



OPEN ACCESS

EDITED BY

Andreas Pfenning,
Carnegie Mellon University, United States

REVIEWED BY

Laura La Paglia,
National Research Council (CNR), Italy
Ettore Mosca,
National Research Council (CNR), Italy

*CORRESPONDENCE

Malik Yousef,
✉ malik.yousef@gmail.com

RECEIVED 06 January 2023

ACCEPTED 05 July 2023

PUBLISHED 21 August 2023

CITATION

Ersoz NS, Bakir-Gungor B and Yousef M (2023), GeNetOntology: identifying affected gene ontology terms via grouping, scoring, and modeling of gene expression data utilizing biological knowledge-based machine learning. *Front. Genet.* 14:1139082. doi: 10.3389/fgene.2023.1139082

COPYRIGHT

© 2023 Ersoz, Bakir-Gungor and Yousef. This is an open-access article distributed under the terms of the [Creative Commons Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

GeNetOntology: identifying affected gene ontology terms via grouping, scoring, and modeling of gene expression data utilizing biological knowledge-based machine learning

Nur Sebnem Ersoz¹, Burcu Bakir-Gungor^{2,3} and Malik Yousef^{4,5*}

¹Department of Bioengineering, Graduate School of Engineering and Science, Abdullah Gul University, Kayseri, Türkiye, ²Department of Computer Engineering, Faculty of Engineering, Abdullah Gul University, Kayseri, Türkiye, ³Department of Bioengineering, Faculty of Life and Natural Sciences, Abdullah Gul University, Kayseri, Türkiye, ⁴Department of Information Systems, Zefat Academic College, Zefat, Israel, ⁵Galilee Digital Health Research Center (GDH), Zefat Academic College, Zefat, Israel

Introduction: Identifying significant sets of genes that are up/downregulated under specific conditions is vital to understand disease development mechanisms at the molecular level. Along this line, in order to analyze transcriptomic data, several computational feature selection (i.e., gene selection) methods have been proposed. On the other hand, uncovering the core functions of the selected genes provides a deep understanding of diseases. In order to address this problem, biological domain knowledge-based feature selection methods have been proposed. Unlike computational gene selection approaches, these domain knowledge-based methods take the underlying biology into account and integrate knowledge from external biological resources. Gene Ontology (GO) is one such biological resource that provides ontology terms for defining the molecular function, cellular component, and biological process of the gene product.

Methods: In this study, we developed a tool named GeNetOntology which performs GO-based feature selection for gene expression data analysis. In the proposed approach, the process of Grouping, Scoring, and Modeling (G-S-M) is used to identify significant GO terms. GO information has been used as the grouping information, which has been embedded into a machine learning (ML) algorithm to select informative ontology terms. The genes annotated with the selected ontology terms have been used in the training part to carry out the classification task of the ML model. The output is an important set of ontologies for the two-class classification task applied to gene expression data for a given phenotype.

Results: Our approach has been tested on 11 different gene expression datasets, and the results showed that GeNetOntology successfully identified important disease-related ontology terms to be used in the classification model.

Discussion: GeNetOntology will assist geneticists and scientists to identify a range of disease-related genes and ontologies in transcriptomic data analysis, and it will also help doctors design diagnosis platforms and improve patient treatment plans.

KEYWORDS

gene ontology, gene expression data analysis, machine learning, feature selection, enrichment analysis, feature scoring, feature grouping, classification

1 Introduction

Precision medicine gained importance in the last decade (König et al., 2017). Molecular abnormalities in disease formation can differ from patient to patient, and a more customized medication plan is required for each individual. Fortunately, today's next-generation sequencing technologies offer several opportunities to quickly generate a series of omics data in order to monitor molecular alterations at different levels. Along this line, next-generation sequencing has been widely used to analyze genetic variations, gene-expression profiling, epigenomes, and genome diversity (Levy and Boone, 2019). Furthermore, with improvement in the technology, the cost of these high-throughput technologies is decreasing. However, the reduction in the cost of next-generation sequencing and other high-throughput technologies creates a burden on the data analysis approaches.

Biological systems are quite complex by their nature. Therefore, one of the difficulties in omics data analysis stems from these complex data, where the information is present at multiple layers. For example, in a biological system, the gene activities under different physiological states are reflected by the gene expression data (at the transcriptome level). On the other hand, the epigenome reflects the effects of environmental factors on gene activities and disease development. Epigenetic modifications such as DNA methylations and histone modifications (acetylation, methylation, phosphorylation, and ubiquitination) can alter the genome and regulate gene expression. Hence, epigenomic variations can control transcriptomes which can cause proteomic dysfunctions and result in disease formation. A similar scenario may also be caused by genomic variations. In addition to that, post-translational modifications play roles in phenotypic differentiation in physiology and pathology. Such complexities of biological systems make it harder to analyze biological data.

In order to enlighten the molecular and functional mechanisms of disease development, one of the widely studied data types is transcriptomic data (Barrett et al., 2012; Perscheid et al., 2019). Researchers analyze gene expression datasets to identify disease-associated genes and to find biomarkers that can aid in early diagnosis and targeted therapies. Various technologies, such as microarray and RNA-seq technology, can detect thousands of transcripts. One of the challenges in gene expression data analysis stems from it being noisy and high dimensional. It has a high number of features (i.e., genes or mRNAs) measured as a result of the experiments conducted at the molecular level, with a small number of samples including the patient group, control group, and treated or non-treated groups. The expression values of the genes are provided to a learning algorithm to accomplish the classification task. However, usually, the majority of the genes (i.e., features) are redundant, noisy, or irrelevant to the learning task, which will influence the learning accuracy and training speed (Ang et al., 2016; Aziz et al., 2017). In other words, only a number of genes are primarily related to disease development. Therefore, it is crucial to find disease-related transcripts (i.e., genes) by applying feature selection methods. In other words, gene selection refers to

biomarker detection via applying feature selection methods on gene expression data (Perscheid, 2021).

Traditional feature selection (FS) approaches are mainly based on statistical tests. In the literature, several comparative studies were carried out on different FS methods (Albattah et al., 2022). Several comprehensive studies of different FS methods are provided for classification problems (Bolón-Canedo et al., 2016). Yet, another recent survey discussed the basics, applications, and challenges of FS methods in the context of high-dimensional data (Abdulwahab et al., 2022). According to the interaction of the FS method with the classification model, Liu and Motoda (1998) originally classified the FS methods into filter, wrapper, and embedded methods (Guyon et al., 2006; Jensen and Shen, 2008; Chandrashekar and Sahin, 2014; You et al., 2014; Tadist et al., 2019; Albattah et al., 2022). Later on, hybrid and ensemble methods were proposed in the literature as variants of them (Bellazzi and Zupan, 2007; Ang et al., 2016; Perscheid et al., 2019; Tadist et al., 2019).

Filter methods evaluate a subset of features or a feature only by using the intrinsic properties of the training samples. These methods can be combined with a variety of classifiers, and therefore, filter methods have a better generalization ability and lower computational complexity. Filter methods are based on F-statistics (ANOVA, t-test, etc.), mutual information, and entropy (Srinivasa et al., 2020), and they evaluate the influence of the input values on the output value. Some examples of the filter methods are information gain (IG) and ReliefF (Perscheid et al., 2019).

Wrapper methods can achieve better classification performance than filter methods because they are specific to a particular classifier (Inza et al., 2004). These methods assess the quality of a candidate subset. The successive feature selection (SFS) approach is an example of the wrapper-type feature selection method (Perscheid et al., 2019). The main disadvantage of the wrapper method is that they are far more time-consuming.

As a special case of the wrapper methods, embedded methods are characterized by an interaction between the FS and the classification algorithm. When embedded methods are used to construct the classifier, feature subsets are generated (Wang et al., 2015). The support vector machine with recursive feature elimination (SVM-RFE) approach is an example of embedded FS methods. As summarized here, traditional methods are fully data-driven approaches, and they neglect biological domain knowledge. For example, when selecting important genes in transcriptome data analysis, the importance of each gene is usually evaluated with a filtering method, ignoring the interactions and relationships between the genes. On the other hand, wrapper methods utilize the learning algorithm while evaluating the features (i.e., genes). Hence, they are able to find optimal feature sets, but they may encounter the overfitting problem.

It is reported that one of the main obstacles of traditional methodologies is that they hardly perform biological interpretation, and hence they do not allow the creation of new biological knowledge (Perscheid et al., 2019; Yousef et al., 2021). Since traditional gene selection approaches have limitations like noise due to the high-dimensional data (Perscheid, 2021), lately, scientists have started to develop integrative gene selection

approaches that incorporate domain knowledge from external biological resources during the gene expression data analysis (Perscheid et al., 2019; Yousef et al., 2021). Genes perform their biological functions in an organized fashion (in terms of metabolic networks and signaling pathways). Hence, scientists attempt to develop new methodologies which employ external biological information such as pathways, interactions, and gene ontology (Yousef et al., 2021). To this end, the integrative gene selection process generates a ranked list of genes according to both statistical metrics and biological background information collected from external resources (Perscheid et al., 2019).

There are several resources, databases, and repositories of biological knowledge. For example, Kyoto Encyclopedia of Genes and Genomes (KEGG) is a widely used external ontology resource, which serves as a pathway knowledge-base for systemic analysis of gene functions, and it provides manually curated pathways (Kanehisa and Goto, 2000). As another resource, The Cancer Genome Atlas (TCGA) (Tomczak et al., 2015) hosts detailed information about oncogenomic expression profiles. On the other hand, miRTarbase (Chou et al., 2018) serves annotated experimentally validated miRNA–target interactions. An integrative gene selection method could also utilize functional information from the UniProt Knowledge Base (UniProtKB) (The UniProt Consortium, 2017). DisGeNET provides a system-level view on the genes and diseases via giving access to other data sources such as RNA and interaction graphs (Perscheid, 2021). As one of the widely used biological knowledge bases, GO intends to unite detailed and standardized terminologies defined for the various levels of molecular biology (Balakrishnan et al., 2013). It supplies tools for exploring these terminologies and for describing biological terms using this vocabulary (The Gene Ontology Consortium, 2019).

The GO Consortium created GO with the aim of presenting a fully defined, organized terminology to describe the gene functions and products in each organism (Ashburner et al., 2000). Primarily, Mouse Genome Informatics (MGI), FlyBase, and *Saccharomyces* Genome Database (SGD) model organism databases were used by the GO Consortium and then were expanded to many organisms. In the ontology, over 45,000 terms have been connected by about 134,000 relations. In addition, more than 7 million genes and the annotations of gene products from over 3,200 organisms are included in the GO knowledge base (The Gene Ontology Consortium, 2019). Three different aspects of genes are covered by GO, i.e., biological process (BP), molecular function (MF), and cellular component (CC). Although MF represents the activity of a gene product at the molecular level, CC represents the cellular localization of the gene product, or where it acts. Lastly, the BP is the larger biological objective that the molecular-level process of the gene product contributes to (The Gene Ontology Consortium (2019)).

GO annotations constructed by connecting specific gene products to terms in the ontology are also contained in the GO knowledge base. Each information includes evidence on which it is based, using the standardized codes, computational analysis evidence codes such as Inferred from Sequence or structural Similarity (ISS), curatorial statement evidence codes such as Inferred by Curator (IC), and electronic annotation evidence code such as inferred from electronic annotation (IEA), where these codes are defined by the Evidence and Conclusion Ontology (ECO) (Chibucos et al., 2017; Guide to GO evidence codes, 2022).

