GediNET- Discover Disease-Disease Gene Associations utilizing Knowledge-based Machine Learning

Malik Yousef (malik.yousef@gmail.com)
Zefat Academic College
Emma Qumsiyeh
Al-Quds University

Article

Keywords:

Posted Date: May 31st, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1643219/v1

Abstract

Most gene expression studies aim to discover genes associated with specific diseases. The standard approaches based on machine learning utilize several feature selection techniques to identify significant genes that can serve as biomarkers for a given disease. In recent times, the integration of prior knowledge-based approaches for biomarker identification has shown much promise in discovering several biomarkers, thus allowing scope for an increase in translational applications. In this study, we developed a novel approach GediNET that integrates prior biological knowledge about genes associated with diseases like cancer to group genes into groups. The novelty of GediNet is that it discovers disease-disease gene associations within groups rather than disease-gene associations.

These groups are later subject to a Scoring component for performing group selections rather than single feature selection. The top-ranked groups are used to train the machine learning model. The process of Grouping and Scoring using the (G-S-M) model is then applied to discover groups of disease genes or biomarkers for a specific disease. One of the outputs of the suggested tool GediNET is a list of significant groups of diseases that combine their associated genes can contribute to developing biomarkers and drugs.

GediNET identifies the relationships between diseases, Disease–disease association (DDA) based machine learning, which explores novel associations of diseases that enhance knowledge of disease relationships, which could further improve approaches to disease diagnosis, prognosis, and treatment.

The GediNET Knime workflow can be downloaded from: https://github.com/malikyousef/GediNET.git or

https://kni.me/w/3kH1SQV_mMUsMTS- .

Introduction

Complex diseases like diabetes, Alzheimer's, and cancer are influenced by genetics, lifestyle, and environmental factors. They do not follow any inheritance patterns. A great effort in research is to target such diseases to reveal their genetic disorders and corresponding disease genes. Understanding disease genetic causes can lead to early diagnosis, prognosis, and an effective drug design [1]. With the advances in bioinformatics, researchers have made a tremendous effort to identify disease-related genes effectively. Biomarker identification and sample classification, based on gene expression data, have become an attractive research area in the field of bioinformatics [2–5].

The increased availability of high-throughput molecular profiling data with reduced costs has triggered researchers to deeply analyze the emerging biological knowledge. Over the last decade, the large availability of datasets has contributed to forming a rich resource cohort. Many resources of biological knowledge and repositories are available, such as miRTarBase [6] for microRNA, Gene Ontology (GO) [7], Gene Expression Omnibus (GEO), which provides access to microarray measurements [8], TCGA - a database for gene expression RNA-seq [9], and KEGG - a knowledge-base of pathways [10]. Another widely used biological resource is DisGeNET, a knowledge-base platform for gene-disease–variant associations [11]. Researchers can leverage these resources for in-silico validation and train statistical machine learning models for classification and biomarker discovery.

Hallmarks of human diseases follow the same rule: the critical perturbation in the gene(s)/protein(s) will have implications in molecular pathways and produce implicated or lethal phenotypes. This is based on the principle of guilt-by-association, which suggests that associated genes share functions such as genetic or physical interactions [12]. In other words, genes responsible for similar diseases are alike. This finding has motivated to shift from the traditional pure data-oriented approaches to knowledge-based integrative approaches to handle the considerable resources. Insights can be better attained when advanced tools exploit biological knowledge for deep analysis rather than the traditional clustering and machine learning approaches [13, 14].

Different studies have invested in identifying genes associated with human diseases. They also shed light on the importance of using computational tools to diagnose diseases and design novel drugs. Although too many publications on computational tools exist in the literature, they differ in their approach and use of resources. Many integrated various biological information about disease genes into machine learning[15, 16]. One sort of integrative approach is by aggregating multiple datasets to increase the statistical power in effectively identifying a small subset of genes to predict disease types [17]. One such method is BioGraph, presented by Liekens et al. [18]. The authors developed a data-mining platform for disease gene prioritization and identification. They integrated 21 curated biomedical databases to rank disease-gene relations. Results exposed disease genes and identified potential susceptibility genes.

Other approaches such as GeP-HMRF integrated Genome-wide association studies (GWAS), expression quantitative trait loci (eQTL), and protein-protein interaction (PPI) data [19]. GeP-HMRF is a unified statistical model to predict disease-related genes. Authors described that their approach outperforms Sherlock [20], COLOC[21], and NetWAS [22] tools. The work of Peng et al. [23] proposed a new network-based disease gene prediction method called SLN-SRW (Simplified Laplacian Normalization-Supervised Random Walk) to generate edge weights of a new biomedical network by integrating heterogeneous sources of biomedical data.

The study by [16] has demonstrated that machine learning classifiers trained on functional gene similarities, using Gene Ontology (GO), can improve the identification of genes involved in complex diseases. The GO annotations were used to compute similarities between genes. The approach was tested on autism spectrum disorder (ASD) candidate genes. Luo et al. [24] proposed EdgCSN, an ensemble learning algorithm that uses protein-protein interaction networks extracted from clinical sample-based networks to predict disease-associated genes.

Many studies that integrate biological knowledge about genes associated with diseases have been examined in the literature. However, a critical component of such research is integrating a profound knowledge base for genes and associated diseases. Such knowledge exists in the DisGeNET database [25]. For example, Hamzeh and Rueda propose a new machine learning method that incorporates the DisGeNET database to detect biomarkers in prostate cancer. A wrapper-based feature-selection approach was used to group genes-related diseases based on their classification accuracy. Results for each iteration were saved for further validation by researchers based on the best AUC or the highest number of detected genes in each group [25].

Yousef et. al. has developed the Grouping-Scoring-Modeling (G-S-M) approach for integrated biological knowledge through different computation tools such as SVM-RCE-R [27, 28] maTE [29], CogNet [30],mirCorrnet [31],miRModuleNet [32], and PriPath [33]. For a review paper on G-S-M approaches, we refer to [34].

SVM-RCE-R [27, 28] tools were the first study by Yousef et al. that considered groups of genes rather than individuals. SVM-RCE (Support Vector Machines -Recursive Cluster Elimination) groups genes based on their gene expression values and scores each cluster of genes by a machine learning algorithm. Moreover, a recent study by Yousef et al. [34] used the G-S-M model to integrate Gene Ontology for grouping the genes. Similarly, SVM-RNE [35] detects gene networks to serve as clusters for ranking and scoring by adopting the G-S-M model. Even though different studies have used mRNA expression data and knowledge bases such as DisGeNet, our pioneer approach is not similar to any of the tools presented before. Using the G-S-M approach, the main objective is to group genes best related to a specific disease. Our novel machine learning approach with two-class classification does not need other data annotation. With Monte Carlo cross-validation (MCCV), fractions of the samples are randomly selected as training data, and the rest is assigned for the test data. The most accurate disease-genes groups are then identified in each iteration, and later accumulative top-ranked groups are combined to train the model. We further examined the results with similar approaches that follow the same merit, such as maTE [29], CogNet [30], mirCornet[31], miRModuleNet[32], and PriPath[33] ; GediNET has shown its superiority against previous state-of-the-art methods. However, the aim of the GediNET is not to compete with other tools in terms of performance; the aim is to discover a novel Disease-Disease association-based machine learning.

Materials And Methods

Datasets

We downloaded 10 human gene expression datasets for different types of complex diseases from GEO [8]. For each dataset, the name of the disease and the number of samples were defined. Moreover, positive and negative samples were available. Table 1 describes the 10 datasets in more detail.

Table 1. Description of the 10 datasets used in the study. Each entry has the GEO accession, the name of the disease, the number of samples, and the data classes.

GEO accession	Title	Disease	#Samples	Classes
GDS1962	Glioma-derived stem cell factor effect on angiogenesis in the brain	Glioma	180	negative = 23 positive = 157
GDS2545	Metastatic prostate cancer (HG-U95A)	Prostate cancer	171	negative = 81 positive = 90
GDS2771	Large airway epithelial cells from cigarette smokers with suspect lung cancer	Lung cancer	192	negative = 90 positive = 102
GDS3257	Cigarette smoking effect on lung adenocarcinoma	Lung adenocarcinoma	107	negative = 49 positive = 58
GDS4206	Pediatric acute leukemia patients with early relapse: white blood cells	Leukemia	197	negative = 157 positive = 40
GDS5499	Pulmonary hypertension: PBMCs	Pulmonary hypertension	140	negative = 41 positive = 99
GDS3837	Non-small cell lung carcinoma in female nonsmokers	Lung cancer	120	negative = 60 positive = 60
GDS4516_4718	Colorectal cancer: laser microdissected tumor tissues	Colorectal cancer	148	negative = 44 positive = 104
GDS2547	Metastatic prostate cancer (HG-U95C)	Prostate cancer	164	negative = 75 positive = 89
GDS3268	Colon epithelial biopsies of ulcerative colitis patients	Colitis	202	negative = 73 positive = 129

DisGeNET disease-gene association dataset

The dataset containing genes and associated diseases was downloaded from DisGeNET version 7.0 [25]. The dataset contains 30,170 diseases and 21,666 genes that form 3,241,576 gene-disease connections. Given the massive dataset

size, two filters were set to reduce the number of associations in terms of practicality and to reduce the computational complexity. The filters were set on the columns *diseaseType* and *diseaseSemanticType* in the DisGeNET dataset. The *diseaseType* column divided the data into three categories - disease, phenotype, and group - and we only chose disease concerning our study. On the column *diseaseSemanticType*, we only chose those rows categorized as *Neoplastic Process* and *Disease*. This was done to increase compatibility and better understand the workflow results. After filtering, only 15991 genes and 3929 diseases related to diseases remained for further analysis, which accounted for 329936 gene-disease associations. Figure 1 illustrates a part of the disease distribution over the number of genes for each disease.

