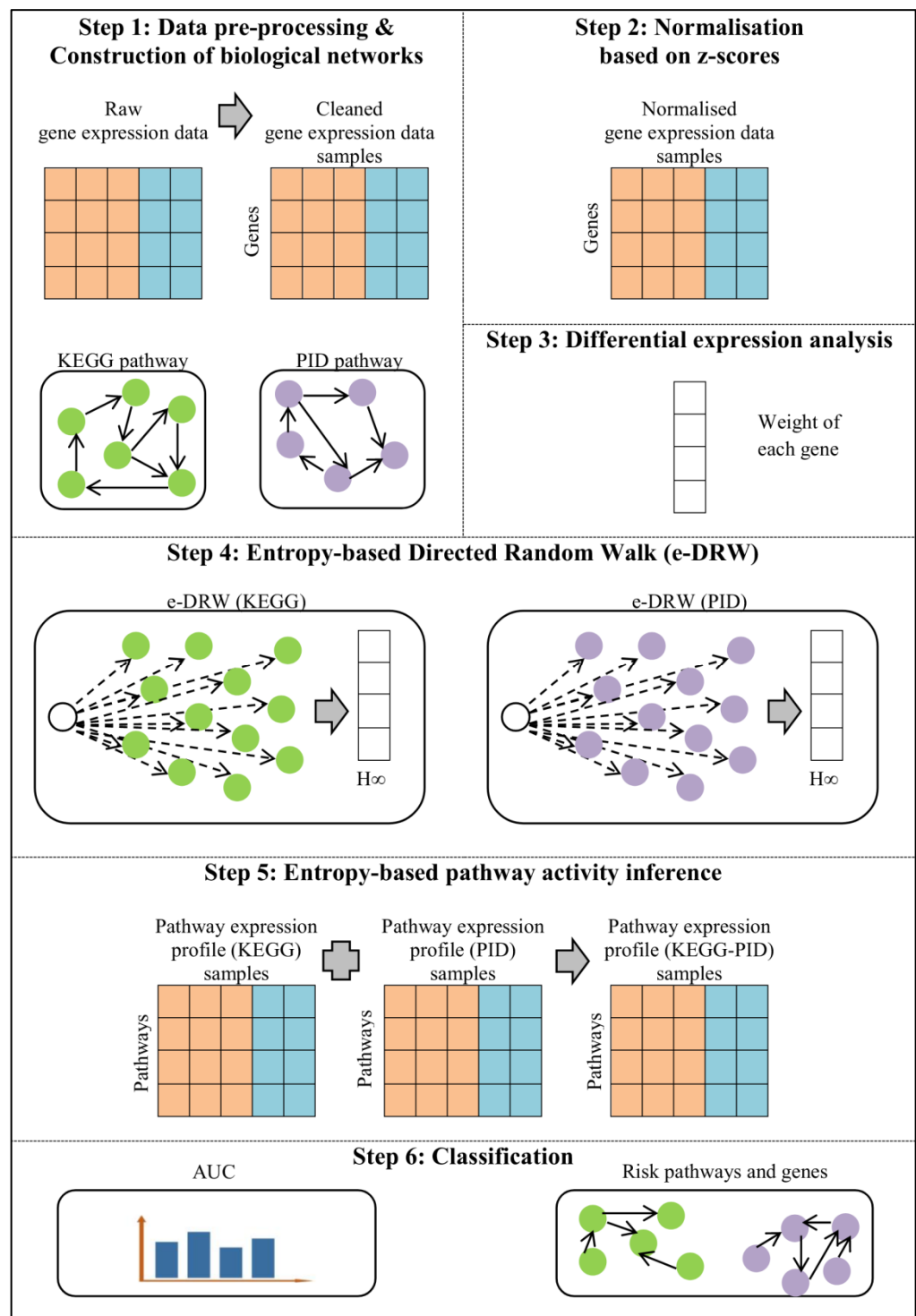


Figure 1. Workflow of e-DRW.

2. Materials and Methods

This section presents the materials and methodology used in e-DRW. Based on Figure 1, each of the steps involved in the workflow of e-DRW will be described thoroughly.

2.1. Data Pre-Processing and Construction of Biological Networks

The first step in e-DRW is data pre-processing and the construction of biological networks. Six gene expression datasets were obtained from the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) database, which are lung [14],

stomach [15], liver [16], kidney [17], thyroid [18], and breast [19] cancer datasets. The collected datasets undergo data pre-processing to produce the cleaned gene expression data. There are two phases involved in data pre-processing: (i) data cleaning and imputation, and (ii) the normalization of the gene expression data. In the first phase, the unwanted and empty values of the attributes were removed. The unwanted attributes include patient biological information and dataset information that is not applicable in cancer classification whereas empty values of attributes refer to the missing values that appeared across the rows in the gene expression dataset. Then, rows with incomplete values of attributes were imputed with mean values to resolve the inconsistencies in data. The completed dataset following the application of the mean imputation was used for inference. However, the rearrangement of data was run through before proceeding to the next phase. The normalization step in the second phase typically included thresholds or flooring to remove poorly detected probes and log2 transformation to normalize the distribution of probes across the intensity range of the experiment. Gene Pattern was used for dataset pre-processing to remove platform noise and genes that have little variation [20]. Table 1 shows the details of the datasets after data preprocessing.

Table 1. Gene expression datasets after pre-processing.

Cancer	GEO ID	Platform ID	Number of Genes		Number of Cancerous Samples	Number of Normal Samples	Total Number of Samples
			Raw	Cleaned			
Lung	GSE10072	GPL96	22283	12986	58	49	107
Stomach	GSE13911	GPL570	54675	12419	38	31	69
Liver	GSE17856	GPL6480	25075	13802	43	44	87
Kidney	GSE15641	GPL96	22283	11593	69	23	92
Thyroid	GSE33630	GPL570	54675	12986	60	45	105
Breast	GSE3494	GPL96	22283	12986	60	176	236

On the other hand, a directed pathway network was constructed based on the pathway information obtained from the KEGG database and PID database. First, each KEGG pathway was converted into a directed graph using the NetPathMiner [21] software package. A total of 328 human pathways were merged to form the KEGG network, covering 6667 nodes and 116,773 directed edges. Subsequently, each PID pathway was converted into another directed graph using the PaxtoolsR [22] software package. A total of 208 human pathways were merged to form the PID network, covering 2817 nodes and 39,289 directed edges. Each node in the graph represented a gene, while each directed edge represented how the genes interacted and controlled each other. The direction of the edge was determined by the type of interaction between the two genes found in the KEGG and PID pathway databases.