Recently, Perscheid published a survey on prior knowledge-based approaches for integrative biomarker detection from gene expression datasets (Perscheid, 2021). In that article, she evaluated the respective characteristics of different integrative gene selection approaches and presented an overview of external knowledge bases that are utilized in these approaches (Perscheid, 2021). The article reported that GO and KEGG resources are predominantly used as external knowledge bases for integrative gene selection. For example, Qi and Tang (2007) showed that incorporating GO as a biological knowledge outperforms traditional gene selection methods in microarray data analysis. Another approach used GO and KEGG ontologies to filter genes more accurately (Fang et al., 2014). GO terms are also used in another study to show the limitations of network-based annotations (Asif et al., 2018).

The same review paper (Perscheid, 2021) noted that although prior knowledge-based approaches offer several advantages for gene selection, these approaches require advanced integration concepts to consider both statistical and biological characteristics. As a result, these approaches have not been widely adopted. In this respect, recently, we proposed a Grouping–Scoring–Modeling (G-S-M) approach (Yousef et al., 2021) for integrating biological knowledge into the machine learning model. The G-S-M approach selects a set of features where different sets can be generated via 1) using pre-existing biological knowledge stored in a database (such as miRTarBase (Chou et al., 2018), DisGeNET (Piñero et al., 2015), and KEGG pathways (Kanehisa and Goto, 2000)) or 2) fully data-driven approach using statistical measures such as Pearson's correlations. The G-S-M approach has been utilized in the development of different computational tools. Examples of such tools are maTE (Yousef et al., 2019) that uses microRNA target gene information for grouping the genes; miRcorrNet (Yousef et al., 2021) and miRModuleNet (Yousef et al., 2022), which detect feature sets via concurrently analyzing mRNA and miRNA expression datasets, respectively; CogNet (Yousef et al., 2021) and PriPath (Yousef et al., 2022) that use KEGG pathway information for grouping the genes; GediNet (Qumsiyeh et al., 2022) that uses disease gene associations from DisGeNET while defining the sets of the genes; and miRdisNET (Jabeer et al., 2023) that uses miRNA target gene information while assigning the genes into sets. As the recent review paper (Perscheid, 2021) points out, biomarker detection only based on statistical analysis is insufficient. To this end, here, we attempt to incorporate external biological knowledge into the selection process, and hence we aim to deliver biologically relevant results. In other words, this study mainly focuses on the detection of disease signatures and on the discovery of novel gene sets with relations across a subset of GO terms for the disease under investigation. In this study, our main aim is to assign genes into groups using the G-S-M approach and to identify the highly correlated sets of GO terms that are related to the disease under investigation. Along this line, in this study, the GeNetOntology algorithm is proposed as a novel algorithm that improves classification performance by utilizing GO as external biological information while selecting the most relevant genes from gene expression datasets. In our experiments, the Monte Carlo cross-validation (MCCV) technique is utilized. Hence, in each iteration, some samples are selected randomly for the training set, and the rest of the samples are selected for the testing set. In each training iteration, the most informative GO term is identified. Later, the genes that are associated with the top-ranked GO term in each iteration is merged to

TABLE 1 Description of the 11 gene expression datasets that have been used in this study. Each entry has the GEO Accession, PMID, disease type, number of healthy samples (controls), and number of patients.

Title	GEO accession	PMID	Disease type	# Of healthy	# Of patients
Glioma-derived stem cell factor effect on angiogenesis in the brain	GDS1962	16616334	Glioma	23	157
Early-stage Parkinson's disease	GDS2519	17215369	Parkinson's disease	23	50
		18669654			
Metastatic prostate cancer (HG-U95C)	GDS2545	17430594	Prostate cancer	81	90
		15254046			
Metastatic prostate cancer (HG-U95C)	GDS2547	17430594	Prostate cancer	75	89
		15254046			
Large airway epithelial cells from cigarette smokers with suspected lung cancer	GDS2771	17334370	Lung cancer	90	102
		20375364			
Cigarette smoking effect on lung adenocarcinoma	GDS3257	18297132	Lung adenocarcinoma	49	58
Colon epithelial biopsies of ulcerative colitis patients	GDS3268	18523026	Colitis	73	129
Non-small cell lung carcinoma in female non-smokers	GDS3837	20802022	Lung cancer	60	60
		25889623			
Pediatric acute leukemia patients with early relapse: white blood cells	GDS4206	21295523	Leukemia	157	40
Colorectal cancer: laser microdissected tumor tissues	GDS4516_4718	21270110	Colorectal cancer	44	104
Pulmonary hypertension: PBMCs	GDS5499	22545094	Pulmonary hypertension	41	99

train the model. In addition, we perform comparative evaluation with other existing methods. The novelty and originality of our approach stems from its capability to explore GO terms to classify and to find the most relevant sets of GO terms associated with the disease under study. In this respect, our approach differs from traditional gene selection approaches where searching is carried out by considering individual genes.

2 Materials and methods

2.1 Gene expression dataset

A total of 11 gene expression datasets for different types of human complex diseases were downloaded from Gene Expression Omnibus (GEO) (<https://www.ncbi.nlm.nih.gov/geo/>). All datasets include both healthy samples (labeled as negative) and patient samples (labeled as positive). These datasets are used to test the performance of GeNetOntology and to compare it with that of other tools. The gene expression dataset is represented as a matrix. In this matrix, while genes (i.e., mRNAs) are shown in the columns, rows represent the samples. This matrix contains a special column called label, which indicates the class annotation for each row. Here, the class labels are either positive, indicating the patient, or negative, indicating the control. Table 1 presents the GEO accession numbers, titles, PubMed Identification numbers (PMID), disease name, and numbers of cases and controls for each one of the 11 gene expression datasets.

TABLE 2 Summary of the GO subsets with the number of terms associated with each GO subset.

Subset of gene ontology (GO)	#Ontology groups (terms)
All ontology gene sets	14,998
GO biological process (BP)	7,481
GO molecular functions (MF)	1,708
GO cellular component (CC)	996
HPO	4,813

2.2 Gene ontology data

The GO ('Gene Ontology Consortium: going forward', 2015) (<http://www.geneontology.org>) database provides the biological knowledge that will be used for the grouping component. The GO and Human Phenotype Ontology (HPO) data are downloaded from the Molecular Signatures Database (*GSEA | MSigDB | Browse Human Gene Sets*) (Liberzon et al., 2015). HPO terms and all GO terms from GO BP, GO CC, and GO MF categories are used in this study. The numbers of the GO terms in each category and the number of HPO terms are listed in Table 2. As illustrated in Supplementary Figure S1, each GO term is associated with one or more genes and it can be represented as a gene set.

We wanted to see the distribution of the number of genes that are associated with each GO term. For this purpose, for each one of

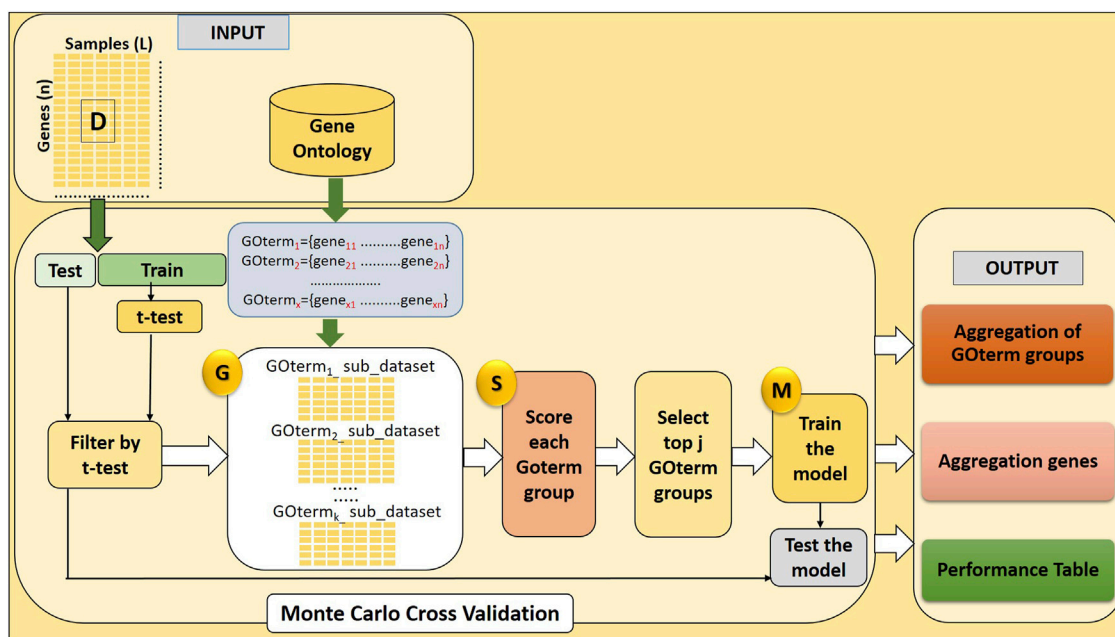


FIGURE 1

GeNetOntology consists of four major components: GO term groups are extracted; sub-datasets for each GO term groups are generated by G component; component S scores and ranks the groups; and component M creates and evaluates the model.

the BP, CC, and MF categories, we further divide GO terms into bins based on the number of genes that are associated with each GO term. [Supplementary Figure S2](#) presents three histograms for each one of the GO BP, CC, and MF categories. In each histogram, we plot the counts of GO terms, where n genes are associated with that GO term and n increases in a window of 20. When the number of genes associated with a GO term (n) is increased by 20, we get 7,481 bins for the BP category, 996 bins for the CC category, and 1,708 bins for the MF category.

Genes that are annotated with the same GO term either share a common function, perform similar activity depending on their responsibility at the molecular or cellular level, or act within the same cellular component. Genes that are annotated with GO Biological Process play roles in biological processes. Furthermore, each gene that is annotated with a specific GO biological process term performs a specific biological process. For example, genes that are annotated with the GOBP_Artery_Morphogenesis term are responsible for artery morphogenesis.

2.3 The general G-S-M model

The main idea of the G-S-M technique is to perform the scoring operation for different sets of features rather than selecting and evaluating individual features. Biological knowledge is used as a function that is applied on the feature space to create sets of features, where each set includes one or more features, i.e., one or more genes in the gene selection problem.

The general G-S-M technique was developed by [Yousef et al. \(2019\)](#) and was embedded in different computational tools, such as SVM-RCE-R ([Yousef et al., 2020](#)), miRcorrNet ([Yousef et al., 2021](#)),

maTE ([Yousef et al., 2019](#)), CogNet ([Yousef et al., 2021](#)), SVM-RCE-R-OPT ([Yousef et al., 2021](#)), Integrating GO-based Grouping and Ranking ([Yousef et al., 2021](#)), PriPath ([Yousef et al., 2023](#)), miRdisNET ([Jabeer et al., 2023](#)), GediNET ([Qumsiyeh et al., 2022](#)), miRModuleNet ([Yousef et al., 2022](#)), AMP-GSM ([Söylemez et al., 2023](#)), and TextNetTopics ([Yousef and Voskergian, 2022](#)). The main idea and most of the relevant tools are reviewed in [Yousef et al. \(2021\)](#).