The merit of our GediNET as Disease-Disease Associations

Let's assume that given a gene expression dataset D, which was designed to study a specific disease R (for example, Lung Cancer or Breast cancer) to detect the significant genes or the biomarker of this disease-based gene expression. The traditional approach of the classification model suggests a list of *k* genes that can serve as a biomarker for predicting the patients with the disease R. In other words, Identifying disease-gene associations. One solution could be a linear function F(X) as :

$F(X) = w_1g_1+w_2g_2+...+w_kg_k$ where w_i are the weights (scores) while the g_i are the gene expression values. The weights could serve as the importance of each gene expression in this equation. For instance, a value weight close to zero indicates that the associated genes contribute less to the equation model. In other words, F(X) describes the biological interaction between those *k* genes to form biomarkers.

Our new approach GediNet is different from those traditional approaches by suggesting the following model equation that Identifies disease-disease gene associations:

$F^*(X) = w_1*grp\_disease_1 + w_2*grp\_disease_2+...+w_p*grp\_disease_p$, where the model consists of *p* groups that were highly scored by the component S of GediNET(See section The main workflow of GediNET). The group $grp\_disease_i$ ($i=1,2,..p$) is a set of genes associated with one disease. F*(x) represents the model by a linear function of groups, also it could be represented as a decision tree, as illustrated in Figure 2 (Right panel). The left panel of Figure 2 illustrates the decision tree model of the significant genes selected by the traditional approach. One needs to take this list and proceed with other functional enrichment processes to discover more biological relationships. On the other hand, the right panel of Figure 2 shows that the decision tree model consists of genes that are associated with the top three GeDiNET ranked diseases. This model contains information about biological knowledge of the diseases showing the disease-disease associations.

For example, considering the datasets GDS1962 that studies the Glioma disease, GediNet might suggest a model of top three significant groups/diseases:

Grp_disease_1 =PAPILLARY RENAL CELL CARCINOMA,

Grp_disease_2= PLASMA CELL,

Grp_disease_3 = NEOPLASM, and ADULT GLIOBLASTOMA.

Where is the following are the sets of genes associated with each disease:

Grp_disease_1 = {SLC16A1, TAGLN2, TIMP3, IGFBP7, TOP2A, TP53, RRM2,..},

Grp_disease_2 = {CD99, TP53, LPL, CD40, CD38, NCAM1, MYC, CSF3, CDKN2A, FGFR3, CCND1},

Grp_disease_3= {EDNRA, CSPG4, MELK, ENPEP, ...}.

Applying GediNet will compute F*(x) that describes the association between the genes, associated with different diseases to the current disease under study and also will describe the relationship between the top detect significant diseases(groups).This might lead to new discoveries that have not been observed before by the traditional approaches. Additionally, the model that GediNET discovers could be combined with the top-ranked group of gene diseases (See the M Component in section number 4), which might be used to explore the relationship between those diseases. One can use those genes in enrichment analysis to discover the different pathways and their roles in each disease and the current disease.

The main workflow of GediNET

The main inspiration for developing our novel tool GeDiNET is the generic approach named G-S-M, which was adopted by different tools such as SVM-RCE [36], SVM-RCE-R [25],SVM-RCE-R-OPT[37], SVM-RNE[35], maTE [27], CogNet [28], miRcorrNet [31], Integrating Gene Ontology Based Grouping and Ranking [34], miRModuleNet [32], PriPath[33] and recently reviewed in Yousef et al. [38]. The main workflow of GeDiNET is illustrated in Figure 3, where the G-S-M approach is presented in the three main sections labeled with the orange section (G), the yellow section (S), and the green section (M), which represent:

1. The G Component (Grouping): where the genes of each disease are grouped.
2. The S Component (Scoring): where the groups are scored and ranked.
3. The M Component (Machine Learning model): where the model is created by training a classifier (Random Forest).

The input for the workflow is gene expression data. The data consist of two classes of samples, classes are control (negative) and disease (positive). The data is split into training and testing. The training data is used to build the final model, while the testing data is used to evaluate the model's performance. The whole workflow is repeated 100 iterations using the cross-validation loop, where the input is randomly split into 90% training and 10% testing in each iteration. A one-way ANOVA test is performed on the training set to filter out the top genes. The top 2000 differentially expressed genes with a P-value less than 0.05 are selected. The selected genes are then used to filter the test dataset to contain the same genes.

The main contribution of the generic approach and the description of each component's functions are explained in detail in the following sections.

Grouping Genes based on Disease (The G component)

The first main component in our tool is the grouping component G (the orange section in Figure 3), which groups genes into groups. The G component can be any algorithm that groups genes. For example, Yousef et al. previously used the maTE algorithm to group the gene expression by the miRNAs that can target them according to the miRTarBase database [29]. In this tool, the G components group genes based on their disease associations extracted from the biological knowledge of the DisGeNET v7 database [39]. The main idea is to group the genes into disease groups. Each group is one of the known diseases where its group members are the genes associated with this disease. Table 2 is an example of such groups where it includes the disease name (group name), the set of genes associated with this disease and the last column is the number of genes that represent the size of the group.

Table 2. An example of groups of diseases with their associated genes. The last column represents the number of genes in each group (group size)

Group Name	Genes	#Genes
Small Cell Carcinoma Of Lung	VPS13B, SLC16A1, ANXA1, CD99, SMARCC1, PCNA…	41
Leukemia, B-Cell	TP53, LAMA4, STK11, CSPG4, CD40, TNFRSF1A…	43
Stage Iii Breast Cancer Ajcc V6	TP53, BRCA2	2
Head And Neck Carcinoma	PRMT5, ANXA1, LGALS1, TIMP3, IGFBP7, PCNA, TNC, TP53…	149
Secondary Malignant Neoplasm Of Bone	ADAM9, SLC16A1, CD99, NME1-NME2, DPYSL3, TNC, TP53, NRAS…	145
Malignant Glioma	TK1, NPAS3, CD63, HMGB1, TAGLN2, TXNIP…	162
Adenocarcinoma, Tubular	PCNA, TP53, EFEMP1, APOE, STK11, PRKD1…	31
Childhood Brain Neoplasm	TP53, NRAS, SOX9, MYC, TNFRSF11B	5
Adult Myelodysplastic Syndrome	CSNK1A1, CTNNA1, HMGB1, PCNA, TOP2A, TP53…	58
Non-Small Cell Lung Cancer Stage I	TP53, PRRX1, IGFBP3, VEGFA, S100A6, GSTK1…	22

Creating Sub-data

Further, a sub-data set needs to be generated for each disease group. This is achieved by extracting the genes belonging to the specific disease and their original class label from the original gene expression training part of the data. Let *f=1,..m* be the number of groups generated by the G component. This stage we will extract or create *m* sub_data named grp_disease_genes_subdata$_1$, grp_disease_genes_subdata$_2$, …, grp_disease_genes_subdata$_m$, that will be serving for the S (Scoring) component. Figure 4 is an example of creating sub_data for four different diseases (groups). For example, the *Well Differentiated Pancreatic Endocrine Tumor* disease group is a group with five genes associated with this specific disease. The genes are RBMS3, TFE3, SSTR2, NTRK1, and PAX8. Moreover, the *X-Linked Lymphoproliferative Disorder* disease is another group with only two genes which are SERPINA4 and NR0B2. The sample class is also extracted and specified for each sample, where *pos* is for the positive class and *neg* for the negative class. Each disease group with its sub-data is the input to the following S component (yellow section in Figure 3) to be scored and sorted.

Scoring the groups of diseases associated with their genes

Consider the gene expression dataset as D, which contains two classes of *s* covariate samples (patients and control) and *n* genes. After applying the grouping component G for each disease, the diseases that are now represented by a sub_data, are scored according to their ability to best differentiate between the two classes after training on the associated sub_data using a Random Forest (RF) classifier. The sub_data is divided into a conventional 80:20 training and testing split. We repeat this procedure r=5 times recording different performance metrics while we use the mean of the accuracy as the assigned score for the specific disease. However, one might use a different combination of those metrics to assign the final score. For more information on such an option, we refer to [37].

Table 3 is an example output of the Scoring component for the GDS2525 dataset.

Table 3: An example of the output of the Scoring S component. The first column is the name of the disease name, the Accuracy column is the score given by the S component, and the Rank is the Rank of the group based on the value of the score.