2.2. Normalization Based on Z-Scores

The second step in e-DRW is normalization based on z-scores. The collected gene expression datasets underwent normalization based on the z-score to produce the normalized gene expression data. This step aimed to normalize the gene expression values over all samples to a scale of mean zero and variance one [2,23]. Normalization based on z-scores can provide a way to standardize data across the gene expression dataset. This is an important step to achieving good classification performance before evaluating the data on machine learning algorithms [24]. The formula of normalization based on z-scores is shown as below:

$$z(gi) = \frac{gene(gi) - \bar{X}(gi)}{S(gi)} \quad (1)$$

where $z(gi)$ is the normalized gene expression values for gene i over all samples; $gene(gi)$ is the gene expression values for gene i over all samples; $\bar{X}(gi)$ is the mean of gene expression values for gene i ; $S(gi)$ is the standard deviation of the gene expression values for gene i ; and i is the number of genes in the gene expression data.

2.3. Differential Expression Analysis

The third step in e-DRW is differential expression analysis. t -test statistics with equal variances [25] and the point biserial correlation (PBC) coefficient [26] were calculated for each gene in the gene expression data. This step aimed to calculate the statistical difference between the normal and disease samples. The formula for calculating the t -test statistics with equal variances of each gene is shown below:

$$t(gi) = \frac{\bar{X}_1 - \bar{X}_2}{S \times \sqrt{\left(\frac{1}{N_1}\right) + \left(\frac{1}{N_2}\right)}} \quad (2)$$

where $t(gi)$ is the t -score of the t -test statistics with equal variances, \bar{X}_1 is the mean for the normal sample; \bar{X}_2 is the mean for the tumor sample; S is the standard deviation of the two samples; N_1 is the number of normal samples; and N_2 is the number of tumor samples. At the same time, PBC was performed to calculate the correlation coefficient for each gene. PBC measures the relationship between two variables (genes) using the formula shown below:

$$p_{pb}(gi) = \frac{M_1 - M_2}{S} \sqrt{pq} \quad (3)$$

where $p_{pb}(gi)$ is the PBC coefficient for each gene gi ; M_1 is the mean for normal samples; M_2 is the mean for the tumor sample, S is the standard deviation of the normal and tumor samples; p is the proportion of cases in normal samples; and q is the proportion of cases in the tumor samples.

This study employed a combination of PBC and t -test scores (PCT scores) as a gene-weighting method. The weighted expressions of the member genes reflected two factors: (1) the degree of the differential expression of genes between the means of the normal and cancer group; and (2) the correlation between a gene expression and class label (normal, cancer). Based on these considerations, a new robust gene-weighting method was proposed in this study. The normalized expression values of gene gi in sample k are defined as:

$$Z(gi) = t(gi)^2 + |p(gi)| \quad (4)$$

where $t(gi)$ is the t -score of gene gi calculated using a two-tailed t -test between two phenotypes, while $\rho(gi)$ is the absolute PBC between gene gi and the class label. $Z(gi)$ represents the weighted normalized expression (PCT scores) of gene gi in sample k , reflecting the differential expression degree of gene gi and its correlation with the phenotype. Larger expression values $Z(gi)$ can be related to higher differential expression and a larger correlation with the phenotype.

2.4. Entropy-Based Directed Random Walk (e-DRW)

The fourth step in e-DRW is the calculation of the genes' weight in the directed graph. Before implementing e-DRW, the initial weight of the genes was first calculated using the formula shown below:

$$W_0 = \frac{\text{absolute}(Z(gi)) - \text{maximum}(Z(gi))}{\text{maximum}(Z(gi)) - \text{minimum}(Z(gi))} \quad (5)$$

where W_0 is the initial weight of genes; $\text{absolute}(Z(gi))$ is the absolute values of PCT score; $\text{maximum}(Z(gi))$ is the maximum values of PCT score; and $\text{minimum}(Z(gi))$ is the minimum values of the PCT score. Then, the entropy [27] of each gene was used as the weight parameter to calculate the distribution of each node in the directed graph. Furthermore, the directed graphs for the KEGG and PID networks were converted to an entropy edge-weighted adjacency matrix (network entropy) to enhance the calculation of the genes' weight in both networks. The calculated node entropy for each gene, KEGG, and PID

network entropy were then implemented in e-DRW for the pathway activity inference. e-DRW is defined as:

$$H_{t+1} = (1-r)E^T H_t + rH_0 \quad (6)$$

where H_t represents the node entropy vector that holds the probability at the specific node at time step t . H_0 is the initial entropy probability vector; E^T is an entropy edge-weighted adjacency matrix developed from the directed graphs (with edges); r denotes the restart probability ranges from 0.1 to 0.9; and H_{t+1} denotes the final entropy probability vector.

Considering the bi-random walk of e-DRW on two inputting networks (KEGG and PID networks), the random walking of e-DRW was implemented on the two networks successively to obtain the separate results. The random walk processes are illustrated by the following equations:

$$\text{KEGG network : } H_{t+1}^G = (1-r)G^E H_t + rH_0 \quad (7)$$

$$\text{PID network : } H_{t+1}^P = (1-r)P^E H_t + rH_0 \quad (8)$$

where G^E represents the entropy edge-weighted adjacency matrix of the KEGG network, and P^E represents the entropy edge-weighted adjacency matrix of the PID network. The separate results were then applied for further pathway activity inference and cancer classifications.