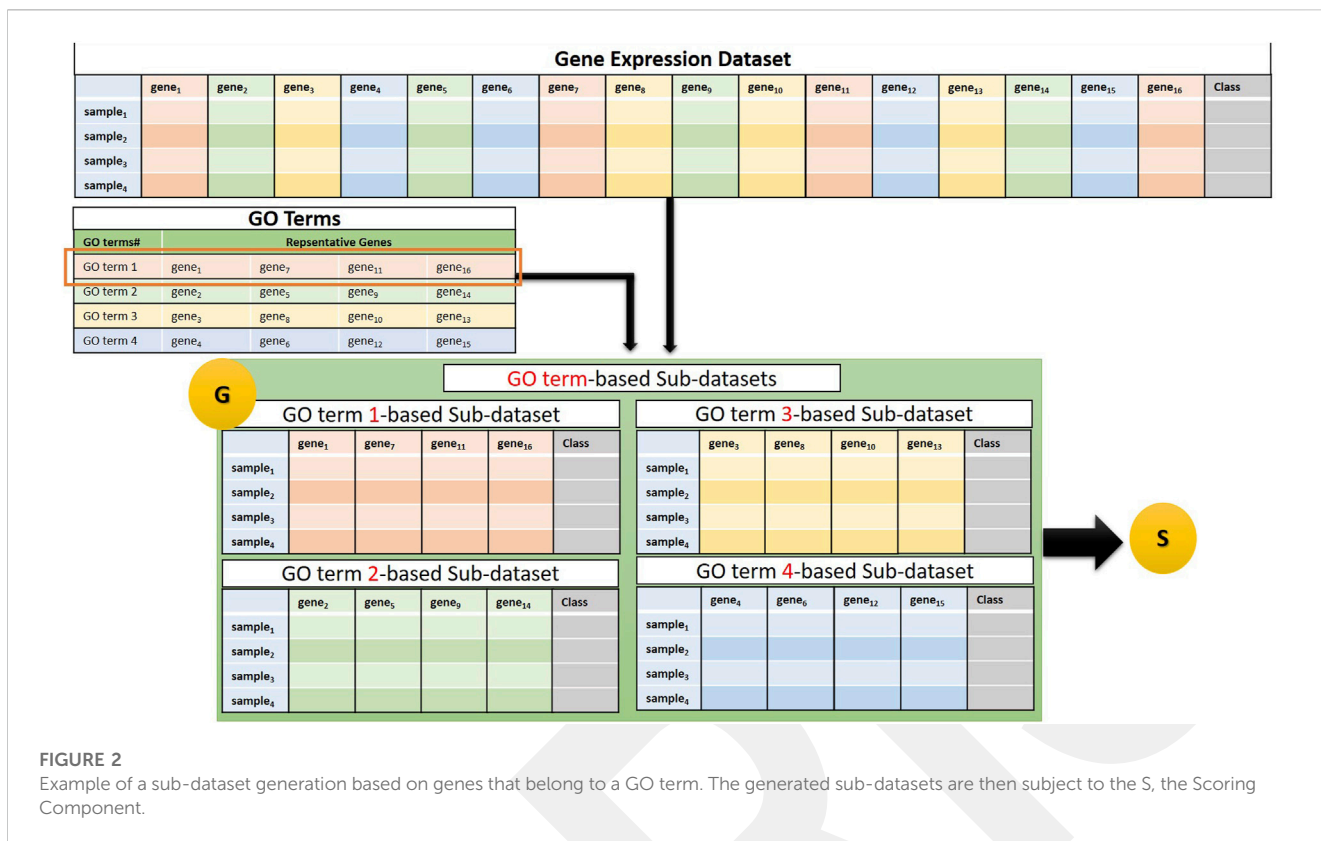
2.3.1 General methodology of GeNetOntology

Here, we develop a novel approach named GeNetOntology. The GeNetOntology consists of three main components (illustrated in [Figure 1](#)):

1. G Component: to generate sub-datasets for each GO term group.
2. S Component: to score GO terms.
3. M Component: to train the classifier (Random Forest) to build the model.

The main purpose of the GeNetOntology is to find significant GO terms (scored in S Component) to be used for training the classifier (M component). In order to evaluate a set of features, for each GO term, a sub-dataset is created by only including the expression values of the genes which are associated with that particular GO term. The pseudo-code of GeNetOntology is presented in [Supplementary Table S1](#), and the algorithm is explained more in detail in the following sections.

The gene expression dataset is represented by C , which consists of two parts, C_{train} and C_{test} . Although C_{train} has been utilized for scoring the GO terms and training the classifier to create a model, C_{test} has been used to test and report the final performance.



2.3.2 Component G: generating sub-datasets

The G component of GeNetOntology creates sub-datasets for each GO term. Several genes are annotated with a specific GO term. For each GO term, the G component extracts a sub-dataset from the gene expression dataset. In other words, each sub-dataset includes 1) expression values just for the genes that are annotated with that particular GO term and 2) class labels (positive or negative) of the samples. We would like to note that each sub-dataset includes the same number of samples, but different numbers of features depending on the number of genes that are annotated with that particular GO term. Figure 2 presents a flowchart for sub-dataset generation based on the genes that are annotated with a specific GO term. Each sub-dataset is named as $GO_{term(i)}_{sub_dataset}$ where i starts from 1 and increases until k . Here, k refers to the number of GO terms exported.

The G component involves two tables:

- (1) The GO Terms table: Each GO term is associated with a set of genes (shown as a row of the GO Terms table in Figure 2).
- (2) Gene expression dataset: The flowchart in Figure 2 illustrates the generation of four sub-datasets, each corresponding to a specific GO term. Next, component S, which is shown as “S” in the last part of Figure 2, performs the scoring step by taking these sub-datasets as an input.

2.3.3 Component S: scoring GO terms

S, as a second component, takes the created sub-datasets as an input from component G and operates an ML algorithm (Random Forest) with an internal MCCV repeated r times (as shown in

Supplementary Figure S3). It has been performed on each sub-dataset to give a score for its related GO term. The scoring is evaluated by testing the capability of each GO term in terms of its classification performance. In other words, the score implies the accuracy of the classification by only using the gene expression values of the genes that are annotated with that specific GO term. In the S component, the classification accuracy is averaged over the r iterations of the MCCV. The mean accuracy value is used as the final score of the particular GO term. The S component ranks all GO terms according to their scores. The top-scoring GO terms are used in the next component to train the model.

Table 3 presents an example of the S output where for each GO term, we assign a score that is calculated as the mean accuracy.

2.3.4 Component M: building the model

The M component trains the classifier and creates the model. The main principle of Component M is illustrated in Supplementary Figure S4. This component trains a classifier (a Random Forest model) using the gene expression values of the genes that are annotated with the top-scoring GO terms. We repeat this procedure in a cumulative manner. In the first iteration, a Random Forest model is built by operating only on the genes that are annotated with the highest scoring GO term. In the second iteration, the M component takes the second highest scoring GO term and merges the genes annotated with this GO term with the gene set annotated with the highest scoring GO term which is identified in the first iteration. In this way, it forms a new sub-dataset that will be subject to training and testing the model. This operation continues in the same way until all GO

TABLE 3 Example of grouping Gene Ontology (GO) terms and their computed scores. Calculated for the GO BP category in the GDS1962 dataset.

GO term#	Score
GOBP_AMEBOIDAL_TYPE_CELL_MIGRATION	1
GOBP_APOPTOTIC_PROCESS_INVOLVED_IN_BLOOD_VESSEL_MORPHOGENESIS	1
GOBP_CANONICAL_WNT_SIGNALING_PATHWAY	0.98
GOBP_CELLULAR_RESPONSE_TO_EXTERNAL_STIMULUS	0.98
GOBP_CELLULAR_RESPONSE_TO_LOW_DENSITY_LIPOPROTEIN_PARTICLE_STIMULUS	0.97
GOBP_CELLULAR_RESPONSE_TO_NUTRIENT	0.96
GOBP_CELLULAR_RESPONSE_TO_PEPTIDE_HORMONE_STIMULUS	0.96
GOBP_CHOLESTEROL_STORAGE	0.95
GOBP_FATTY_ACID_BIOSYNTHETIC_PROCESS	0.93
GOBP_FOAM_CELL_DIFFERENTIATION	0.92

TABLE 4 Sample output of GeNetOntology. Averages over 10-fold MCCV are presented for different performance metrics. Obtained cumulatively for the top 10 ranked Gene Ontology BP groups using the GDS1962 dataset.

#Groups	#Genes	Accuracy	Sensitivity	Specificity	AUC
10	133.9	0.94	0.96	0.9	1
9	122.9	0.95	0.96	0.95	0.995
8	114.9	0.94	0.96	0.9	0.985
7	107.8	0.95	0.96	0.95	0.98
6	93.8	0.94	0.98	0.85	0.99
5	88.6	0.95	0.98	0.9	1
4	74.7	0.98	0.98	1	1
3	62.4	0.95	0.96	0.95	1
2	48.2	0.94	0.98	0.85	0.98
1	31.9	0.91	0.96	0.8	0.97

terms are processed. Hence, we calculate the cumulative performance of the model. Through this approach, it becomes possible to plot the performance results over different feature sets (i.e., the highest scoring GO term, the top two highest scoring GO terms, until the top 10 highest scoring GO terms). In this way, one can discover the best feature set which is defined as the combination of the genes annotated with the top scoring GO terms.

A sample output of the M component is shown in Table 4. GeNetOntology presents the average performance metrics of the

classification using the genes that are associated with the top 10 scoring GO terms, averaged over 10-fold MCCV.

2.4 Design and implementation of GeNetOntology

Figure 1 and Supplementary Figure S5 demonstrate the workflow of GeNetOntology. Two input files are required by GeNetOntology: gene expression data and the biological knowledge file (GO terms and a set of genes annotated with each GO term). In this study, GO terms are used for grouping the genes and for grouping their associated gene expression values. In this way, we generate different sets of features using biological knowledge, and then we evaluate the contribution of each feature set to the classification task.

The gene expression data are separated randomly into training and testing parts. The testing data are only used to evaluate the model's performance. The G-S-M components are followed sequentially. The whole workflow is repeated N times, where we set N to 10 in this study. This repetitive part is shown in the MCCV box in Figure 1 and in Supplementary Figure S5.

At the initial steps of GeNetOntology, in order to filter the least significant genes on the training set, a *t*-test is performed. Additionally, to be able to have the same representation of the genes in the test dataset, only the selected genes within the training set are used in the test set. The test set is only used at the M component for testing the model. The MCCV loop creates N lists of different outputs, such as lists of performance tables, lists of ranked GO terms, and ranked genes. The average is calculated over all lists of performance tables to create a final performance table that also contains the standard deviation. The robust rank aggregation approach was applied to the other lists (i.e., ranked GO terms and ranked genes) to aggregate them into a final list, as shown in Table 5. The ranked GO terms and the genes that are annotated with these terms are shown in the final list. All those final lists or tables are visualized in the output panel of GeNetOntology, as shown in Figure 1 and Supplementary Figure S5.

GeNetOntology has been implemented in the free and open-source KNIME Analytics Platform, which is a data analysis, reporting, and integration tool under the General Public License (GNU) (Berthold et al., 2009). KNIME is able to utilize scripts in both R and Python. In the KNIME workflow, there are several nodes with their own functions. Nodes have been collected under meta-nodes that fulfill a specific task. The GeNetOntology KNIME workflow is available publicly in <https://github.com/malikyousef/GeNetOntology.git>.

2.5 Model performance evaluation of GeNetOntology

We have evaluated a set of statistical measures such as specificity, sensitivity, and accuracy for each model to score model efficiency. The following formulations were used to calculate the statistics:

TABLE 5 Example of the robust rank aggregation output for the GDS1962 dataset where the ranked GO BP terms (groups) and their associated genes are shown.