Disease	Score as Accuracy	Rank
PAPILLARY RENAL CELL CARCINOMA	0.98	1
PLASMA CELL NEOPLASM	0.98	1
ADULT GLIOBLASTOMA	0.97	2
INTESTINAL CANCER	0.97	2
MALIGNANT NEOPLASM OF COLON STAGE IV	0.97	2
DERMATOFIBROSARCOMA	0.95	3

Implementation of GeDiNET

We have implemented the GeDiNET tool using the free and open-source platform KNIME [41] due to its simple and intuitive graphical user interface. KNIME is a highly integrative platform that has enabled the scope to utilize scripts in both python and R in tandem to implement our tool as a KNIME workflow.

The workflow created on KNIME comprises several nodes with their separate functions. Meta-nodes are created as a collection of nodes that perform specific tasks.

The KNIME workflow for GeDiNET is presented in Figure 5. It starts by uploading a list of the names of the dataset via the "List Files/Folders" node. Then a loop over those datasets is run to read each dataset by the node "Table Reader," which is then processed by the meta-node "FilterMissingValues" to remove and or filter out rows with missing values. It then sends the filtered data as input to the GeDiNET meta-node. While the "Integer Input" node allows modifying the number of iterations, the tool should be used while training the model.

The flowchart for the GeDiNET tool is presented in Figure 3, while in Figure 5, the implementation of the GeDiNET as a KNIME workflow along with its meta-nodes and components are shown. The left outport (input ports) brings 3 inputs into the GeDiNET node (Top: Variable Flow port, which contains path/location of datasets and output files, Middle: Input dataset with missing values removed, and Bottom: Number of iterations). The input dataset is passed to the "GroupTargetsByDisease" node, which acts as the biological grouping function by grouping all genes concerning their corresponding diseases. The dataset is then normalized and passed onto the "ClassificationBasedDiseaseRanks" component for ranking via a "Partitioning" node that segments the input dataset into training and a testing set.

The "ClassificationBasedDiseaseRanks" node is expanded in Figure 6, which shows the two further meta-nodes, "Genes filter ttest" which performs a one-way ANOVA test to filter out top genes which are further used in the "Rank and Classify."

Finally, the "SaveResults" node collects all the results after each iteration of the "Loop End" node to process and save the results.

The GediNET Knime workflow could be downloaded from: https://github.com/malikyousef/GediNET.git or https://kni.me/w/3kH1SQV_mMUsMTS-

Model Performance Evaluation

We used the Random Forest Classifier while splitting the data into 80% training and 20% testing. Since the datasets are imbalanced, meaning the dataset's target class has an uneven distribution of observations, we employed the under-sampling method. Such a method deals with the imbalanced datasets by pertaining all of the data in the minority class while decreasing the size of the majority class. Besides, for model training, we applied 10-fold Monte Carlo cross-validation (MCCV) [42]. With Monte Carlo cross-validation (MCCV), fractions of the samples are randomly selected as

training data, and the rest is assigned for the test data. The performance measures are computed as the average of 100-fold MCCV.

To evaluate the performance of RF model, several quantitative metrics were calculated, such as Accuracy, Sensitivity, Specificity, and Precision [43], using the following formulations:

Sensitivity (SE, Recall) = TP/(TP + FN)

Specificity (SP) = TN/(TN + FP)

Accuracy (ACC) = (TP + TN)/(TP + TN + FP + FN

Where TP: true positive; FP: false positive, TN: true negative; and FN: false negative. Moreover, the Area Under the Curve (AUC) measures the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve [44]. We used the AUC to evaluate the performance results.

In each iteration, our approach generates lists of diseases groups and their associated genes that are slightly different. Hence, there is a need to apply a prioritization approach on those lists. As utilized in miRcorrNet, we have used rank aggregation methods. In this respect, we have embedded the RobustRankAggreg R package[40], developed by (Kolde et al., 2012) into GediNET workflow. The RobustRankAggreg assigns a p-Value to each element in the aggregated list, which describes how good each element/entity was ranked compared to the expected value.

Results

Performance Evaluation of GediNET

Table 4 presents an example of the average 100-fold MCCV performance table of GeDiNET for aggregated top-ranked 10 groups for the GDS1962 dataset. The last row presents the performance of the top-ranked group (#Groups=1). The AUC obtained is 97% using 21.61 genes on average. The row of #Groups=2 presents the performance metrics obtained for the top 2 groups, where the genes of the first top-ranked group and the second-highest scoring group are aggregated together. That is to say that GeDiNET reports the performance results for the top 10 groups cumulatively.

Table 4. An example average of 100 MCCV performance table of GeDiNET for top-ranked 10 groups for GDS1962 dataset cumulatively.

#Groups	#Genes	Accuracy	Sensitivity	Specificity	AUC
10	136.74	0.928	0.93	0.92	0.98
9	127.68	0.93	0.93	0.92	0.98
8	116.02	0.93	0.94	0.92	0.98
7	111.16	0.93	0.93	0.91	0.98
6	102.02	0.93	0.9	0.92	0.98
5	92.88	0.93	0.93	0.93	0.98
4	78.37	0.93	0.93	0.92	0.98
3	62.47	0.93	0.94	0.92	0.98
2	45.57	0.93	0.93	0.93	0.97
1	21.61	0.92	0.93	0.92	0.97

Table 5. Performance Results of GeDiNET over the top-ranked group. ACC stands for Accuracy, SEN stands for Sensitivity, SPE stands for Specificity, FM stands for F-Measure, and AUC stands for Area Under the ROC Curve.

GEO Accession	#Genes	ACC	SEN	SPE	AUC
GDS1962	45.57	0.93	0.93	0.93	0.97
GDS2545	113.76	0.73	0.72	0.74	0.81
GDS2771	97.83	0.64	0.69	0.59	0.70
GDS3257	74.81	0.97	0.99	0.94	0.99
GDS3837	21	0.92	0.83	1	0.92
GDS4206	83	0.66	0.3	0.82	0.58
GDS4516_4718	40.72	0.99	0.99	0.99	1
GDS2574	102.49	0.76	0.77	0.76	0.83
GDS3268	115.7	0.67	0.7	0.63	0.73
GDS5499	80.23	0.9	0.96	0.77	0.95

Table 5 shows the GediNET performance over 10 datasets for the top 2 groups. All values are the results of an average of 100-MCCV while considering the AUC for presenting the performance. The complete performance results are attached in the supplementary. The table shows the GEO accession in the first column, the number of genes in column #Genes while ACC is the accuracy, SEN is the sensitivity, SPE is the specificity, and the AUC is the area under the curve. We see only one unsuccessful result for the dataset GDS4206. However, a similar observation was made when applying other tools to this specific dataset, as illustrated in Figure 7.

The average number of genes associated with the top 2 groups is slightly high because the distribution of genes over the disease is slightly high compared, for example, to other biological knowledge such as microRNA target or KEGG

pathways. Moreover, this number of genes could be reduced by removing the least contributed genes by processing each group. This step will be considered in the future version of the algorithm. Also, one can use additional biological knowledge to filter out more genes from the group by, for example, leaving the most associated genes with the disease. The last suggestion requires other biological resources to be embedded into the GeDiNET.

Comparative Evaluation with other biological G-S-M

For comparison, we have considered similar tools that apply the G-S-M approach by integrating biological knowledge for grouping the genes and performing the scoring on the group, such as CogNet, maTE, and PriPath. We have recorded the AUC values for the top 1-10 groups ranked by the scoring component for each tool by applying 100-MCCV. More specifically, we considered the top two groups for comparison purposes.

Figure 7 illustrates the mean AUC values of the four tools for the 10 datasets. Meanwhile, Figure 8 plots the mean number of genes for the four tools. As apparent in Figure 7, the AUC values of GeDiNET, CogNet, maTE, and PriPath for 10 different datasets for the top two clusters are nearly similar. Thus the performance of those tools is comparable. This close performance indicates that the developed tool GediNET is consistent and robust. However, the outcome of each tool is different as each one of those tools has its merit and its aim of detecting significant groups related to specific pre-biological knowledge. Interestingly it is to develop a tool that integrates the outcome of all those tools to shed light on a new discovery.

Figure 8 implies that, on average, GediNET uses a 10-fold higher number of genes than other tools. This is due to the fact that the groups of genes associated with the diseases are much higher than others.

One of the tool's outputs is a list of ranked disease groups that were assigned a p-value by the robust rank aggregation package [40]. Table 6 is an example of this tool for the GDS1962 dataset.