2.5. Entropy-Based Pathway Activity Inference

The fifth step in e-DRW is entropy-based pathway activity inference. The normalized gene expression data were first split into three subsets whereby 60% of the datasets was used as the training set, 20% used as the validation sets, and another 20% used as the test sets. The three subsets were then utilized for entropy-based pathway activity inference. Genes with p -values less than 0.05 for each pathway in the pathway data were chosen to construct the pathway activities. Entropy-based pathway activity inference for the training, validation, and test sets for the KEGG and PID networks is defined as:

$$a(P_j) = \frac{\sum_{i=1}^{n_j} H\infty(gi) \times PCTscore(gi) \times Z(gi)}{\sqrt{\sum_{i=1}^{n_j} (H\infty(\frac{1-gi}{sum(1-gi)}))^2}} \quad (9)$$

where $a(P_j)$ is the pathway activity (or expression value vector); $H\infty$ is the output of genes (or weight vector calculated from e-DRW); $PCTscore(gi)$ is the summation of PBC between gene gi and class label (normal and tumor samples) and the t -test statistics of gene gi from a two-tailed t -test with equal variances in the expression values between two classes. $Z(gi)$ is a normalized value vector of gene gi across the whole dataset, and $H\infty(\frac{1-gi}{sum(1-gi)})$ is the entropy weight of gene gi . The calculated KEGG and PID pathway expression profiles for the training, validation, and test sets were then combined respectively for the pathway selections. The top 50 pathways ranked by the t -test statistics for the training, validation, and test sets were selected to construct the final pathway expression profiles for further classification.

2.6. Classification

The final step in e-DRW is classification. Within-dataset experiments were implemented for the six cancer datasets. The R caret [28] package was utilized to obtain the classification accuracy. Three classifiers were selected to evaluate the performance of e-DRW, which were Naïve Bayes (NB), K-nearest neighbors (KNN), and logistic regression (GLM). e-DRW implemented stratified 10-fold cross validation on the training set to evaluate the performance of the classifier. The top 50 pathways in the training dataset were used as candidate features to build the model. Subsequently, pathways were added sequentially to train the model. The performance of the classifier was measured by evaluating the area under the receiver operating characteristic curve (AUC). The added pathway marker was

maintained in the feature set if the AUC increased, but was removed if otherwise [2]. This process was repeated for the top 50 pathway markers to optimize the classifier and to yield the best feature set. The performance of the optimized classifier was evaluated on the test set using pathway markers from the best feature set. This process was repeated 10 times to ensure unbiased evaluation and to estimate the variation of the AUC. As the final step, the mean AUC across 10 classifiers was estimated to represent the overall performance of the classification method.

3. Results

This section presents the classification performance within-dataset experiments. For comparison with other pathway activity inference methods, five pathway-based classification methods were chosen, namely, the DRW method [2], sDRW method [29], iDRW method [30], PAC method [4], and principal component analysis (PCA method) [7]. The experimental setting was the same for the DRW, sDRW, iDRW, and PAC methods. The PCA method was implemented as the pathway-based classification method by applying the proposed KEGG network to calculate the pathway expression profiles. Classification accuracy and robustness of the predicted risk-active pathways were chosen as the performance measurements of cancer classification.

3.1. Classification Performance on Within-Dataset Experiments

Table 2 presents the mean AUCs of e-DRW with varying restart probabilities (0.1–0.9) across the six cancer datasets using three different classifiers (NB, KNN, LR).

Table 2. Mean AUC of e-DRW.

Restart Probabilities	Classifiers	Datasets					
		Lung	Stomach	Liver	Kidney	Thyroid	Breast
0.1	NB	0.878505	0.846377	0.87931	0.866304	0.939048	0.762712
	KNN	0.918692	0.858066	0.855172	0.838043	0.946667	0.751695
	LR	0.864486	0.795652	0.871264	0.893478	0.875238	0.761017
0.2	NB	0.917757	0.866667	0.872414	0.820652	0.949524	0.75678
	KNN	0.929907	0.915942	0.870115	0.834783	0.917143	0.762288
	LR	0.860748	0.797101	0.851724	0.846739	0.912381	0.758051
0.3	NB	0.966355	0.86087	0.885057	0.802174	0.957143	0.761017
	KNN	0.946729	0.933333	0.931034	0.858696	0.954286	0.760169
	LR	0.873832	0.797101	0.918391	0.868478	0.939048	0.761017
0.4	NB	0.935514	0.926087	0.87931	0.873913	0.96	0.755932
	KNN	0.948598	0.908696	0.855172	0.883696	0.950476	0.762712
	LR	0.914953	0.842029	0.871264	0.926087	0.898095	0.751695
0.5	NB	0.962617	0.905797	0.83908	0.83913	0.950476	0.752542
	KNN	0.980374	0.886957	0.855172	0.897826	0.954286	0.769068
	LR	0.899065	0.831884	0.837931	0.871739	0.931429	0.755085
0.6	NB	0.969159	0.917391	0.874713	0.863043	0.937143	0.758898
	KNN	0.971028	0.955072	0.873563	0.829348	0.868571	0.75678
	LR	0.909346	0.868116	0.862069	0.891304	0.86	0.758475
0.7	NB	0.969159	0.849275	0.868966	0.804348	0.941905	0.761864
	KNN	0.961682	0.83913	0.906897	0.795652	0.94	0.747034
	LR	0.903738	0.792754	0.851724	0.823913	0.900952	0.760593
0.8	NB	0.961682	0.897101	0.918391	0.856522	0.942857	0.754237
	KNN	0.930841	0.892754	0.905747	0.88913	0.921905	0.751271
	LR	0.87757	0.785507	0.827586	0.861957	0.932381	0.755085
0.9	NB	0.931776	0.894203	0.928736	0.86413	0.950476	0.751695
	KNN	0.919626	0.926087	0.910345	0.906522	0.950476	0.747034
	LR	0.909346	0.857971	0.931034	0.894565	0.900952	0.753814

Bold values: the highest values. Refer to Supplementary Table S1 for more details.