GO groups	p-value (Score)	#Genes	Genes
GOBP_ARTERY_MORPHOGENESIS	1.30E-07	17	ADAMTS9, APOE, PRRX1, EFEMP2, VEGFA...
GOBP_CHROMATIN_SILENCING	4.44853E-07	9	SMCHD1, HMGB1, EZH2...
GOBP_CAMERA_TYPE_EYE_DEVELOPMENT	5.61732E-07	49	RDH10, MEGF11, ATOH7...
GOBP_GLAND_DEVELOPMENT	7.07998E-07	85	PRMT5, IQGAP3, MSN...
GOBP_RESPONSE_TO_HORMONE	1.65122E-06	123	IDH1, ADAM9, GPR173...
GOBP_NEGATIVE_REGULATION_OF_EPITHELIAL_CELL_MIGRATION	1.8769E-06	14	HMGB1, SP100, APOE...

- Sensitivity (Recall) = True Positive/(True Positive + False Negative)
- Specificity = True Negative/(True Negative + False Positive)
- Accuracy = (True Positive + True Negative)/#All examples,

In addition, the area under the receiver operating characteristic (ROC) curve (AUC) is used to approximate the probability of a classifier which would score a randomly selected positive instance higher than a randomly selected negative instance. All reported performance measures indicate the average of 10-fold MCCV. We have performed an under-sampling approach to deal with the imbalanced dataset problem. This approach decreases the number of samples in the majority class to the number of samples in the minority class. In that way, we can reduce the bias in the size distribution of datasets and overcome the imbalanced class distribution problem. The under-sampling ratio is chosen as 1:2. Graphs and figures have been generated using the software GraphPad Prism 8.

2.6 Protein–protein interaction (PPI) network analysis

Network analysis was performed using Cytoscape (Franz et al., 2016). Using Cytoscape, we have visualized the PPI networks of the genes which are annotated with the most significant GO term. Cytoscape imports the human PPI network from the STRING database. The betweenness centrality of nodes was calculated in Cytoscape using the built-in NetworkAnalyzer (Cytoscape App Store - NetworkAnalyzer, 2021). For each GO category (BP, CC, and MF), genes that are annotated with the top 10 scoring GO terms were selected. For each protein, we have computed the betweenness centrality which indicates the amount of control that this node exerts over the interactions of other nodes in the network (Yoon et al., 2006). The color and size of the node are used to represent the betweenness centrality. Bigger and darker colored nodes (proteins) in the PPI network have higher betweenness centrality.

3 Results

3.1 Model performance evaluation of GeNetOntology

GeNetOntology is tested on 11 different gene expression datasets, where the characteristics of the datasets are presented in

Table 1. For each dataset, for different numbers of feature sets, the accuracy, sensitivity, specificity, and AUC values have been calculated as the mean of the values obtained in 10 iterations of the cross-validation procedure. For each feature set, GeNetOntology reports the number of features (i.e., genes) included in the set (i.e., GO term). In addition, the average gene number over 10 iterations is reported. Table 4 shows the performance metrics of GeNetOntology for the GDS1962 dataset for the top 10 scoring GO terms. For example, there are 31.9 genes on average as shown in the # of the Genes column of the last row (top scoring GO term) in Table 4. There are 48.2 genes on average as shown in the # of the Genes column of the 2nd last row (top two scoring GO terms cumulatively). In other words, the model that is generated via only using the gene expression values of 48.2 genes can successfully predict glioma patients with 0.98 AUC score.

In our experiments on various gene expression datasets, three different GO categories, i.e., BP, CC, MF, and all GO terms are utilized. Table 6 summarizes the performance metrics obtained for 11 different gene expression datasets using only the top two scoring GO terms. AUC, accuracy, and sensitivity values for GDS2519 and GDS4206 are not as high as in other datasets. However, AUC, accuracy, and sensitivity values for GDS 1962, GDS3837, GDS4516_718, and GDS5499 are quite high.

Additionally, to be able to compare our approach with other solutions, we have performed additional experiments. In this respect, we have applied different traditional feature selection methods such as eXtreme Gradient Boosting (XGB) (Athanasios et al., 2020; Li et al., 2020), Information Gain (IG) (Lei, 2012), Select K Best (SKB) (Pedregosa et al., 2011), and Fast Correlation-Based Filter (FCBF) (Senliol et al., 2008); and different classifiers such as Adaboost, Decision Tree (DT), LogitBoost, Random Forest (RF), SVM_opt, Stack_Logitboost_Kmeans, and Stack_SVM_Kmeans on the same 11 gene expression datasets using 10-fold MCCV. For each dataset, we evaluate the performance of each classifier and each feature selection method with the same number of features used by GeNetOntology. For most of the tested classifiers, the AUC of the XGB FS method (Supplementary Table S2) showed higher performance than the IG FS method (Supplementary Table S3), SKB (Supplementary Table S4), and FCBF FS method (Supplementary Table S5). For each FS method, we have plotted the average AUC values over seven different classifiers for 11 different datasets. As shown in Supplementary Figure S6, GeNetOntology generates similar AUC values compared with

TABLE 6 Performance results table including accuracy, sensitivity, specificity, and AUC value of GeNetOntology for 11 different datasets.

Accession Numbers in GEO	#Genes	Accuracy	Sensitivity	Specificity	AUC
GDS1962	48.2	0.94	0.98	0.85	0.98
GDS2519	170.4	0.5	0.6	0.41	0.58
GDS2545	110.5	0.72	0.72	0.7375	0.78
GDS2547	81.7	0.75	0.725	0.7875	0.83
GDS2771	68.4	0.66	0.73	0.6	0.68
GDS3257	20.6	1	1	1	1
GDS3268	90.3	0.66	0.62	0.71	0.73
GDS3837	64.6	0.98	0.96	1	0.99
GDS4206	46.1	0.63	0.25	0.8	0.55
GDS4516_718	50.6	1	1	1	1
GDS5499	52.1	0.9	0.94	0.8	0.95

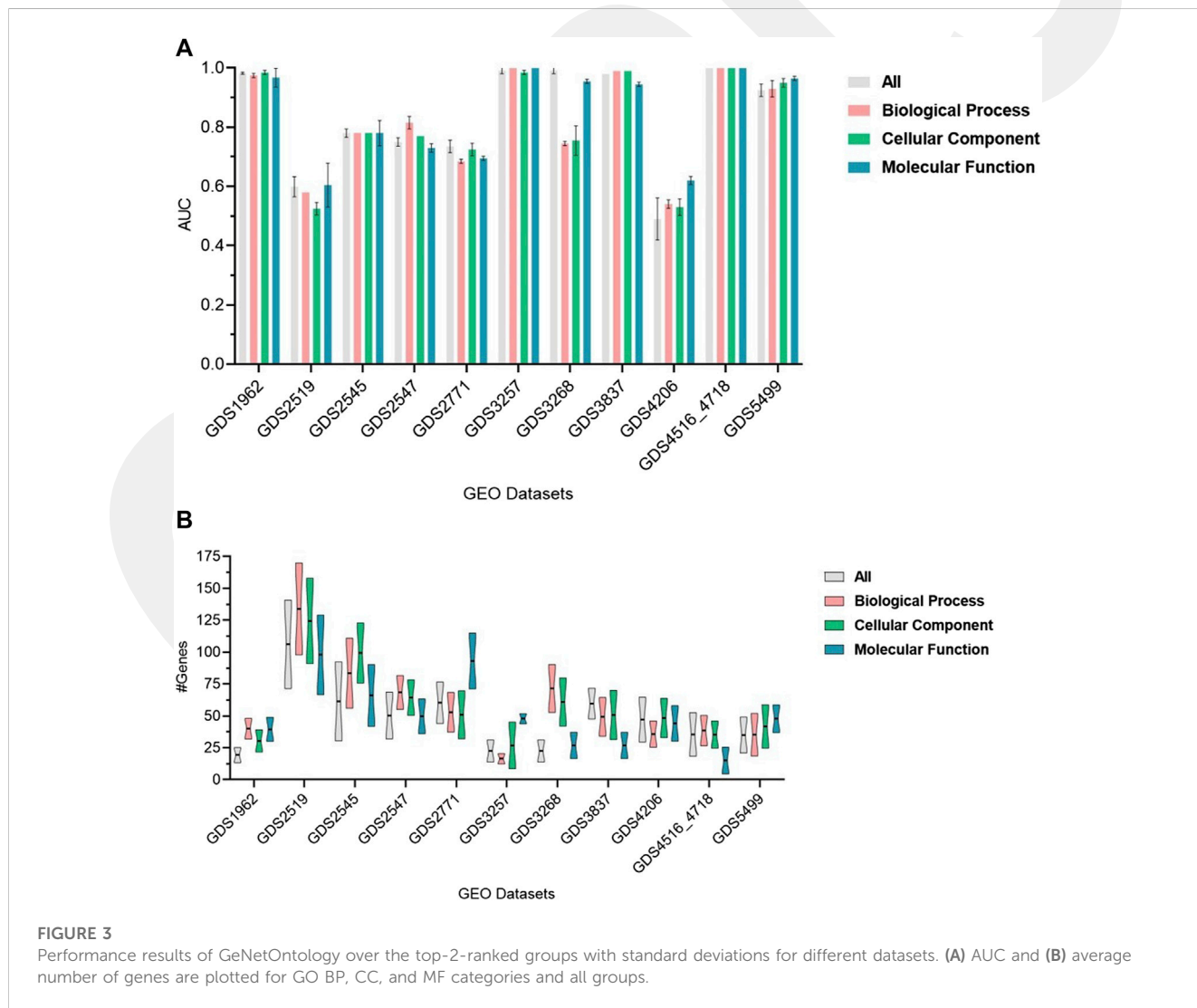
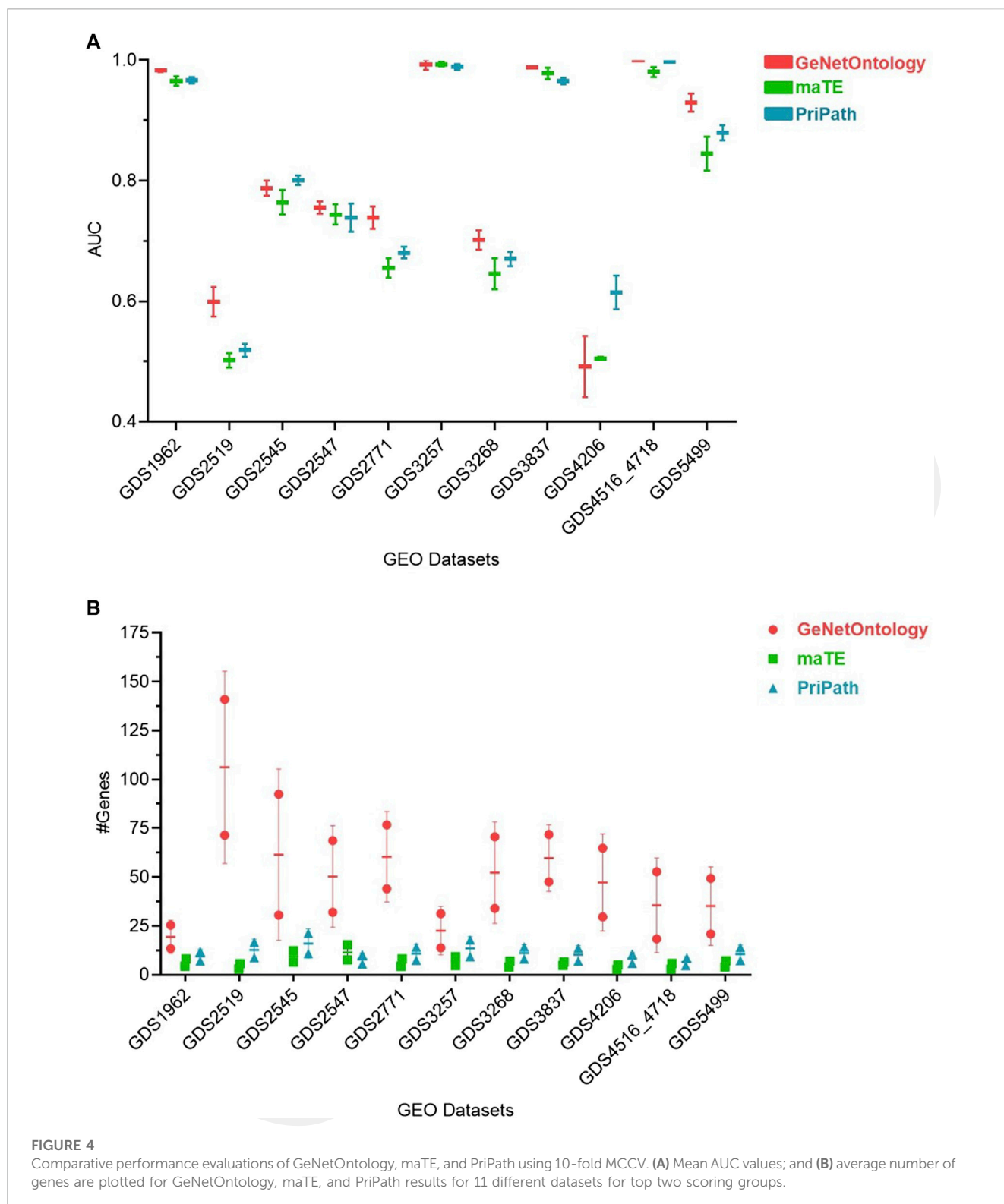