Table 6. An output of the RobustRankAggreg tool for the GDS1962

GDS1962			
Disease Name	**p-value**	**#Genes**	**List of genes**
PAPILLARY RENAL CELL CARCINOMA	0.00052	22	SLC16A1, TAGLN2, TIMP3, IGFBP7,...
PLASMA CELL NEOPLASM	0.0010	11	CD99, TP53, LPL, CD40,...
COMMON ACUTE LYMPHOBLASTIC LEUKEMIA	0.001772	3	KNG1, MME, BCL2
DUCTAL BREAST CARCINOMA	0.002363	13	TCF21, AFAP1L2, PLG,...
GASTRIC MUCOSA-ASSOCIATED LYMPHOID TISSUE LYMPHOMA	0.002953	2	BCL2, EPCAM
INTRAHEPATIC CHOLANGIOCARCINOMA	0.003544	27	SHBG, BAX, TYMS, GPC3,...
LYMPHOMA, NON-HODGKIN	0.004135	44	BAX, SLC23A1, MME, TYMS, …
MALIGNANT NEOPLASM OF COLON STAGE IV	0.004725	7	TYMS, MYCN, KLK6, NDRG1, …
NEUROECTODERMAL TUMOR, PRIMITIVE	0.005316	14	SFRP1, PCSK2, MYCN, CAPS,...
PAPILLARY THYROID CARCINOMA	0.005907	75	BAX, PKHD1L1, MME, GPC3, …

This is a novel output of feature selection techniques that our tools provide. This table will be used to analyze the relationship between the diseases further. For example, Table 6 raises a biological question about the association

between the top-ranked disease (PAPILLARY RENAL CELL CARCINOMA, PLASMA CELL NEOPLASM, ..) and the target disease of the study (data set GDS1962 with target disease Glioma). Additionally, GediNET provides a list of significant genes that also was aggregated by the RobyutRankAgreg tool. While scoring each group, the gene associated with the group is scored with the same score as the group. This list with its scores is aggregated at the end to compile and report a list of significant genes as described in Table 7.

Table 7. Top 10 significant genes that were aggregated by the RobustRankAggreg tool for the GDS1962 dataset.

Gene	p-value
GALNT13	0.0449
C1R	0.1448
NUP35	0.1482
KDELR2	0.1504
MCUB	0.1664
PHYHIPL	0.1673
GNAI3	0.1758
OXCT1	0.1774
ANXA2P2	0.1821
TUBB6	0.1824

The user can consider the list of significant genes for functional and enrichment analysis as was done in similar studies such as PriPath and miRmodulnet using different tools such as David [45], EnrichR [46], and GeneMANIA [47].

Biological Interpretations

One of the outputs of GediNet is a list of significant diseases which had been scored by the S component, as illustrated in Table 6. This list is ranked by p-value (ranked by RobustRankAggreg).

For all the 10 GEO datasets, the top 2 diseases and their set of genes were considered to perform pathway enrichment analysis. Their total number of distinct genes is 1184.

The web tool, EnrichR [46] was used to perform the pathway enrichment. The tool was run to collect the top enriched pathways for each disease-gene group per dataset, and the top pathways (with the least p-value) were selected. WikiPathway database [48] version 2021 for human genes was used to select our results. The top cell signaling pathways' names for the 10 GEO datasets, p-values, adjusted p-value, and associated genes are illustrated in Table 8. Evidence from literature was then gathered for the dataset cancer and the top-performing disease, along with the enriched genes and pathways found from the enrichment analysis.

Table 8: The top cell signaling pathways' names for the 10 GEO datasets. The first column is the name of the cell signaling pathway, the second column is the p-values, the third column is the adjusted p-value, the Genes column represents an example of the associated genes, and finally, the last column is the total number of associated genes.

Cell signaling pathways term	P-value	Adjusted P-value	List of Genes	#Genes
Head and Neck Squamous Cell Carcinoma WP4674	2.24E-13	6.31E-11	CCND1;CDKN2A; AKT1…	9
DNA damage response (only ATM dependent) WP710	2.95E-16	1.08E-13	GSK3B;SMAD4;CDKN1A,...	14
VEGFA-VEGFR2 Signaling Pathway WP3888	1.66E-10	6.37E-08	LRRC59;NRP2;PRKAA2;...	27
VEGFA-VEGFR2 Signaling Pathway WP3888	1.05E-11	2.59E-09	HSP90AA1;ANXA1;...	18
Lung fibrosis WP3624	6.32E-09	1.73E-06	GREM1;CSF3;IL6;PLAU;EGF;MUC5B;MMP9	7
IL-18 signaling pathway WP4754	2.33E-17	1.05E-14	GSK3B;CEBPB;CXCL8;...	29
Effects of nitric oxide WP1995	2.93E-05	0.00310457	NOS1;XDH	2
TP53 network WP1742	2.14E-13	9.13E-11	CDKN1A;CDKN2A;MYC;...	9
Apoptosis WP254	1.88E-06	4.25E-04	CASP10;MYC;PMAIP1;...	6
Hepatitis C and Hepatocellular Carcinoma WP3646	5.41E-12	2.07E-09	CDKN1A;IL6;CXCL8;...	10

Next, we used the Cytoscape tool [49] to visualize the correlation network between the cell signaling pathways with the overlapping genes for all the top enriched pathways from the previous step. In total, we took the most 10 significant pathways that were enriched among the 20 disease-gene group pairs to visualize. Figure 11 represents the signaling pathway networks with overlapping genes across different GEO datasets.

As we have stated, we examine 10 different GEO gene expression datasets, studying mostly different diseases. Figure 11 illustrates the most significant pathways related to all given datasets, indicating that disease genes are correlated and associated even when studying different diseases. The network in Figure 11 shows that GediNet discovers important biological information related to various diseases. Moreover, We have studied the significance of GediNet on the data GDS3257 by considering the top 2 significant diseases having 12 distinct genes. Figure 12 illustrates the network of the most significant pathways and their related genes.

Disease-Disease Associations

We assume that disease is represented by a set of genes. The simple approach for finding a disease-disease association is by applying different association indices that consider the number of shared genes between the two diseases. For example, one might use the Jaccard Simpson, Geometric, Cosine, and even Pearson correlation coefficient (PCC) [34,35].

Recently, different efforts toward Disease-Disease associations (DDA) are gaining attention for their importance in exploring novel associations of diseases and enhancing knowledge of disease relationships, which could further improve approaches to disease diagnosis, prognosis, and treatment. Yet, shared genes offer only limited information about the relationship between two diseases.

The number of known DDA and reliable associations is very small. Thus it suggests that more efforts are required for DDA detections.

Disease-disease relationships through the incomplete human interactome [50] are computational approaches that derive mathematical conditions for the identifiability of disease modules and show that the network-based location of each disease module determines its pathobiological relationship to other diseases.

Suratanee A, Plaimas K. [51] have developed a novel network-based scoring algorithm called DDA to identify the relationships between diseases in a large-scale study. Their method is developed based on a random walk prioritization in a protein-protein interaction network.

DisGeNET provides through its API disease-disease associations that have been obtained by computing the number of shared genes and shared variants between pairs of diseases by source. DisGeNet uses two metrics to compute the DDA. The first one is the Jaccard Index (JI) $Jaccard_G = \frac{G_1 \cap G_2}{G_1 \cup G_2}$, G_1 is the set of genes associated with Disease 1, and G_2 is the set of genes related to Disease 2.

The second one is Jaccard variance $Jaccard_G = \frac{G_1 \cap G_2}{G_1 \cup G_2}$, V_1 is the set of variants associated with Disease 1, and V_2 is the set of variants associated with Disease 2.

In order to compute for each dataset the standard DDA in GediNET, we have computed the fraction of the number of shared genes for each pair of the top-scored disease group for 4 datasets, as illustrated in Figure 13.

The suggested tool GediNET is different from the tools mentioned above in that it is based on machine learning for detecting the relationships between diseases, Disease–disease association (DDA), which detects novel and not known associations of diseases that might enhance knowledge of disease relationships, which could further improve approaches to disease diagnosis, prognosis, and treatment. We have conducted a further analysis to explore if GediNet suggests new unknown relationships between diseases using DisGeNET API.

Table 9 illustrates for each data set its three top detected diseases by DisGeNET API and the top 3 ranked diseases by GeDiNET. For each detected disease by DisGeNet, we have looked up the disease in the list of ranked diseases by GeDiNET to examine the two tools.

Table 9. illustrates the three top detected diseases by DisGeNET API and the top 3 ranked diseases by GeDiNET for each GEO dataset. For each detected disease by DisGeNet, we have looked up the disease in the list of robust ranked aggregated disease results by GeDiNET. The values in parenthesis for the rows of DisGeNET are the position of the disease and the p-value assigned by GediNET.