Based on Table 3, the KNN classifier showed the highest mean AUCs across four cancer datasets, which were the lung cancer dataset, stomach cancer dataset, liver cancer dataset, and breast cancer dataset. Hence, the KNN classifier was chosen to evaluate the classification performance of the other methods. For a fair and effective comparison with other methods, within-dataset experiments similar to those used in Liu et al. (2013) [2] were implemented to evaluate the classification performance. Figure 2 illustrates the comparison of the classification performance for the six methods on the within-dataset experiments.

Table 3. Mean AUC of e-DRW.

Risk Pathways	Datasets					
	Lung	Stomach	Liver	Kidney	Thyroid	Breast
PI3K-Akt signaling pathway	✓*	✓	✓	✓	✓	
Pathways in cancer	✓	✓	✓	✓	✓	
Human papillomavirus infection	✓	✓	✓			
Calcium signaling pathway				✓	✓	✓
ECM-receptor interaction	✓				✓	
Lipid and atherosclerosis	✓					✓
Apelin signaling pathway	✓			✓		
Focal adhesion	✓				✓	
Hippo signaling pathway		✓				✓
Wnt signaling pathway		✓				✓
Small cell lung cancer			✓		✓	
Neuroactive ligand–receptor interaction				✓	✓	
Integrin-linked kinase signaling		✓		✓		
Adrenergic signaling in cardiomyocytes				✓		✓
cGMP-PKG signaling pathway				✓		✓

* Detected significant pathways.

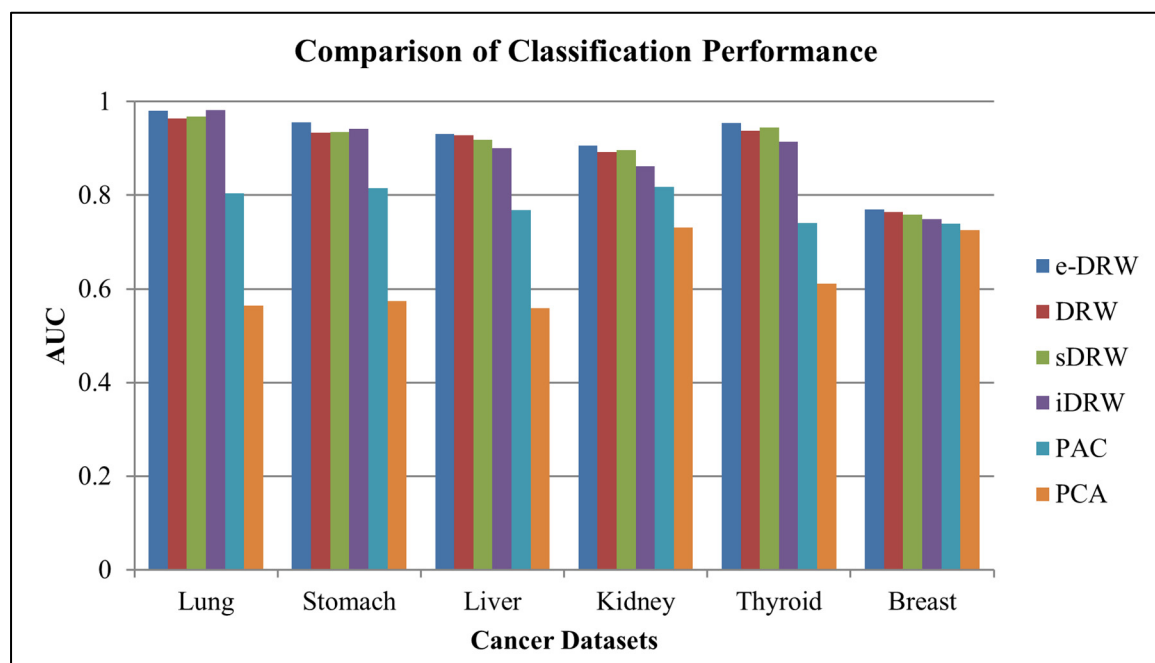


Figure 2. Comparison of classification performance.

Based on Figure 2, the proposed e-DRW method obtained mean AUCs of 0.980374 for the lung cancer dataset, 0.95072 for the stomach cancer dataset, 0.931034 for the liver cancer dataset, 0.906522 for the kidney cancer dataset, 0.954286 for the thyroid cancer dataset, and 0.769068 for the breast cancer dataset. By comparing the mean AUCs with other methods,

e-DRW achieved the highest mean AUCs across all datasets, except for the lung cancer dataset, which was slightly lower than the iDRW (0.981259) method. This indicates that e-DRW-based pathway markers are quite competent in discriminating between different disease phenotypes. It also demonstrated the best overall classification performance on the within-dataset experiments when compared with other methods.

3.2. Robustness of Predicted Risk-Active Pathways

The detection of robust risk-active pathways is important in cancer studies. Risk-active pathways detected across 10 experiments for each cancer dataset are provided in Table S2. Genes in the risk-active pathways were extracted and provided in Table S3. Table 3 lists the top 15 most predicted cancer-related pathways involved in various biological processes studied by e-DRW across the six datasets.

Based on Table 3, pathways specific to the phenotype of classification were identified. Among these pathways, the PI3K-Akt signaling pathway and pathways in cancer identified in most cancer datasets reported the relations of these pathways with cancer [2,31–33]. Furthermore, the human papillomavirus infection pathway and calcium signaling pathway are known cancer pathways, as reported in multiple studies [34–39]. The ECM–receptor interaction pathway and focal adhesion pathway also suggest their important roles in lung and thyroid cancer based on pertinent studies [40,41]. In addition, several extensively researched cancer-related pathways were identified as risk-active pathways in multiple cancers such as the lipid and atherosclerosis pathway, Apelin signaling pathway, Hippo signaling pathway, and Wnt signaling pathway [42–49]. Furthermore, the predictions of the small cell lung cancer pathway, neuroactive ligand–receptor interaction pathway, and integrin-linked kinase signaling pathway were consistent with several studies [50–55]. Relevant studies have validated the coactive effect of adrenergic signaling in the cardiomyocyte pathway and the cGMP-PKG signaling pathway with cancers [56–58].