FIGURE 3 Performance results of GeNetOntology over the top-2-ranked groups with standard deviations for different datasets. (A) AUC and (B) average number of genes are plotted for GO BP, CC, and MF categories and all groups.



other methods, averaged over different datasets. We would like to emphasize that the aim of GeNetOntology is not to compete with other feature selection (FS) approaches. Our aim is to select significant ontology terms that have biological meaning. Even if the performance of GeNetOntology is similar with that of other FS methods or even slightly less than that of other FS methods, the

contribution of the tool is to find the most informative GO terms that can help the researchers understand the biological background of the disease under study.

The performance of GeNetOntology over 11 datasets by using GO BP, CC, and MF categories and all GO terms are summarized in Figure 3. AUC is considered the performance metric, and all values

in Figure 3 are the mean of 10-MCCV iterations, shown together with standard deviations. Figure 3A implies that except for the GDS3268, GDS2519, and GDS4206 datasets, the choice of GO category does not affect the performance in terms of AUC. Figure 3B implies that, except for the GDS2771, GDS3257, and GDS3268 datasets, the choice of the GO category does not affect the average number of genes in the top two selected GO terms by GeNetOntology.

3.2 Comparative performance evaluation of GeNetOntology with other GSM-based tools

PriPath (Yousef et al., 2023) and maTE (Yousef et al., 2019) are two other G-S-M-based tools that incorporate biological domain knowledge. Although PriPath utilizes KEGG pathways as the biological knowledge, maTE uses miRNA information as pre-existing biological knowledge. maTE integrates information about miRNA target genes with gene expression data, and multiple high-scoring miRNAs are utilized while building the final classifier. PriPath tries to detect dysregulated pathways by using KEGG pathways as the grouping information and insert this information into an ML algorithm for selecting the most significant KEGG pathways in the gene expression dataset. We have compared GeNetOntology results with PriPath and maTE results (Figure 4). Although maTE, PriPath, and GeNetOntology generate distinct output tables, they all produce a table which demonstrates the tool's performance. AUC performance metrics of GeNetOntology, PriPath, and maTE have been comparatively evaluated for 11 different datasets. We have considered the AUC values of the top two scoring sets for each tool by applying 10-fold MCCV. The mean AUC values of the three tools for the 11 datasets are represented in Figure 4A. One can deduce from Figure 4A that for GDS2519, GDS2771, and GDS5499 datasets, GeNetOntology performed higher than maTE and PriPath. For the remaining datasets, the AUC values are comparable. The mean number of genes utilized by the tools is also plotted in Figure 4B.

3.3 Biological validation of the GeNetOntology: analysis of the diagnostic model developed for the glioma dataset

In the previous section, we assessed the performance of GeNetOntology from a computational point of view using different computational performance evaluation metrics. In this section, we evaluate GeNetOntology findings from a biological point of view using the glioma dataset (GDS 1962).

3.3.1 Correlation between top scored gene ontology terms for glioma dataset

One of the outputs of the GeNetOntology tool is a list of ranked GO terms for the disease under study. The robust rank aggregation method of GeNetOntology calculates a p -value for each GO term, which shows the significance of the GO term in distinguishing the cases from controls for the disease under investigation. At the final step, the GO terms are ranked based on this p -value. In Figure 5, we plotted the top 10 important GO terms for the GDS1962 dataset

associated with glioma. Although the robust rank aggregation p values are converted to $-\log_{10}$ scale and shown in the x -axis, the GO terms are represented in the y -axis, and some examples of the genes annotated with the specified GO terms are represented on the bars. GO terms are ranked separately for BP, CC, and MF categories and shown in Figures 5A–C, respectively.

As depicted in Figure 5A, for the glioma dataset (GDS 1962), artery morphogenesis is the top ranked GO term in the BP category. Chromatin silencing, camera-type eye development, gland development, response to hormone, negative regulation of epithelial cell migration, animal organ morphogenesis, and cell morphogenesis involved in differentiation have similar p values; and lipid catabolic process and cellular lipid metabolic process have the lowest $-\log_{10}$ p -value among the top 10 ranked GO BP terms. As illustrated in Figure 5B, for the same dataset, the cell cortex has the highest score in the top 10 ranked GO CC category. Envelope, actin filament, blood microparticle, protein DNA complex, inclusion body, vacuole, condensed chromosome centromeric region, coated vesicle, and nuclear protein-containing complex are the other top ranked terms in the GDS1962 gene expression dataset. As shown in Figure 5C, primary active transmembrane transporter activity and carbohydrate binding GO terms have the highest importance for the GDS1962 gene expression dataset in the MF category. Hormone binding, lipid binding, hydrolase activity acting on ester bonds, molecular adapter activity, isomerase activity, carboxylic ester hydrolase activity, protein kinase c binding, and active transmembrane transporter activity are the other GO MF terms selected among the top 10 for the glioma dataset. These results guided us to apply further analysis to investigate the possible relationships between these GO terms.

It has been mentioned in literature that GO terms facilitate our understanding of the disease development and progression at the molecular level (Denny et al., 2018). Studying the associations between GO terms may further help us enlighten the relations of the GO terms with the disease mechanisms. In Figures 6A–C, for the GDS1962 glioma dataset, we present the pairwise correlations among the top 10 scoring gene ontology terms for BP, CC, and MF categories, respectively. In the BP category, organ morphogenesis and gland development terms have moderate correlations in terms of their shared genes) (as shown in a heatmap in Figure 6A). In the GO CC category, no significant correlation is observed between the top 10 identified GO terms (as depicted in Figure 6B). In the GO MF category, primary active transmembrane transporter activity and active transmembrane transporter activity have a moderate relationship (as displayed in Figure 6C). Figure 6 implies that the set of genes in each one of the top 10 scoring GO terms are nearly unique and there is minimum redundancy between the genes that are annotated with the selected GO terms. Hence, each selected GO term contributes to the classification. In other words, there are not so many overlapping genes in the top 10 selected GO terms. This finding is independent from the BP, CC, and MF categories, as shown with low or medium pairwise correlations in Figures 6A–C, respectively. Figure 6D shows the axis names of GO terms for BP, CC, and MF.

3.3.2 PPI network analysis of the genes included in the top 10 scoring GO terms

We have collected the genes that are annotated with the top 10 scoring GO terms identified by GeNetOntology for the

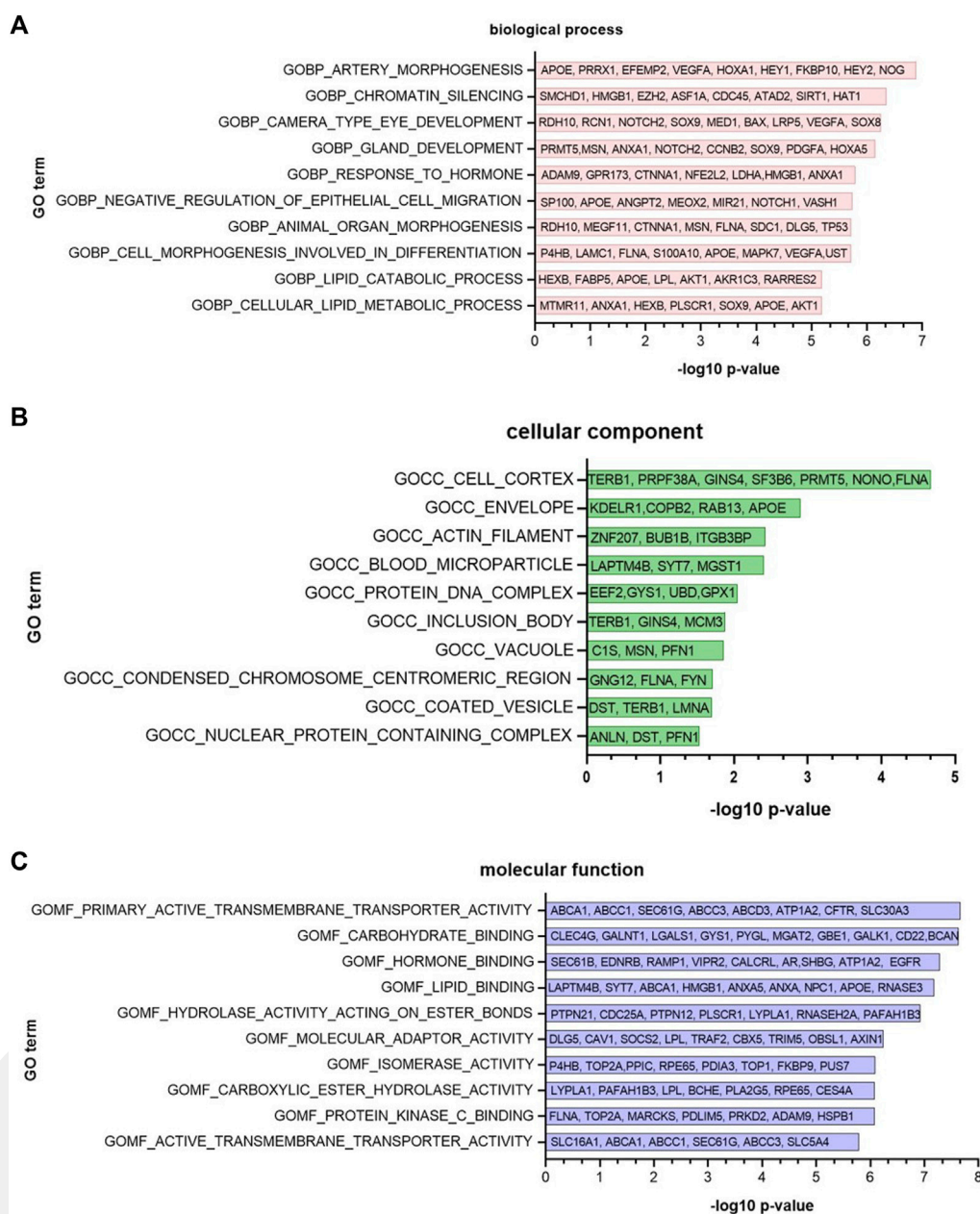


FIGURE 5 Top 10 important Gene Ontology terms in (A) BP, (B) CC, and (C) MF categories, identified by GeNetOntology for the GDS1962 glioma dataset. Although $-\log_{10} p$ -values are represented on the x -axis, GO terms are represented on the y -axis, and some examples of the genes that belong to the associated GO terms are represented along the bars.