GEO Data Set/ Target Disease	The Data Disease	Top 1 Disease name	Top 2 Disease name	Top 3 Disease name
GDS1962/ BrainStem Glioblastoma	DisGeNET	Recurrent Endometrial Cancer (#193, pv=0.16)	Adult Astrocytic Tumor (#253, pv=0.22)	ALPHA-THALASSEMIA/MENTAL RETARDATION SYNDROME, NONDELETION TYPE, X-LINKED
	GediNET	PAPILLARY RENAL CELL CARCINOMA	PLASMA CELL NEOPLASM	ADULT GLIOBLASTOMA
GDS2545/ Metastatic prostate cancer	DisGeNET	Metastasis from malignant tumor of prostate (#25, pv=0.01)	Hormone refractory prostate cancer (#274, pv=0.34)	Secondary malignant neoplasm of bone (#62, pv=0.04)
	GediNET	CHILDHOOD RHABDOMYOSARCOMA	RHABDOMYOSARCOMA	SECONDARY MALIGNANT NEOPLASM OF LIVER
GDS2771/ Lung Cancer	DisGeNET	Primary malignant neoplasm of lung (#50, pv=0.03)	Carcinoma of lung (#97, pv=0.08)	Non-Small Cell Lung Carcinoma (#141, pv=0.14)
	GediNET	MANTLE CELL LYMPHOMA	GASTROINTESTINAL CARCINOID TUMOR	MUCINOUS ADENOCARCINOMA
GDS3257/ Lung Adenocarcinoma	DisGeNET	Non-small cell lung cancer recurrent (#116, pv=0.11)	Adenosquamous cell lung cancer (#137, pv=0.15)	Adenocarcinoma, metastatic (#200, 0.22)
	GediNET	ACOUSTIC NEUROMA	ADENOCARCINOMA OF COLON	ADENOCARCINOMA OF ESOPHAGUS
GDS4206/ Pediatric acute leukemia patients with early relapse: white blood cells	DisGeNET	Childhood Leukemia (#96, pv=0.13)	Melanoma (#29, pv=0.03)	Glioblastoma Multiforme (#115, pv=0.18)
	GediNET	ACUTE LEUKEMIA	ADULT DIFFUSE LARGE B-CELL LYMPHOMA	ESOPHAGEAL CARCINOMA
GDS5499/ Pulmonary hypertension	DisGeNET	Idiopathic pulmonary hypertension	Vascular Diseases	Endothelial dysfunction
	GediNET	CHOLANGIOCARCINOMA	HEPATOCARCINOGENESIS	PAPILLOMA
GDS3837/ Non-small cell lung carcinoma in female nonsmokers	DisGeNET	Primary malignant neoplasm of lung	Carcinoma of lung (#10, pv=0.009)	Neoplasm Metastasis
	GediNET	EARLY-STAGE BREAST CARCINOMA	MENINGIOMA, BENIGN, NO ICD-O SUBTYPE	COLORECTAL CARCINOMA

GDS4516_4718/ Colorectal Carcinoma	DisGeNET	Malignant neoplasm of colon and/or rectum (#3, pv=0.002)	Carcinogenesis	Neoplasm Metastasis
	GediNET	ACUTE LEUKEMIA	ACUTE LYMPHOCYTIC LEUKEMIA	Malignant neoplasm of colon and/or rectum
GDS2547/ Metastatic prostate cancer	DisGeNET	Metastasis from malignant tumor of prostate (#27, pv=0.02)	Hormone refractory prostate cancer (#91, pv=0.1)	Secondary malignant neoplasm of bone (#123, pv=0.18)
	GediNET	MALIGNANT NEOPLASM OF LUNG	CARCINOMA OF BLADDER	PROSTATE CARCINOMA
GDS3268/ Ulcerative Colitis	DisGeNET	Crohn Disease	Inflammatory Bowel Diseases	Colitis
	GediNET	MALIGNANT NEOPLASM OF THYROID	ADENOMATOUS POLYPOSIS COLI	LEUKEMIA, MYELOCYTIC, ACUTE

In Table 9, we have included additional information, the values in parenthesis for the rows of DisGeNET are the position of the disease and the p-value assigned by GediNET. Interestingly, excluding just one disease, all the top three significant diseases detected by GediNET are novel. This suggests that the tool detects a new biological knowledge that the biology researcher should consider.

Discussion And Conclusion

In this study, we have developed a novel approach for discovering disease-disease associations and detecting biomarkers of genes associated with the disease.

The approach is based on grouping the genes by their disease association and then scoring those groups in terms of classification significance to train the machine learning model. For example, if the given data is about a specific disease, let's say lung cancer, then the model created from genes that are associated with groups of genes that are related to different diseases will open a biological question about the relationship between those diseases. The traditional approach of searching for genes that could be used as a biomarker in most cases yields a list of significant genes that solve the computational problem and does not take into account any prior knowledge about those genes, as such, their association with diseases or even with other biological knowledge such as microRNA target (see maTE tool [29]), or Pathways(See CogNet tool [30]), GeneOntology (See tool [34]).

Our tool is different in that the search for the significant genes or biomarkers is among groups representing the genes associated with the disease. The final list of genes is the disease-disease associations as presented in Fig. 2, right panel. The knowledge of our tool is more specific and more direct to the relationship between different disease genes and the target disease that is under study.

GediNET identifies the relationships between diseases, Disease–disease association (DDA) based machine learning, which explores novel associations of diseases that enhance knowledge of disease relationships, which could further improve approaches to disease diagnosis, prognosis, and treatment. As we had shown before, GediNET discovered a new unknown relationship between diseases based on the model G-S-M.

References

1. Wang, X.; Gulbahce, N.; Yu, H. Network-Based Methods for Human Disease Gene Prediction. Briefings in Functional Genomics 2011, *10*, 280–293, doi:10.1093/bfgp/elr024.
2. Chen, B.; Shang, X.; Li, M.; Wang, J.; Wu, F.-X. Identifying Individual-Cancer-Related Genes by Rebalancing the Training Samples. IEEE Transactions on NanoBioscience 2016, *15*, 1–1, doi:10.1109/TNB.2016.2553119.
3. Browne, F.; Wang, H.; Zheng, H. A Computational Framework for the Prioritization of Disease-Gene Candidates. BMC Genomics 2015, doi:10.1186/1471-2164-16-S9-S2.
4. Navlakha, S.; Kingsford, C. The Power of Protein Interaction Networks for Associating Genes with Diseases. Bioinformatics 2010, *26*, 1057–1063, doi:10.1093/bioinformatics/btq076.
5. Advances in Translational Bioinformatics: Computational Approaches for the Hunting of Disease Genes | Briefings in Bioinformatics | Oxford Academic Available online: https://academic.oup.com/bib/article/11/1/96/193936 (accessed on 30 November 2021).
6. MiRTarBase 2016: Updates to the Experimentally Validated MiRNA-Target Interactions Database | Nucleic Acids Research | Oxford Academic Available online: https://academic.oup.com/nar/article/44/D1/D239/2503072 (accessed on 30 November 2021).
7. Gene Ontology: Tool for the Unification of Biology | Nature Genetics Available online: https://www.nature.com/articles/ng0500_25/ (accessed on 30 November 2021).
8. Clough, E.; Barrett, T. The Gene Expression Omnibus Database. Methods Mol Biol 2016, *1418*, 93–110, doi:10.1007/978-1-4939-3578-9_5.
9. Tomczak, K.; Czerwińska, P.; Wiznerowicz, M. The Cancer Genome Atlas (TCGA): An Immeasurable Source of Knowledge. *Contemp Oncol (Pozn)* 2015, *19*, A68–A77, doi:10.5114/wo.2014.47136.
10. From Genomics to Chemical Genomics: New Developments in KEGG | Nucleic Acids Research | Oxford Academic Available online: https://academic.oup.com/nar/article/34/suppl_1/D354/1133379 (accessed on 30 November 2021).
11. Piñero, J.; Bravo, À.; Queralt-Rosinach, N.; Gutiérrez-Sacristán, A.; Deu-Pons, J.; Centeno, E.; García-García, J.; Sanz, F.; Furlong, L.I. DisGeNET: A Comprehensive Platform Integrating Information on Human Disease-Associated Genes and Variants. Nucleic Acids Research 2016, *45*, doi:10.1093/nar/gkw943.
12. Gillis, J.; Pavlidis, P. "Guilt by Association" Is the Exception Rather Than the Rule in Gene Networks. PLOS Computational Biology 2012, *8*, e1002444, doi:10.1371/journal.pcbi.1002444.
13. Ben-dor, A. Gene-Expression Profiles in Hereditary Breast Cancer. Advances in Anatomic Pathology 2002.
14. Bittner, M.; Meltzer, P.; Chen, Y.; Jiang, Y.; Seftor, E.; Hendrix, M.; Radmacher, M.; Simon, R.; Yakhini, Z.; Ben-Dor, A.; et al. Molecular Classification of Cutaneous Malignant Melanoma by Gene Expression Profiling. Nature 2000, *406*, 536–540, doi:10.1038/35020115.
15. van Driel, M.A.; Brunner, H.G. Bioinformatics Methods for Identifying Candidate Disease Genes. Hum Genomics 2006, *2*, 429, doi:10.1186/1479-7364-2-6-429.
16. Asif, M.; Martiniano, H.F.M.C.M.; Vicente, A.M.; Couto, F.M. Identifying Disease Genes Using Machine Learning and Gene Functional Similarities, Assessed through Gene Ontology. PLoS ONE 2018, *13*, e0208626, doi:10.1371/journal.pone.0208626.
17. Multi-View Based Integrative Analysis of Gene Expression Data for Identifying Biomarkers | Scientific Reports Available online: https://www.nature.com/articles/s41598-019-49967-4 (accessed on 30 November 2021).
18. Liekens, A.M.; De Knijf, J.; Daelemans, W.; Goethals, B.; De Rijk, P.; Del-Favero, J. BioGraph: Unsupervised Biomedical Knowledge Discovery via Automated Hypothesis Generation. Genome Biology 2011, *12*, R57, doi:10.1186/gb-2011-12-6-r57.