4. Discussion

In the literature, multiple existing pathway-based methods incorporate pathway topological information to identify important genes within pathways. For instance, Guo et al. [3] employed the mean or median expression value of the member genes to infer the pathway activity. Bild et al. [7] used the first principal component of the expression profile of member genes to evaluate the activity of a given pathway (PCA method). Lee et al. [4] proposed pathway activity inference using only a subset of genes in the pathway, called the condition responsive genes (CORGs), in which the combined expression levels can accurately discriminate the phenotypes of interest (PAC method). However, these methods simply consider pathways as simple gene sets but ignore significant individual genes and interactions between genes, which are essential to infer a more robust pathway activity [1].

A comprehensive pathway topology is important to clarify the roles that the genes play in the pathway and weight the genes more precisely [2]. Several pathway-based methods utilize pathway topology information collected from pathway databases for analysis. For example, Liu et al. [2] constructed the global-directed pathway network, which covers 300 pathways collected from the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway database. Seah et al. [29] built a directed graph using 300 pathway datasets obtained from the KEGG pathway database. Kim et al. [30] selected 327 human pathways to construct a directed gene–gene graph for pathway activity inference. Lee et al. [4] collected 472 canonical metabolic and signaling pathways from MsigDB v1.0 for cancer classification. However, the limitations of these methods lie in the coverage of human pathway information. The complete biological pathway information not only enables a more accurate prediction of disease status, but also paves the way to unveiling novel functional pathways or complexes [59–62].

This study proposed an entropy-based pathway activity inference scheme to identify reproducible pathway biomarkers for clinical cancer applications. Previous literature has revealed that individual gene markers are less reliable compared to pathway markers,

and thus are unable to effectively capture the biological interpretation of gene expression in functional categories [3–6]. The proposed entropy-based pathway activity inference method conducted a bi-random walk of e-DRW on two separated networks for pathway activity inference. A robust gene-weighting method was proposed that incorporates PBC and the *t*-test to calculate the weight of each gene. Considering the effectiveness of entropy as weight variables, entropy was implemented as a weight parameter to enhance the weight initialization scheme in e-DRW [63]. The entropy weight metric was also applied in entropy-based pathway activity inference to enhance e-DRW pathway activities for cancer classification.

Based on the classification performance, the mean AUCs of the e-DRW method were significantly higher and more robust across the experiments. The reliable performance of the e-DRW pathway activities could be attributed to the construction of the directed pathway network and gene-weighting method. The proposed biological networks provide larger pathway topology for the random walking of e-DRW on the KEGG and PID networks. Furthermore, the gene-weighting method based on PBC and the *t*-test can greatly magnify the signals of essential genes whose expression levels may have a large impact on the pathway while weakening the differential expression of genes that only appear downstream or have a minor impact on the system. Therefore, the e-DRW approach could alleviate the noise caused by sample heterogeneity or technical measurements, resulting in more reproducible pathway activities.

Moreover, the mean AUCs of e-DRW were better in terms of cancer classification due to higher accuracy compared to other pathway-based analysis methods. Results on the top 15 known cancer-related pathways showed that the performance of most pathways was very close to the best performance. This indicates that the proposed e-DRW was even robust on many cancer-related pathways. Additionally, we found that the proposed e-DRW could achieve a satisfactory performance for all datasets through the PI3K-Akt signaling pathway and pathways in cancer. Overall, e-DRW was more effective in pathways and gene prediction as it was more robust compared to the other methods.

5. Conclusions

In cancer studies, an accurate prediction of cancer is crucial for the diagnosis and prognosis of clinical therapy. An e-DRW on two separated networks for cancer classification was proposed. The two enhancements based on Liu et al.'s work [2] and the proposed e-DRW were proven to be effective in inferring pathway activities and accurate cancer classification. The proposed enhancements included (1) the construction of the directed pathway network (KEGG and PID networks), and (2) gene-weighting based on the PBC and *t*-test. Two biological networks (KEGG and PID networks) were constructed to increase the coverage of human pathway information. A gene-weighting method in e-DRW incorporating the *t*-test statistics scores and correlation coefficient values to weigh each gene in the directed pathway network was also proposed. This weighting strategy not only reflects the degree of the differential expression of genes between the normal and cancer groups, but also considers the correlation coefficient values between genes in the gene expression data. Additionally, the weight initialization of genes and the scoring of pathways were further enhanced by the calculation of gene expression entropy, which implicitly increased the accuracy of cancer classification. Finally, stratified 10-fold cross-validation was utilized to train the classifier and classify the significant pathways detected by e-DRW. In conclusion, the proposed approach was more effective and feasible for cancer classification compared to other pathway-based methods.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/genes14030574/s1>, Table S1: The complete mean AUCs of the e-DRW with varying restart probabilities (0.1–0.9) across the six cancer datasets. Table S2: Pathway markers detected across 10 experiments for each cancer dataset. Table S3: The genes in the risk active pathways. e-DRW R package: R source code of e-DRW for analysis.

Author Contributions: Conceptualization, X.H.T., T.S., S.K., M.F.M.F., R.H., E.A.P.A., N.A., and C.S.S.; Methodology, X.H.T., T.S., S.K., M.F.M.F., R.H., E.A.P.A., N.A., and C.S.S.; Software, X.H.T.; Validation, T.S., S.K., M.F.M.F., R.H., E.A.P.A., N.A., and C.S.S.; Formal analysis, X.H.T., T.S., S.K., M.F.M.F., R.H., E.A.P.A., N.A., and C.S.S.; Investigation, X.H.T., T.S., S.K., M.F.M.F., R.H., E.A.P.A., N.A., and C.S.S.; Resources, X.H.T.; Data curation, X.H.T.; Writing—original draft preparation, X.H.T.; Writing—review and editing, X.H.T., S.K., and M.F.M.F.; Visualization, X.H.T.; Supervision, T.S., S.K., M.F.M.F., R.H., E.A.P.A., N.A., and C.S.S.; Project administration, X.H.T.; Funding acquisition, S.K., M.F.M.F., R.H., E.A.P.A., and N.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Fundamental Research Grant Scheme from the Ministry of Higher Education, grant number H888, Universiti Tun Hussein Onn Malaysia REGG FASA 1/2021 (VOT NO. H888), and grant number H995, Universiti Tun Hussein Onn Malaysia.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data analyzed in this paper are available in the Gene Expression Omnibus (GEO) repository at NCBI.