GDS1962 glioma dataset. Then, we analyzed the topological properties of these proteins within the human PPI network. To this end, Figure 7 represents the PPI network of the genes annotated with the top 10 scoring GO terms detected by GeNetOntology for the GDS1962 dataset. We repeated this procedure separately for GO BP, CC, and MF categories and illustrated in Figures 7A–C, respectively. The PPI networks including 460 genes with 2,809 interactions for the BP category, 236 genes with 990 interactions for the MF category, and 284 genes with 777 interactions for the CC category are represented in Figures 7A–C, respectively. For each protein, we have computed the

betweenness centrality which indicates the amount of control that this node exerts over the interactions of other nodes in the network (Yoon et al., 2006). Bigger and darker colored nodes in Figure 7 represent the proteins with higher betweenness centrality. One can easily observe from Figure 7A that TP53, AKT1, VEGFA, IDH1, MYC, APOE, NOTCH1, SOX2, FGF2, CAV1, and CCND1 have high betweenness centrality in the BP category for the glioma dataset. It implies that for glioma, these genes play key roles between other nodes (proteins) as connective proteins. In terms of the GO CC category, APOE, CCNB1, LMNA, RHOC, GNS, ANLN, CAV1, SIRT1, CTSO, and LMNA proteins are found

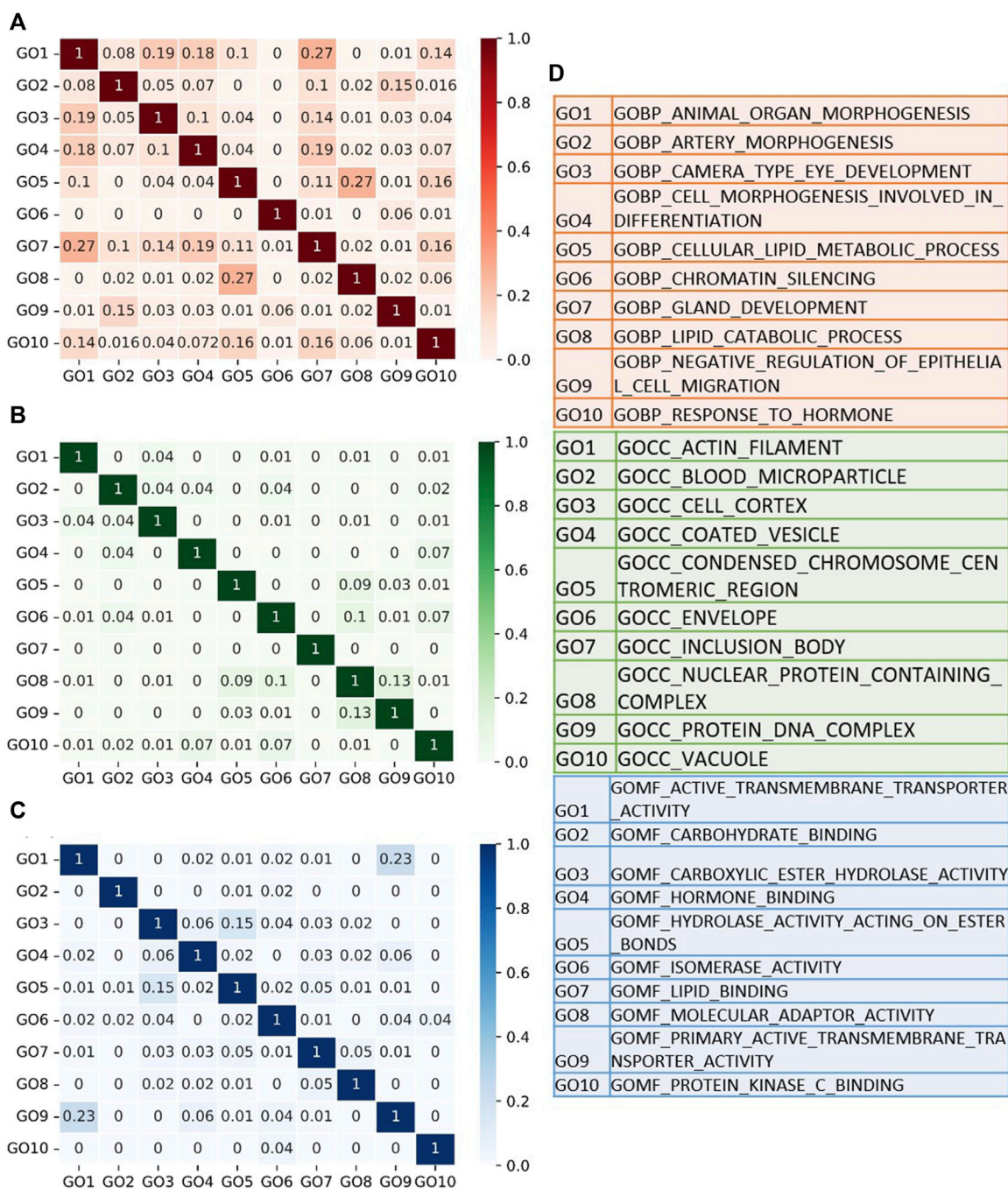


FIGURE 6 Correlations among the top 10 Gene Ontology terms identified by GeNetOntology for the GDS1962 glioma dataset. The pairwise correlations are calculated based on the number of shared genes within GO term pairs. Heatmaps for pairwise correlations of the GO terms are visualized in (A) BP; (B) CC; (C) and MF categories. (D) Full names of the GO terms.

to have high betweenness centrality in the analysis of GeNetOntology on the glioma dataset. One can imply from Figure 7B that these core proteins may play an important role for glioma. As visualized in Figure 7C, AKT1, MAPT, MYC, APOE, JUN, CAV1, and EGFR are hub proteins in the PPI network that are generated according to the top 10 scoring GO terms identified by GeNetOntology on the MF category for the glioma dataset.

3.3.3 Frequency and word cloud analysis of the genes associated with top 10 gene ontology terms in the glioma dataset

Proteins can play different roles in the organism, and hence they may be included in different GO terms. We have collected the genes that are annotated with the top 10 scoring GO terms identified by GeNetOntology for the GDS1962 glioma dataset. This time, instead of having three different gene sets for BP, CC, and MF categories, we

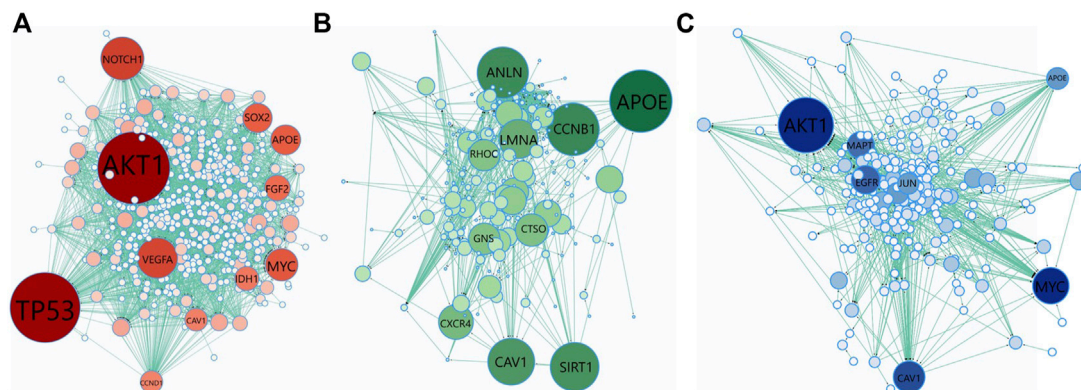


FIGURE 7

Protein-protein interaction (PPI) network of the genes included in the top-10 scoring GO terms detected by GeNetOntology for the GDS1962 dataset. Genes are collected from the top 10 GO terms in (A) BP; (B) CC; and (C) MF categories. Color and node size denote betweenness centrality. Bigger and darker colored nodes have higher betweenness centrality in the PPI network.

merged these lists and obtained one gene set. In other words, if a gene is annotated with at least one of the top 10 scoring GO BP, CC, or MF terms for the glioma dataset, we include this gene into our final list. For the identified genes, we also keep track of the frequencies or how many times a gene is observed in any one of the top 10 scoring GO BP, CC, or MF terms for the glioma dataset (as shown partially in Figure 8A). One can observe from Figure 8A that APOE, PSEN1, RPE65, PTEN, and SRC have the highest frequencies. These five genes are annotated with eight GO terms among the top 10 scoring GO terms in any one of the BP, CC, and MF categories. Here, the top 10 scoring GO terms refer to the terms that are identified by GeNetOntology while analyzing the GDS1962 glioma dataset. We have also visualized the frequency of the genes from the top 30 scoring GO terms as a word cloud. To this end, Figure 8B presents the top 100 frequent genes from the top 30 scoring GO terms, where the top 10 scoring GO terms are identified by GeNetOntology for each one of the BP, CC, and MF categories for the glioma dataset.

4 Discussion

4.1 Computational performance evaluation of GeNetOntology

GeNetOntology is a novel and highly effective approach that predicts disease-causing genes by modeling and analyzing gene expression data with GO terms. In GeNetOntology, we have implemented an ML algorithm to select the most significant GO terms. Yet, when more terms are included, we did not obtain any statistically significant improvement in the performance metrics. Therefore, it is possible to construct a model with less number of genes, which facilitates the interpretation of the generated model. The GeNetOntology was tested on 11 different gene expression datasets, including cancer, hypertension, and other diseases. For the present analysis, we have tested GeNetOntology using i) all GO terms as all terms; ii) terms in BP; iii) in CC; and iv) in MF categories. Under these four settings, the performance results (in terms of AUC metric) using the top two scoring GO terms from different categories for different

datasets are shown in Figure 3. In Figure 3, we present the mean of the AUC values for 10-fold cross-validations for all, BP, CC, and MF categories and the mean number of genes for each one of the 11 datasets.

As shown in Figure 3, GeNetOntology performs well on all datasets, except GDS2519 and GDS4206. When we analyzed different performance metrics of GeNetOntology over the top two ranked GO terms for different datasets, we noticed one more time that GeNetOntology performs well on all datasets, except GDS2519 and GDS4206, as shown in Table 6. In addition, when these datasets are analyzed with other tools such as maTE and PriPath, it has been observed that other tools also generated poor performance on these two datasets (as shown in Figure 4). As it can be seen in Supplementary Tables S2-S5, the AUC scores of all tested classifiers (Adaboost, DT, LogitBoost, RF, SMV_opt, Stack_Logitboost_Kmeans, and Stack_SVM_Kmeans) and all tested feature selection methods (XGB, IG, SKB, and FCBF) are also low for GDS2519 and GDS4206 datasets.