19. Wang, J.; Zheng, J.; Wang, Z.; Li, H.; Deng, M. Inferring Gene-Disease Association by an Integrative Analysis of EQTL Genome-Wide Association Study and Protein-Protein Interaction Data. Hum Hered 2018, *83*, 117–129, doi:10.1159/000489761.
20. He, X.; Fuller, C.K.; Song, Y.; Meng, Q.; Zhang, B.; Yang, X.; Li, H. Sherlock: Detecting Gene-Disease Associations by Matching Patterns of Expression QTL and GWAS. The American Journal of Human Genetics 2013, *92*, 667–680, doi:10.1016/j.ajhg.2013.03.022.
21. Giambartolomei, C.; Vukcevic, D.; Schadt, E.E.; Franke, L.; Hingorani, A.D.; Wallace, C.; Plagnol, V. Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. PLoS Genet 2014, *10*, e1004383, doi:10.1371/journal.pgen.1004383.
22. Greene, C.S.; Krishnan, A.; Wong, A.K.; Ricciotti, E.; Zelaya, R.A.; Himmelstein, D.S.; Zhang, R.; Hartmann, B.M.; Zaslavsky, E.; Sealfon, S.C.; et al. Understanding Multicellular Function and Disease with Human Tissue-Specific Networks. Nat Genet 2015, *47*, 569–576, doi:10.1038/ng.3259.
23. Peng, J.; Bai, K.; Shang, X.; Wang, G.; Xue, H.; Jin, S.; Cheng, L.; Wang, Y.; Chen, J. Predicting Disease-Related Genes Using Integrated Biomedical Networks. BMC Genomics 2017, *18*, 1043, doi:10.1186/s12864-016-3263-4.
24. Luo, P.; Tian, L.-P.; Chen, B.; Xiao, Q.; Wu, F.-X. Ensemble Disease Gene Prediction by Clinical Sample-Based Networks. BMC Bioinformatics 2020, *21*, 79, doi:10.1186/s12859-020-3346-8.
25. Piñero, J.; Bravo, À.; Queralt-Rosinach, N.; Gutiérrez-Sacristán, A.; Deu-Pons, J.; Centeno, E.; García-García, J.; Sanz, F.; Furlong, L.I. DisGeNET: A Comprehensive Platform Integrating Information on Human Disease-Associated Genes and Variants. Nucleic Acids Res 2017, *45*, D833–D839, doi:10.1093/nar/gkw943.
26. Hamzeh, O.; Rueda, L. A *Gene-Disease-Based Machine Learning Approach to Identify Prostate Cancer Biomarkers*; 2019; p. 638; ISBN 978-1-4503-6666-3.
27. Yousef, M.; Bakir-Gungor, B.; Jabeer, A.; Goy, G.; Qureshi, R.; C. Showe, L. Recursive Cluster Elimination Based Rank Function (SVM-RCE-R) Implemented in KNIME. *F1000Res* 2020, *9*, 1255, doi:10.12688/f1000research.26880.1.
28. SVM-RCE-R-OPT: Optimization of Scoring Function for SVM-RCE-R Available online: https://www.springerprofessional.de/en/svm-rce-r-opt-optimization-of-scoring-function-for-svm-rce-r/19677024 (accessed on 15 March 2022).
29. Yousef, M.; Abdallah, L.; Allmer, J. MaTE: Discovering Expressed Interactions between MicroRNAs and Their Targets. Bioinformatics 2019, *35*, 4020–4028, doi:10.1093/bioinformatics/btz204.
30. Yousef, M.; Ülgen, E.; Uğur Sezerman, O. CogNet: Classification of Gene Expression Data Based on Ranked Active-Subnetwork-Oriented KEGG Pathway Enrichment Analysis. PeerJ Computer Science 2021, *7*, e336, doi:10.7717/peerj-cs.336.
31. Yousef, M.; Goy, G.; Mitra, R.; Eischen, C.M.; Jabeer, A.; Bakir-Gungor, B. MiRcorrNet: Machine Learning-Based Integration of MiRNA and MRNA Expression Profiles, Combined with Feature Grouping and Ranking. PeerJ 2021, *9*, e11458, doi:10.7717/peerj.11458.
32. Yousef, M.; Goy, G.; Bakir-Gungor, B. MiRModuleNet: Detecting MiRNA-MRNA Regulatory Modules. Front. Genet. 2022, *13*, 767455, doi:10.3389/fgene.2022.767455.
33. Yousef, M.; Ozdemir, F.; Jaaber, A.; Allmer, J.; Bakir-Gungor, B. *PriPath: Identifying Dysregulated Pathways from Differential Gene Expression via Grouping, Scoring and Modeling with an Embedded Machine Learning Approach*; In Review, 2022;
34. Yousef, M.; Sayıcı, A.; Bakir-Gungor, B. Integrating Gene Ontology Based Grouping and Ranking into the Machine Learning Algorithm for Gene Expression Data Analysis. In *Database and Expert Systems Applications - DEXA 2021 Workshops*; Kotsis, G., Tjoa, A.M., Khalil, I., Moser, B., Mashkoor, A., Sametinger, J., Fensel, A., Martinez-Gil, J., Fischer, L., Czech, G., Sobieczky, F., Khan, S., Eds.; Communications in Computer and Information Science; Springer International Publishing: Cham, 2021; Vol. 1479, pp. 205–214 ISBN 978-3-030-87100-0.

35. Yousef, M.; Ketany, M.; Manevitz, L.; Showe, L.C.; Showe, M.K. Classification and Biomarker Identification Using Gene Network Modules and Support Vector Machines. BMC bioinformatics 2009, *10*, 337, doi:10.1186/1471-2105-10-337.
36. Yousef, M.; Jung, S.; Showe, L.C.; Showe, M.K. Recursive Cluster Elimination (RCE) for Classification and Feature Selection from Gene Expression Data. BMC Bioinformatics 2007, *8*, 144, doi:10.1186/1471-2105-8-144.
37. Yousef, M.; Jabeer, A.; Bakir-Gungor, B. SVM-RCE-R-OPT: Optimization of Scoring Function for SVM-RCE-R. In *Database and Expert Systems Applications - DEXA 2021 Workshops*; Kotsis, G., Tjoa, A.M., Khalil, I., Moser, B., Mashkoor, A., Sametinger, J., Fensel, A., Martinez-Gil, J., Fischer, L., Czech, G., Sobieczky, F., Khan, S., Eds.; Communications in Computer and Information Science; Springer International Publishing: Cham, 2021; Vol. 1479, pp. 215–224 ISBN 978-3-030-87100-0.
38. Yousef, M.; Kumar, A.; Bakir-Gungor, B. Application of Biological Domain Knowledge Based Feature Selection on Gene Expression Data. Entropy (Basel) 2020, *23*, E2, doi:10.3390/e23010002.
39. Piñero, J.; Bravo, À.; Queralt-Rosinach, N.; Gutiérrez-Sacristán, A.; Deu-Pons, J.; Centeno, E.; García-García, J.; Sanz, F.; Furlong, L.I. DisGeNET: A Comprehensive Platform Integrating Information on Human Disease-Associated Genes and Variants. Nucleic Acids Res 2017, *45*, D833–D839, doi:10.1093/nar/gkw943.
40. Kolde, R.; Laur, S.; Adler, P.; Vilo, J. Robust Rank Aggregation for Gene List Integration and Meta-Analysis. Bioinformatics 2012, *28*, 573–580, doi:10.1093/bioinformatics/btr709.
41. Berthold, M.R.; Cebron, N.; Dill, F.; Gabriel, T.R.; Kötter, T.; Meinl, T.; Ohl, P.; Sieb, C.; Thiel, K.; Wiswedel, B. KNIME: The Konstanz Information Miner. In Proceedings of the Data Analysis, Machine Learning and Applications; Preisach, C., Burkhardt, H., Schmidt-Thieme, L., Decker, R., Eds.; Springer: Berlin, Heidelberg, 2008; pp. 319–326.
42. Xu, Q.-S.; Liang, Y.-Z. Monte Carlo Cross Validation. Chemometrics and Intelligent Laboratory Systems 2001, *56*, 1–11, doi:10.1016/S0169-7439(00)00122-2.
43. El-Hadj Imorou, S. Socio-Economic and Health Determinants of Rural Households Consent to Prepay for Their Health Care in N'Dali (North of Benin). JSS 2020, *08*, 348–360, doi:10.4236/jss.2020.85024.
44. Hand, D.; Till, R. A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems. *undefined* 2004.
45. DAVID: Functional Annotation Tools Available online: https://david.ncifcrf.gov/tools.jsp (accessed on 8 April 2022).
46. Kuleshov, M.V.; Jones, M.R.; Rouillard, A.D.; Fernandez, N.F.; Duan, Q.; Wang, Z.; Koplev, S.; Jenkins, S.L.; Jagodnik, K.M.; Lachmann, A.; et al. Enrichr: A Comprehensive Gene Set Enrichment Analysis Web Server 2016 Update. Nucleic Acids Res 2016, *44*, W90–W97, doi:10.1093/nar/gkw377.
47. GeneMANIA Available online: https://genemania.org/ (accessed on 8 April 2022).
48. Martens, M.; Ammar, A.; Riutta, A.; Waagmeester, A.; Slenter, D.N.; Hanspers, K.; A. Miller, R.; Digles, D.; Lopes, E.N.; Ehrhart, F.; et al. WikiPathways: Connecting Communities. Nucleic Acids Research 2021, *49*, D613–D621, doi:10.1093/nar/gkaa1024.
49. Franz, M.; Lopes, C.T.; Huck, G.; Dong, Y.; Sumer, O.; Bader, G.D. Cytoscape.Js: A Graph Theory Library for Visualisation and Analysis. Bioinformatics 2016, *32*, 309–311, doi:10.1093/bioinformatics/btv557.
50. Menche, J.; Sharma, A.; Kitsak, M.; Ghiassian, S.D.; Vidal, M.; Loscalzo, J.; Barabási, A.-L. Disease Networks. Uncovering Disease-Disease Relationships through the Incomplete Interactome. Science 2015, *347*, 1257601, doi:10.1126/science.1257601.
51. Suratanee, A.; Plaimas, K. DDA: A Novel Network-Based Scoring Method to Identify Disease-Disease Associations. Bioinform Biol Insights 2015, *9*, BBI.S35237, doi:10.4137/BBI.S35237.