Acknowledgments: The authors would like to thank the Ministry of Education of Malaysia, Universiti Tun Hussein Onn Malaysia (UTHM) under REGG, for their support in making this research a success.

Conflicts of Interest: The authors declare no conflict of interest.

References

- Xu, P.; Zhao, G.; Kou, Z.; Fang, G.; Liu, W. Classification of cancers based on a comprehensive pathway activity inferred by genes and their interactions. *IEEE Access* **2020**, *8*, 30515–30521. [[CrossRef](#)]
- Liu, W.; Li, C.; Xu, Y.; Yang, H.; Yao, Q.; Han, J.; Shang, D.; Zhang, C.; Su, F.; Li, X.; et al. Topologically inferring risk-active pathways toward precise cancer classification by directed random walk. *Bioinformatics* **2013**, *29*, 2169–2177. [[CrossRef](#)] [[PubMed](#)]
- Guo, Z.; Zhang, T.; Li, X.; Wang, Q.; Xu, J.; Yu, H.; Zhu, J.; Wang, H.; Wang, C.; Topol, E.J.; et al. Towards precise classification of cancers based on robust gene functional expression profiles. *BMC Bioinform.* **2005**, *6*, 58. [[CrossRef](#)] [[PubMed](#)]
- Lee, E.; Chuang, H.Y.; Kim, J.W.; Ideker, T.; Lee, D. Inferring pathway activity toward precise disease classification. *PLoS Comput. Biol.* **2008**, *4*, e1000217. [[CrossRef](#)]
- Su, J.; Yoon, B.J.; Dougherty, E.R. Accurate and reliable cancer classification based on probabilistic inference of pathway activity. *PLoS ONE* **2009**, *4*, e8161. [[CrossRef](#)]
- Kim, S.; Kon, M.; DeLisi, C. Pathway-based classification of cancer subtypes. *Biol. Direct* **2012**, *7*, 21. [[CrossRef](#)]
- Bild, A.H.; Yao, G.; Chang, J.T.; Wang, Q.; Potti, A.; Chasse, D.; Joshi, M.B.; Harpole, D.; Lancaster, J.M.; Berchuck, A.; et al. Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature* **2006**, *439*, 353–357. [[CrossRef](#)]
- Rapaport, F.; Zinovyev, A.; Dutreix, M.; Barillot, E.; Vert, J.P. Classification of microarray data using gene networks. *BMC Bioinform.* **2007**, *8*, 35. [[CrossRef](#)]
- Tomfohr, J.; Lu, J.; Kepler, T.B. Pathway level analysis of gene expression using singular value decomposition. *BMC Bioinform.* **2005**, *6*, 225. [[CrossRef](#)]
- Bao, Z.; Zhu, Y.; Ge, Q.; Gu, W.; Dong, X.; Bai, Y. GwSPIA: Improved signaling pathway impact analysis with gene weights. *IEEE Access* **2019**, *7*, 69172–69183. [[CrossRef](#)]
- Fang, Z.; Tian, W.; Ji, H. A network-based gene-weighting approach for pathway analysis. *Cell Res.* **2012**, *22*, 565–580. [[CrossRef](#)] [[PubMed](#)]
- Kanehisa, M.; Goto, S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **2000**, *28*, 27–30. [[CrossRef](#)] [[PubMed](#)]
- Schaefer, C.F.; Anthony, K.; Krupa, S.; Buchhoff, J.; Day, M.; Hannay, T.; Buetow, K.H. PID: The pathway interaction database. *Nucleic Acids Res.* **2009**, *37* (Suppl. 1), D674–D679. [[CrossRef](#)]
- Landi, M.T.; Dracheva, T.; Rotunno, M.; Figueroa, J.D.; Liu, H.; Dasgupta, A.; Mann, F.E.; Fukuoka, J.; Hames, M.; Bergen, A.W.; et al. Gene expression signature of cigarette smoking and its role in lung adenocarcinoma development and survival. *PLoS ONE* **2008**, *3*, e1651. [[CrossRef](#)] [[PubMed](#)]
- D’Errico, M.; de Rinaldis, E.; Blasi, M.F.; Viti, V.; Falchetti, M.; Calcagnile, A.; Sera, F.; Saieva, C.; Ottini, L.; Palli, D.; et al. Genome-wide expression profile of sporadic gastric cancers with microsatellite instability. *Eur. J. Cancer* **2009**, *45*, 461–469. [[CrossRef](#)] [[PubMed](#)]
- Tsuchiya, M.; Parker, J.S.; Kono, H.; Matsuda, M.; Fujii, H.; Rusyn, I. Gene expression in nontumoral liver tissue and recurrence-free survival in hepatitis C virus-positive hepatocellular carcinoma. *Mol. Cancer* **2010**, *9*, 74. [[CrossRef](#)] [[PubMed](#)]