We have also tested for the effect of cumulatively adding top scoring GO terms (increasing the gene number) in terms of the performance of GeNetOntology. In Table 4, we present the performance metrics of GeNetOntology averaged over 10-fold MCCV iterations for the aggregated top-10 scoring GO terms for the GDS1962 dataset. For example, GeNetOntology has an AUC value of 0.97 when on average 31.9 genes are used, as shown in the last row of Table 4 (results of the top scoring GO term). On the other hand, the AUC value of GeNetOntology becomes 1 when 133.9 genes from the top-10 scoring GO terms are cumulatively used (as shown in the first row of Table 4). Instead of checking the gene expression values of 134 genes, one can prefer to check the expression values for only 32 genes to predict whether the sample has glioma while sacrificing 3% of AUC. To put it another way, the model that is generated via only using the gene expression values of 32 genes can successfully predict glioma patients with an AUC score of 0.97, which is quite a satisfying result.

Additionally, GeNetOntology is comparatively evaluated with other G-S-M-based tools. For 11 different gene expression datasets, the mean AUC values of 10-fold cross-validations for GeNetOntology, maTE, and PriPath were compared using the genes from the top two

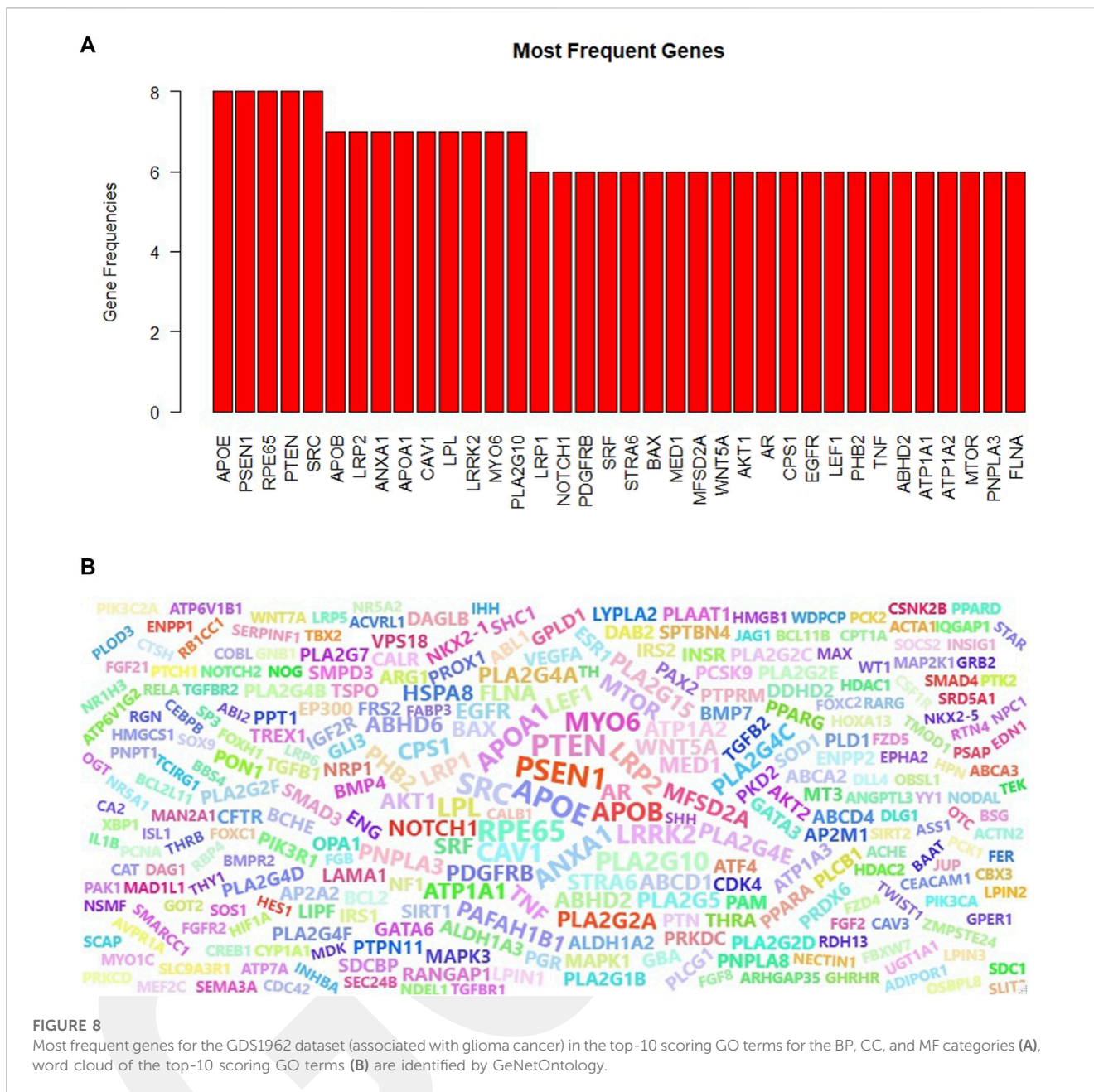


FIGURE 8 Most frequent genes for the GDS1962 dataset (associated with glioma cancer) in the top-10 scoring GO terms for the BP, CC, and MF categories (A), word cloud of the top-10 scoring GO terms (B) are identified by GeNetOntology.

scoring feature sets. One can obtain from Figure 4A that for GDS2519, GDS2771, and GDS5499 datasets, GeNetOntology performed better than maTE and PriPath. For the remaining datasets, the AUC values were comparable. On the other hand, the outcome and contribution of each tool is different because each of these tools aims to identify important feature sets via using a different biological knowledge. For example, while defining important feature sets, PriPath makes use of KEGG pathway information, and the final output of PriPath is the top scoring pathways and their associated genes. In maTE, miRNA–target gene information is utilized, and the final outcome is top scoring miRNAs and their associated genes. However, GeNetOntology exploits GO as biological knowledge; and as an output, it identifies top scoring GO terms and their associated genes that can best distinguish patients from healthy samples for the specific disease under study.

4.2 Biological evaluation of GeNetOntology findings

4.2.1 Top scoring gene ontology terms for the glioma dataset

As the performance metrics shown in Table 6, Figure 3A, Figure 4A implies, GeNetOntology is capable of discovering significant gene ontology terms that can perform as a marker for classification. In the previous section, GeNetOntology is comparatively evaluated with other G-S-M-based tools. Here, we evaluated whether the top scoring GO terms identified by GeNetOntology are supported by previous experimental research.

For the glioma dataset, the top 10 significant GO BP terms include morphogenesis, development, and differentiation (Figures 6A, D). In the literature, it has been shown that morphogenesis is highly correlated

with cancer invasion to other tissues. Lipids are a complex group of biomolecules, and they form the main structure of biological membranes, an energy source, and also act as signaling molecules (Snaebjornsson et al., 2020). Lipid metabolism has an impact on cancer formation and promotion since lipid metabolism, carcinogenesis, and cancer metastasis are related with abnormal levels of lipids (Jiang et al., 2020). Chromatin is a critical and dynamic major regulator of transcription. Studies showed that deregulation of chromatin guides gene activation alteration and/or improper gene silencing and is able to promote oncogenesis by altering chromatin structures (Nair and Kumar, 2012). Numerous studies have demonstrated that epigenetic gene silencing is a critical mechanism for the loss of gene function in many cancers (Baylin and Ohm, 2006). Several GO BP terms such as tissue formation, embryological development, inflammation, immune defense, and cancer progression, are highly related with cell migration (Pijuan et al., 2019). Unusual regulation of epithelial cell migration has a significant role in pathological processes such as cancer metastasis and tissue fibrosis development (Magliozzi et al., 2013). Cancer cell invasion mechanisms and metastasis based on cellular motility is a complicated process (Lorente et al., 2014). An essential characteristic of tumor cell invasion and metastasis is the ability of improved motility and migration capability of the tumor cells. It is important to understand the specific mechanisms of cellular motility in cancer development because a loss or deficiency of the mechanisms that regulate cytoskeletal remodeling might result in tumor development and metastasis (Lorente et al., 2014). Therefore, actin filament formation and cell cortex and envelope formation could potentially have a role in cancer formation. These terms are identified by GeNetOntology among top 10 significant GO CC terms, as shown in Figures 6B, D. Blood microparticles originate from the endothelial lining of blood vessels and cellular components of blood. Almost all cells, when they are exposed to several stress conditions such as apoptosis and cellular activation, shed parts of their plasma membranes, and these are called cellular microparticles (MPs) (Shet, 2008). DNA packaging and gene expression regulation are performed by chromatin, which is a protein–DNA complex (de Brot et al., 2018). For the glioma dataset, the top-10 scoring GO MF terms of GeNetOntology include carbohydrate-, lipid-, and hormone-binding molecular functions which regulate several intracellular and extracellular signaling pathways (Figures 6C, D).

4.2.2 Top scoring genes for the glioma dataset

We have collected the genes that are annotated with the top 10 scoring GO terms identified by GeNetOntology for the GDS1962 glioma dataset. As shown in Figures 7A–C, several well-known cancer driver genes are present in the PPI network generated by using the genes from the top scoring GO BP, CC, and MF terms, respectively. TP53 is a well-known oncogene, and it is highly correlated with many cancer subtypes. AKT1 plays a vital role in many signaling pathways and in growth factor-induced neuronal survival during nervous system development (Dudek et al., 1997). VEGFA is a critical modulator of angiogenesis, and it has been shown in the literature that VEGFA expression is high in cancer tissue, and this is correlated with its aggressive characteristics (Sa-nguanraksa and O-charoenrat, 2012). Mutation in the IDH1 gene has been found in many genetic conditions and cancer types, such as acute myeloid leukemia, glioblastoma, and myelodysplastic syndromes (Dang et al., 2016). MYC is a proto-oncogene, and it is overexpressed in several

tumors. It can escape from several tumor-suppressing checkpoint mechanisms such as apoptosis senescence, proliferative arrest, and induces tumorigenesis (Li et al., 2014). In several tumors, APOE overexpression is related to poor prognosis and aggressive biological behaviors (Zhao et al., 2018). NOTCH1 is associated with numerous signaling pathways in tumorigenesis, and it is involved in many types of cancer, including brain tumors, leukemia, breast cancer, and several other cancer types (Gharaibeh et al., 2020). SOX2 is dysregulated during gene amplification and promotes metastasis, drug resistance, and survival. Therefore, its overexpression is associated with a poor survival rate in cancer patients (Zhang et al., 2020). FGF2 gene expression is correlated with several cancer types, including colorectal cancer (Caiado et al., 2020). With their roles in top 10 scoring GO BP terms, these genes are emphasized in our analysis on the glioma dataset, and these genes are shown with bigger and darker nodes in Figure 7A based on their high betweenness centrality.

APOE plays an important role in GO CC as well (Figure 7B). CAV1 functions both as a tumor suppressor and metastasis promoter membrane-associated scaffolding protein, and it has shown that CAV1 is downregulated in human tumors (Díaz et al., 2020). CCND1, known as a proto-oncogene, switches to proximal APA sites in cancer cells and acts as the G1-S phase of the cell cycle regulator (Wang et al., 2018). It has been shown that in lung cancer, breast cancer, cervical cancer, and melanoma and esophageal squamous cell carcinoma, CCNB1 expression is relatively high (Li et al., 2019). LMNA functions as an oncogene in many cancer cell types, especially in hepatocellular carcinoma (Liu et al., 2020). Understanding the role of RhoC-regulated migration processes is crucial to deal with cancer metastasis mechanisms (Lou et al., 2021). In many types of site-specific cancerous tumors, such as bone marrow, brain, breast, colorectal, pancreas, and lung cancer, ANLN is highly expressed (Hall et al., 2005; Olakowski et al., 2009; Uhlén et al., 2015; Tuan and Lee, 2020). Several studies have presented that SIRT1 can function as a tumor promoter or tumor suppressor depending on its targets in specific cancer and signaling pathways (Lin and Fang, 2013). CTSO is a biomarker that can predict which women will emanate the highest benefit from a selective estrogen receptor modulator (SERM) therapy (Brentnall et al., 2016). With their roles in top 10 scoring GO CC terms, these genes are emphasized in our analysis on the glioma dataset, and these genes are shown with bigger and darker nodes in Figure 7B based on their high betweenness centrality.