Figures

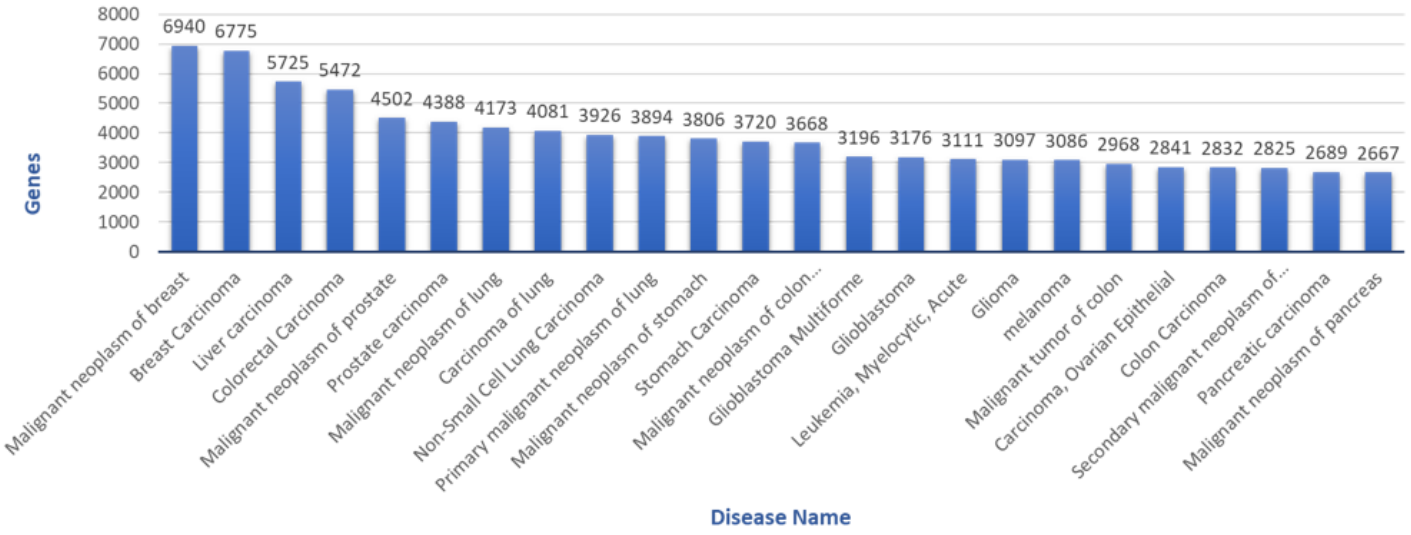

Figure 1

Histogram frequency plot shows the number of genes associated with each disease, where the X-axis is the disease name, and Y-axis is the number of genes.

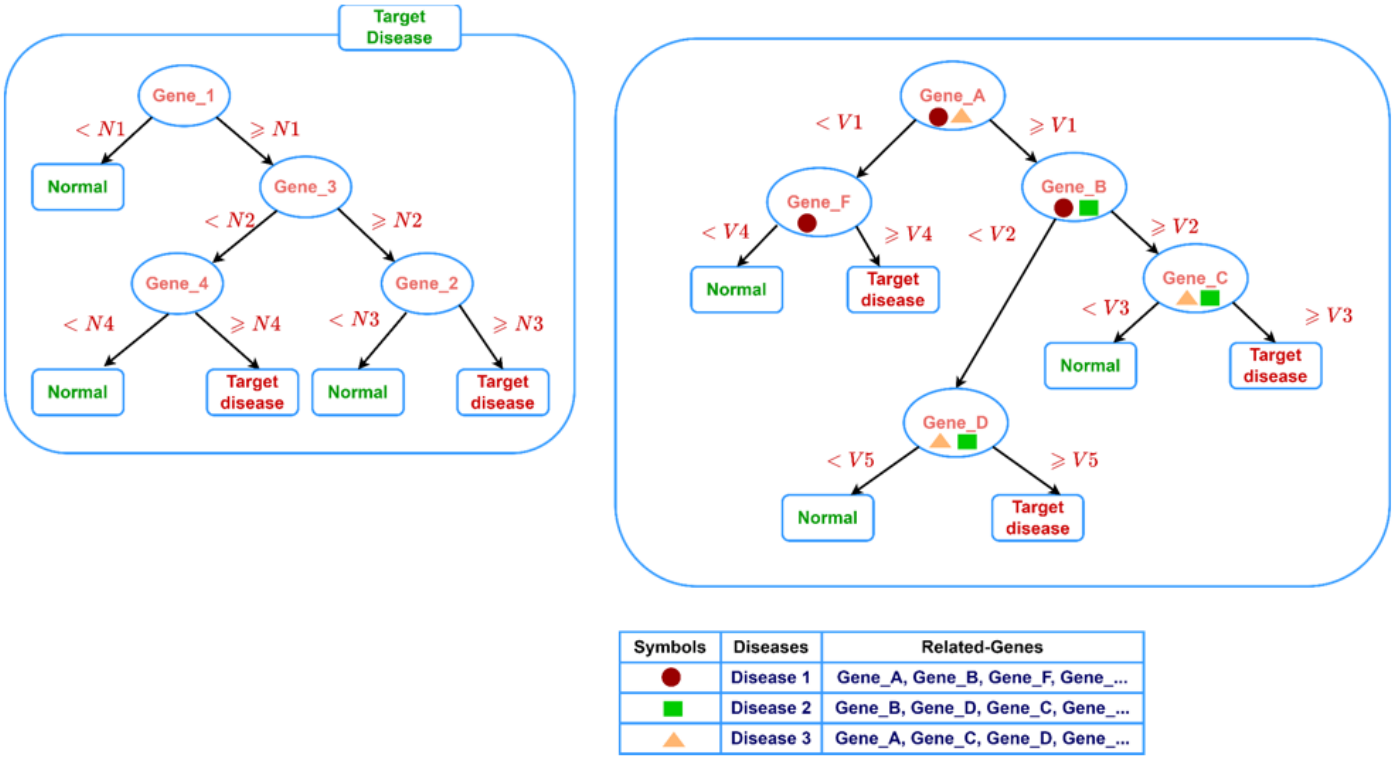

Figure 2

The left panel illustrates the traditional approach that detects gene-disease associations, while the right panel illustrates the disease-disease association as the output of GeDiNET.

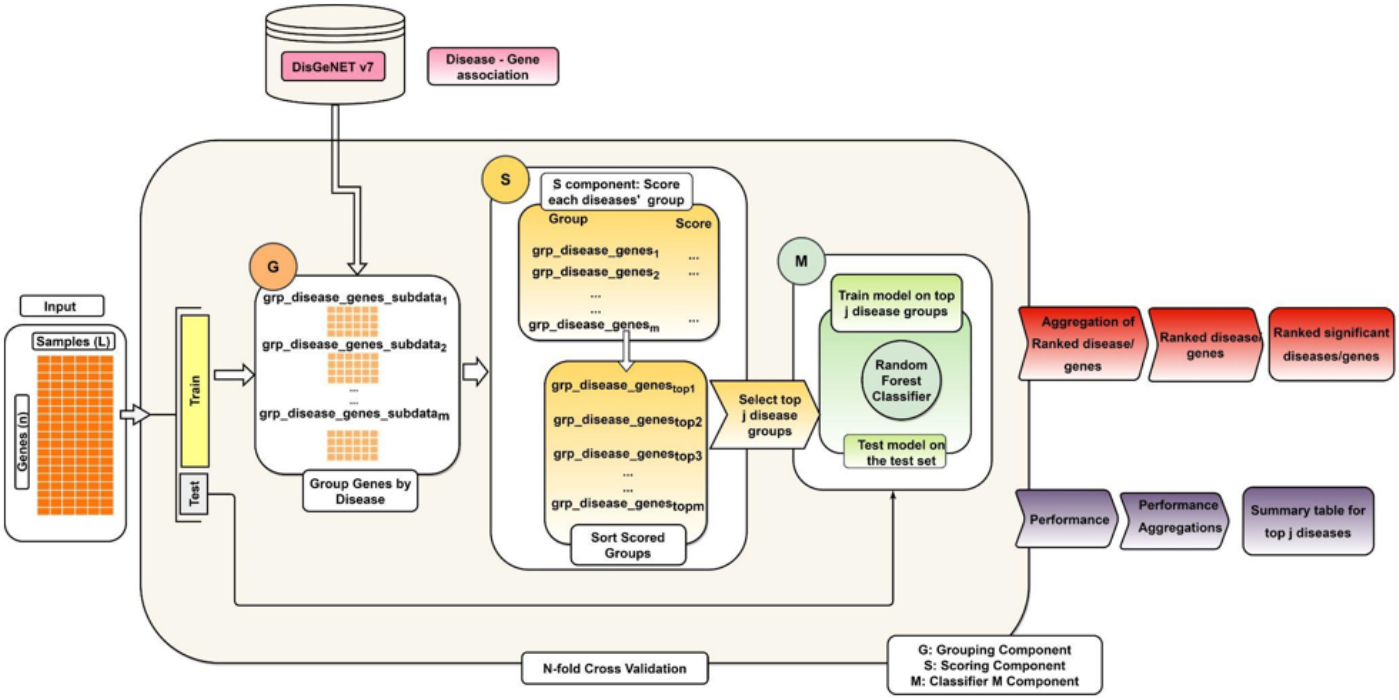

Figure 3

GeDiNET workflow. The main workflow for integrating biological information for grouping genes based on Disease-Gene association is derived from the DisGeNET v7 database.