17. Jones, J.; Otu, H.; Spentzos, D.; Kolia, S.; Inan, M.; Beecken, W.D.; Fellbaum, C.; Gu, X.; Joseph, M.; Pantuck, A.J.; et al. Gene signatures of progression and metastasis in renal cell cancer. *Clin. Cancer Res.* **2005**, *11*, 5730–5739. [[CrossRef](#)]
18. Tomás, G.; Tarabichi, M.; Gacquer, D.; Hebrant, A.; Dom, G.; Dumont, J.E.; Keutgen, X.; Fahey, T.; Maenhaut, C.; Detours, V. A general method to derive robust organ-specific gene expression-based differentiation indices: Application to thyroid cancer diagnostic. *Oncogene* **2012**, *31*, 4490–4498. [[CrossRef](#)]
19. Miller, L.D.; Smeds, J.; George, J.; Vega, V.B.; Vergara, L.; Ploner, A.; Pawitan, Y.; Hall, P.; Klaar, S.; Liu, E.T.; et al. An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 13550–13555. [[CrossRef](#)]
20. Kuehn, H.; Liberzon, A.; Reich, M.; Mesirov, J.P. Using GenePattern for gene expression analysis. *Curr. Protoc. Bioinform.* **2008**, *22*, 7–12. [[CrossRef](#)]
21. Mohamed, A.; Hancock, T.; Nguyen, C.H.; Mamitsuka, H. NetPathMiner: R/Bioconductor package for network path mining through gene expression. *Bioinformatics* **2014**, *30*, 3139–3141. [[CrossRef](#)] [[PubMed](#)]
22. Luna, A.; Babur, O.; Aksoy, A.B.; Demir, E.; Sander, C. PaxtoolsR: Pathway Analysis in R Using Pathway Commons. *Bioinformatics* **2015**, *32*, 1262–1264. [[CrossRef](#)] [[PubMed](#)]
23. Usoskin, D.; Furlan, A.; Islam, S.; Abdo, H.; Lönnerberg, P.; Lou, D.; Hjerling-Leffler, J.; Haeggström, J.; Kharchenko, O.; Kharchenko, P.V.; et al. Unbiased classification of sensory neuron types by large-scale singlecell RNA sequencing. *Nat. Neurosci.* **2015**, *18*, 14. [[CrossRef](#)] [[PubMed](#)]
24. Singh, D.; Singh, B. Investigating the impact of data normalization on classification performance. *Appl. Soft Comput.* **2020**, *97*, 105524. [[CrossRef](#)]
25. Kim, T.K. T test as a parametric statistic. *Korean J. Anesthesiol.* **2015**, *68*, 540–546. [[CrossRef](#)] [[PubMed](#)]
26. Brown, J.D. Point-biserial correlation coefficients. *Statistics* **2001**, *5*, 12–16.
27. Chen, Z.; Dehmer, M.; Emmert-Streib, F.; Shi, Y. Entropy of Weighted Graphs with Randi c Weights. *Entropy* **2015**, *17*, 3710–3723. [[CrossRef](#)]
28. Kuhn, M. Caret package. *J. Stat. Softw.* **2008**, *28*, 1–26.
29. Seah, C.S.; Kasim, S.; Fudzee, M.F.M.; Ping, J.M.L.T.; Mohamad, M.S.; Saedudin, R.R.; Ismail, M.A. An enhanced topologically significant directed random walk in cancer classification using gene expression datasets. *Saudi J. Biol. Sci.* **2017**, *24*, 1828–1841. [[CrossRef](#)]
30. Kim, S.Y.; Jeong, H.H.; Kim, J.; Moon, J.H.; Sohn, K.A. Robust pathway-based multi-omics data integration using directed random walks for survival prediction in multiple cancer studies. *Biol. Direct* **2019**, *14*, 8. [[CrossRef](#)]
31. Liu, K.-Q.; Liu, Z.-P.; Hao, J.-K.; Chen, L.; Zhao, X.-M. Identifying dysregulated pathways in cancers from pathway interaction networks. *BMC Bioinf.* **2012**, *13*, 126.
32. Vara, J.Á.F.; Casado, E.; de Castro, J.; Cejas, P.; Belda-Iniesta, C.; González-Barón, M. PI3K/Akt signalling pathway and cancer. *Cancer Treat. Rev.* **2004**, *30*, 193–204.
33. Noorolyai, S.; Shajari, N.; Baghbani, E.; Sadreddini, S.; Baradaran, B. The relation between PI3K/AKT signalling pathway and cancer. *Gene* **2019**, *698*, 120–128. [[PubMed](#)]
34. Cheng, Y.W.; Chiou, H.L.; Sheu, G.T.; Hsieh, L.L.; Chen, J.T.; Chen, C.Y.; Su, J.M.; Lee, H. The association of human papillomavirus 16/18 infection with lung cancer among nonsmoking Taiwanese women. *Cancer Res.* **2001**, *61*, 2799–2803. [[PubMed](#)]
35. Zeng, Z.M.; Luo, F.F.; Zou, L.X.; He, R.Q.; Pan, D.H.; Chen, X.; Xie, T.T.; Li, Y.Q.; Peng, Z.G.; Chen, G. Human papillomavirus as a potential risk factor for gastric cancer: A meta-analysis of 1917 cases. *OncoTargets Ther.* **2016**, *9*, 7105. [[CrossRef](#)] [[PubMed](#)]
36. Scinicariello, F.; Sato, T.; Lee, C.S.; Hsu, H.C.; Chan, T.S.; Tyring, S.K. Detection of human papillomavirus in primary hepatocellular carcinoma. *Anticancer. Res.* **1992**, *12*, 763–766.
37. Cui, C.; Merritt, R.; Fu, L.; Pan, Z. Targeting calcium signaling in cancer therapy. *Acta Pharm. Sin. B* **2017**, *7*, 3–17.
38. Kohn, A.D.; Moon, R.T. Wnt and calcium signaling: β -catenin-independent pathways. *Cell Calcium* **2005**, *38*, 439–446.
39. So, C.L.; Saunus, J.M.; Roberts-Thomson, S.J.; Monteith, G.R. Calcium signalling and breast cancer. In *Seminars in Cell & Developmental Biology*; Academic Press: Cambridge, MA, USA, 2019; Volume 94, pp. 74–83.
40. Cho, J.H.; Gelinis, R.; Wang, K.; Etheridge, A.; Piper, M.G.; Batte, K.; Dakhlallah, D.; Price, J.; Bornman, D.; Zhang, S.; et al. Systems biology of interstitial lung diseases: Integration of mRNA and microRNA expression changes. *BMC Med. Genom.* **2011**, *4*, 8.
41. Zhang, H.; Teng, X.; Liu, Z.; Zhang, L.; Liu, Z. Gene expression profile analyze the molecular mechanism of CXCR7 regulating papillary thyroid carcinoma growth and metastasis. *J. Exp. Clin. Cancer Res.* **2015**, *34*, 16.
42. Meagher, E.; Rader, D.J. Antioxidant therapy and atherosclerosis: Animal and human studies. *Trends Cardiovasc. Med.* **2001**, *11*, 162–165. [[PubMed](#)]
43. Vinitha, R.; Thangaraju, M.; Sachdanandam, P. Effect of tamoxifen on lipids and lipid metabolising marker enzymes in experimental atherosclerosis in Wistar rats. *Mol. Cell. Biochem.* **1997**, *168*, 1–7. [[CrossRef](#)]
44. Ran, J.; Li, Y.; Liu, L.; Zhu, Y.; Ni, Y.; Huang, H.; Liu, Z.; Miao, Z.; Zhang, L. Apelin enhances biological functions in lung cancer A549 cells by downregulating exosomal miR-15a-5p. *Carcinogenesis* **2021**, *42*, 243–253. [[PubMed](#)]
45. Chapman, F.A.; Nyimanu, D.; Maguire, J.J.; Davenport, A.P.; Newby, D.E.; Dhaun, N. The therapeutic potential of apelin in kidney disease. *Nat. Rev. Nephrol.* **2021**, *17*, 840–853. [[PubMed](#)]

46. Zhou, G.X.; Li, X.Y.; Zhang, Q.; Zhao, K.; Zhang, C.P.; Xue, C.H.; Yang, K.; Tian, Z.B. Effects of the hippo signaling pathway in human gastric cancer. *Asian Pac. J. Cancer Prev.* **2013**, *14*, 5199–5205.
47. Wei, C.; Wang, Y.; Li, X. The role of Hippo signal pathway in breast cancer metastasis. *OncoTargets Ther.* **2018**, *11*, 2185.
48. Koushyar, S.; Powell, A.G.; Vincan, E.; Phesse, T.J. Targeting Wnt signaling for the treatment of gastric cancer. *Int. J. Mol. Sci.* **2020**, *21*, 3927.
49. Howe, L.R.; Brown, A.M. Wnt signaling and breast cancer. *Cancer Biol. Ther.* **2004**, *3*, 36–41. [[CrossRef](#)]
50. Brown, Z.J.; Heinrich, B.; Steinberg, S.M.; Yu, S.J.; Greten, T.F. Safety in treatment of hepatocellular carcinoma with immune checkpoint inhibitors as compared to melanoma and non-small cell lung cancer. *J. Immunother. Cancer* **2017**, *5*, 93.
51. Katsenos, S.; Archondakis, S.; Vaias, M.; Skoulikaris, N. Thyroid gland metastasis from small cell lung cancer: An unusual site of metastatic spread. *J. Thorac. Dis.* **2013**, *5*, E21.
52. Liu, X.; Wang, J.; Sun, G. Identification of key genes and pathways in renal cell carcinoma through expression profiling data. *Kidney Blood Press. Res.* **2015**, *40*, 288–297. [[PubMed](#)]
53. Tang, J.; Kong, D.; Cui, Q.; Wang, K.; Zhang, D.; Yuan, Q.; Liao, X.; Gong, Y.; Wu, G. Bioinformatic analysis and identification of potential prognostic microRNAs and mRNAs in thyroid cancer. *PeerJ* **2018**, *6*, e4674. [[PubMed](#)]
54. Ito, R.; Oue, N.; Zhu, X.; Yoshida, K.; Nakayama, H.; Yokozaki, H.; Yasui, W. Expression of integrin-linked kinase is closely correlated with invasion and metastasis of gastric carcinoma. *Virchows Arch.* **2003**, *442*, 118–123. [[PubMed](#)]
55. Engelman MD, F.B.; Grande, R.M.; Naves, M.A.; de Franco, M.F.; de Paulo Castro Teixeira, V. Integrin-linked kinase (ILK) expression correlates with tumor severity in clear cell renal carcinoma. *Pathol. Oncol. Res.* **2013**, *19*, 27–33.
56. Yan, C.; Yang, Y.; Tang, Y.; Zheng, X.; Xu, B. The Critical Gene Screening to Prevent Chromophobe Cell Renal Carcinoma Metastasis through TCGA and WGCNA. *J. Oncol.* **2022**, *2022*, 2909095.
57. Shen, K.; Johnson, D.W.; Gobe, G.C. The role of cGMP and its signaling pathways in kidney disease. *Am. J. Physiol.-Ren. Physiol.* **2016**, *311*, F671–F681.
58. Lv, Y.; Wang, X.; Li, X.; Xu, G.; Bai, Y.; Wu, J.; Piao, Y.; Shi, Y.; Xiang, R.; Wang, L. Nucleotide de novo synthesis increases breast cancer stemness and metastasis via cGMP-PKG-MAPK signaling pathway. *PLoS Biol.* **2020**, *18*, e3000872.
59. Chen, J.; Yuan, B. Detecting functional modules in the yeast protein-protein interaction network. *Bioinformatics* **2006**, *22*, 2283–2290.
60. Ideker, T.; Ozier, O.; Schwikowski, B.; Siegel, A.F. Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics* **2002**, *18* (Suppl. 1), S233–S240.
61. Segal, E.; Wang, H.; Koller, D. Discovering molecular pathways from protein interaction and gene expression data. *Bioinformatics* **2003**, *19* (Suppl. 1), i264–i271. [[CrossRef](#)]
62. Sharan, R.; Suthram, S.; Kelley, R.M.; Kuhn, T.; McCuine, S.; Uetz, P.; Sittler, T.; Karp, R.M.; Ideker, T. Conserved patterns of protein interaction in multiple species. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 1974–1979. [[PubMed](#)]
63. Tay, X.H.; Sutikno, T.; Kasim, S.; Fudzee, M.F.; Hassan, R.; Seah, C.S. A Direct Proof of Entropy-Based Directed Random Walk. 2022. Available online: <https://crim.utem.edu.my/wp-content/uploads/2022/09/204-414-4151.pdf> (accessed on 1 February 2023).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.