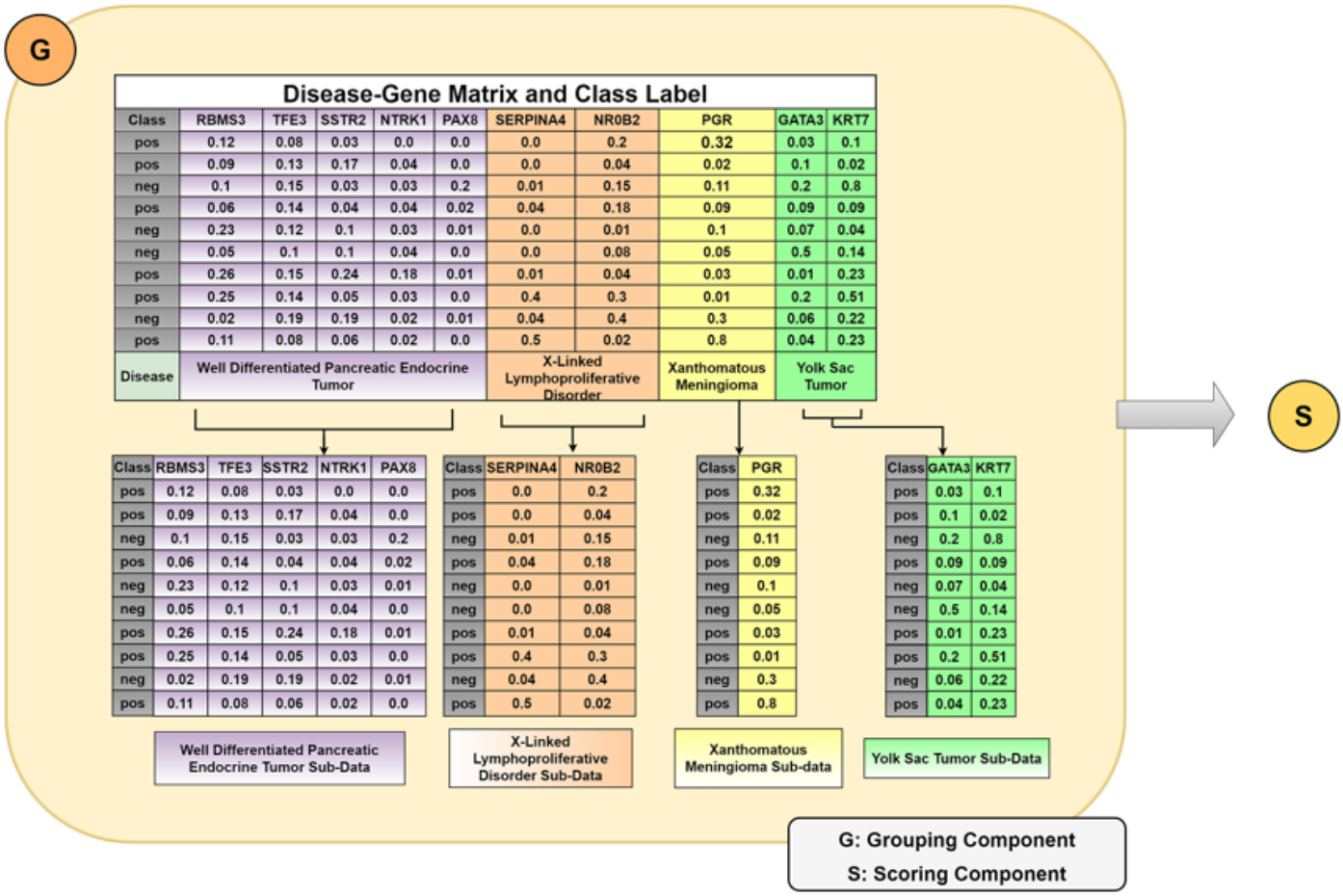

Figure 4

An example of creating sub-data extracted according to disease-group names. These sub-datasets will be subject to the S component for scoring.

Figure 5

GeDiNET Workflow in KNIME

Figure 6

Expanded workflow (meta-nodes) for GeDiNET.

Figure 7

The mean AUC values of GeDiNET, CogNet, maTE, and PriPath for 10 different datasets for the top two clusters.

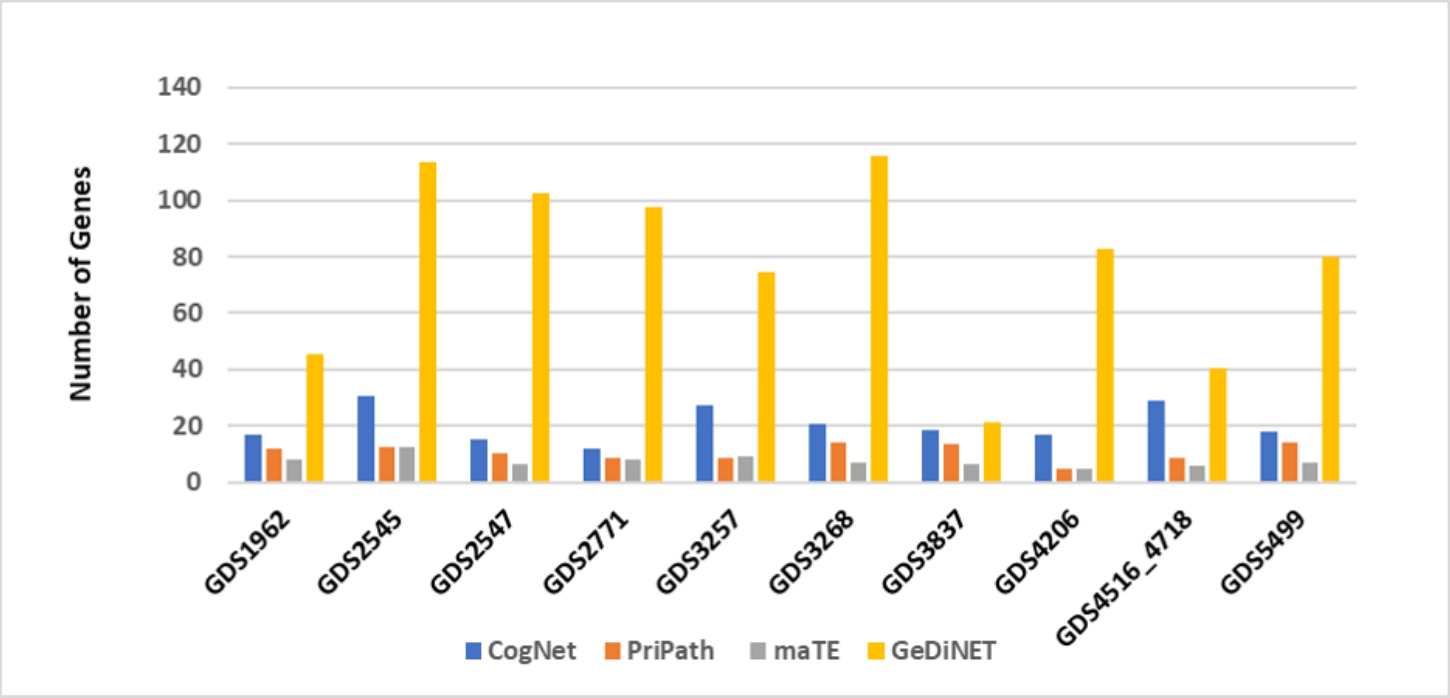

Figure 8

The mean number of genes of GeDiNET, CogNet, maTE, and PriPath tools for 10 different datasets for the top two clusters.

Figure 9

Figure 11. Network visualization of the Gene interaction for the cell signaling pathway with overlapping genes for the 10 GEO datasets.

Figure 10

Figure 12. Network visualization of the cell signaling pathway with overlapping genes for the GDS3257 dataset.

Figure 11

Figure 13. An example of the DDA for 4 datasets in Gedinet where the number of shared genes for the top-scored disease group is represented. The upper panel shows the DDA for GDS1962, GDS3257, GDS2771, and GDS5499 datasets. The lower panel shows the annotations used in the DDA illustration formation.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryData.docx