AKT1, MAPT, MYC, APOE, JUN, CAV1, and EGFR are identified as hub proteins in the PPI network in Figure 7C, which is generated using the proteins associated with the top 10 scoring GO MF terms. The associations of most of these genes with glioma are also reported in literature as follows. Tauopathies, known as neurodegenerative disorders, are characterized by abnormal tau protein deposition in the brain, and MAPT expression is a biomarker for tauopathies and an increased survival rate and low-grade glioma (Zaman et al., 2019). JUN is a proto-oncogene transcription factor and regulates transcription-caused cancer formation (Expression of JUN in cancer - Summary - The Human Protein Atlas, 2021). EGFR is a driver of tumorigenesis and is identified as a biomarker of resistance in tumors, especially in glioblastoma, breast, and lung cancer (Sigismund et al., 2018). In the literature, it has been shown that the APOE gene is related with tumorigenesis and progression, such as cell proliferation, angiogenesis, and metastasis (Zhao et al., 2018; Adaku et al., 2022).

4.2.3 The most frequent genes and word cloud of the genes associated with top 10 GO terms in the glioma dataset

Genes might have several functions in the living organism and act with multiple roles. Therefore, a gene can be annotated with different GO terms. When a gene is mutated, gain or lose function, several biological processes and molecular functions would be affected. To this end, we analyzed the frequency of the genes that are annotated with the top 10 scoring ontology terms. Figure 8A showed that APOE, PSEN1, PTEN, RP65, and SRC genes play a role in eight different top scoring GO terms for the glioma dataset. Although the importance of the APOE gene for glioma is discussed in the aforementioned section, the other frequently observed genes are potentially associated with glioma as follows. PSEN1 gene missense mutation is a well-known cause of a neurological disorder, Alzheimer's disease (Randa, 2019). The RPE65 gene mutation causes several inherited retinal diseases because it provides instructions to make essential proteins for normal vision (Sodi et al., 2021; *RPE65 gene: MedlinePlus Genetics*). PTEN can function as a tumor suppressor in a PI3K signaling pathway, and when the tumor suppressor function of the PTEN enzyme is disrupted by mutations, it causes cells to grow and uncontrolled division and contributes to a cancerous tumor formation (Milella et al., 2015). Several human cancers, such as colorectal, lung, breast, and prostate cancer, have been strongly related to SRC, which promotes maintenance, progression, development, and metastasis of cancers (Wheeler et al., 2009). We extended the most frequently observed gene list and compiled the list of the top 100 genes which are annotated with the top 10 GO terms identified by GeNetOntology for the glioma dataset. The word cloud visualized in Figure 8B shows that this list included several glioma-associated genes, where several of these genes are discussed in the previous section. Taken together with our previous results, GeNetOntology findings imply that the identified genes have a significant impact on disease development and progression for glioma.

5 Conclusion

The current advancements in next-generation sequencing and other high-throughput technologies make it possible to acquire gene expression profiles from tissue samples at quite low expenses. Various gene expression datasets were publicly available right after these technologies were developed, and extracting knowledge from these datasets became a major challenge. In this study, we have introduced a computational tool that uses biological knowledge from GO, which is implemented into the ML algorithm performing gene selection. Our methodology is different from the standard approaches where the analysis is carried out by considering individual genes; however, GeNetOntology has focused on the investigation of the ontology terms to rank and discover the most influential feature sets. Performance evaluations over 11 different datasets showed that the GeNetOntology tool is consistent and robust. We have compared the performance of GeNetOntology with that of PriPath and maTE, which are similar in their merits. The results show that, in most cases, GeNetOntology outperforms maTE and PriPath, depending on the gene expression dataset. We believe that GeNetOntology will assist scientists, medical geneticists, and physicians in studying and analyzing their gene expression datasets and in better understanding disease-related genes and the main mechanisms

behind disease development and progression. GeNetOntology could help researchers define dysregulated genes and gene ontology terms in BP, CC, and MF categories, which can be potentially applied to medical diagnostics. As a future work, we intend to make improvements in our proposed approach in a way that allows us to perform patient stratification based on gene expression, and it allows us to determine druggable targets toward precision medicine.

Data availability statement

The GeNetOntology Knime workflow is freely available: <https://github.com/malikyousef/GeNetOntology.git>. All the datasets used in this study are publicly available at the gene omnibus at NCBI and can be retrieved using the cited accession numbers. GEO accession numbers of datasets (GDS1962, GDS2519, GDS2545, GDS2547, GDS2771, GDS3257, GDS3268, GDS3837, GDS4206, GDS4516_4718, GDS5499) are shown in the second column of Table 1.

Author contributions

NSE performed the experiments, analyzed the data and results, prepared figures and tables and wrote the manuscript. MY conceived the ideas, designed the study, analyzed the results and wrote the manuscript. BB-G designed the study, analyzed the results and wrote the manuscript. All authors read and approved the final manuscript.

Funding

The work of MY has been supported by the Zefat Academic College. The work of BB-G has been supported by the Abdullah Gul University Support Foundation (AGUV).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2023.1139082/full#supplementary-material>

References

- Abdulwahab, H. M., Ajitha, S., and Saif, M. A. N. (2022). Feature selection techniques in the context of big data: taxonomy and analysis. *Appl. Intell.* 52 (12), 13568–13613. doi:10.1007/s10489-021-03118-3
- Adaku, N., Ostendorf, B., and Tavazoie, S. (2022). *Apolipoprotein E2 promotes melanoma growth, metastasis, and protein synthesis via the LRP1 receptor*. doi:10.1101/2022.10.03.510632
- Albattah, W., Khan, R. U., Alsharekh, M. F., and Khasawneh, S. F. (2022). Feature selection techniques for big data Analytics. *Electronics* 11 (19), 3177. doi:10.3390/electronics11193177
- Ang, J. C., Mirzal, A., Haron, H., and Hamed, H. N. A. (2016). Supervised, unsupervised, and semi-supervised feature selection: a review on gene selection. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* 13 (5), 971–989. doi:10.1109/TCBB.2015.2478454
- Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene ontology: tool for the unification of biology. The gene ontology Consortium. *Nat. Genet.* 25 (1), 25–29. doi:10.1038/75556
- Asif, M., Martiniano, H. F. M. C. M., Vicente, A. M., and Couto, F. M. (2018). Identifying disease genes using machine learning and gene functional similarities, assessed through Gene Ontology. *PLOS ONE* 13 (12), e0208626. doi:10.1371/journal.pone.0208626
- Athanasiou, M., Sfrintzeri, K., Zarkogianni, K., Thanopoulou, A., and Nikita, K. S. (2020). 'An explainable XGBoost-based approach towards assessing the risk of cardiovascular disease in patients with Type 2 Diabetes Mellitus', in Proceedings of the 2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE), Cincinnati, OH, USA, October 2020, pp. 859–864. doi:10.1109/BIBE50027.2020.00146
- Aziz, R., Verma, C., and Srivastava, N. (2017). Dimension reduction methods for microarray data: a review. *AIMS Bioeng.* 4 (1), 179–197. doi:10.3934/bioeng.2017.1.179
- Balakrishnan, R., Harris, M. A., Huntley, R., Van Aukun, K., and Cherry, J. M. (2013). 'A guide to best practices for Gene Ontology (GO) manual annotation'. *Database J. Biol. Databases Curation* 2013, bat054. doi:10.1093/database/bat054
- Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., et al. (2012). NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res.* 41 (D1), D991–D995. doi:10.1093/nar/gks1193
- Baylin, S. B., and Ohm, J. E. (2006). Epigenetic gene silencing in cancer – A mechanism for early oncogenic pathway addiction? *Nat. Rev. Cancer* 6 (2), 107–116. doi:10.1038/nrc1799
- Bellazzi, R., and Zupan, B. (2007). Towards knowledge-based gene expression data mining. *J. Biomed. Inf.* 40 (6), 787–802. doi:10.1016/j.jbi.2007.06.005
- Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Kötter, T., Meinl, T., et al. (2009). Knime - the konstantz information miner: version 2.0 and beyond. *ACM SIGKDD Explor. Newsl.* 11 (1), 26–31. doi:10.1145/1656274.1656280
- Bolón-Canedo, V., Sánchez-Maróño, N., and Alonso-Betanzos, A. (2016). Feature selection for high-dimensional data. *Prog. Artif. Intell.* 2 (5), 65–75. doi:10.1007/s13748-015-0080-y
- Brentnall, A. R., Cuzick, J., Byers, H., Segal, C., Reuter, C., Detre, S., et al. (2016). Relationship of ZNF423 and CTSO with breast cancer risk in two randomised tamoxifen prevention trials. *Breast Cancer Res. Treat.* 158 (3), 591–596. doi:10.1007/s10549-016-3885-x
- Caiado, H., Conceição, N., Tiago, D., Marreiros, A., Vicente, S., Enriquez, J. L., et al. (2020). Data on the evaluation of FGF2 gene expression in Colorectal Cancer. *Data Brief* 31, 105765. doi:10.1016/j.dib.2020.105765
- Chandrashekar, G., and Sahin, F. (2014). A survey on feature selection methods. *Comput. Electr. Eng.* 40 (1), 16–28. doi:10.1016/j.compeleceng.2013.11.024
- Chibucos, M. C., Siegle, D. A., Hu, J. C., and Giglio, M. (2017). "The evidence and conclusion ontology (ECO): supporting GO annotations," in *The gene ontology handbook*. Editors C. Dessimoz and N. Škunca (New York, NY: Springer), 245–259. doi:10.1007/978-1-4939-3743-1_18
- Chou, C.-H., Shrestha, S., Yang, C. D., Chang, N. W., Lin, Y. L., Liao, K. W., et al. (2018). miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions. *Nucleic Acids Res.* 46, D296–D302–D302. doi:10.1093/nar/gkx1067
- Cytoscape App Store - NetworkAnalyzer Cytoscape App store-NetworkAnalyzer.2021 Available at: <https://apps.cytoscape.org/apps/networkanalyzer> (Accessed 30 December 2022).
- Dang, L., Yen, K., and Attar, E. C. (2016). IDH mutations in cancer and progress toward development of targeted therapeutics. *Ann. Oncol.* 27 (4), 599–608. doi:10.1093/annonc/mdw013
- de Brot, S., Rutland, C. P. S., Mongan, N. P., and Victoria, J. (2018). "Chapter 20 - epigenetic control of MicroRNA expression and cancer," in *Cancer and noncoding RNAs*. Editors Dr. J. Chakrabarti and Dr. S. Mitra (Boston: Academic Press), 373–380. doi:10.1016/B978-0-12-811022-5.00020-6
- Denny, P., Feuermann, M., Hill, D. P., Lovering, R. C., Plun-Favreau, H., and Roncaglia, P. (2018). Exploring autophagy with gene ontology. *Autophagy* 14 (3), 419–436. doi:10.1080/15548627.2017.1415189
- Díaz, M. I., Díaz, P., Bennett, J. C., Urrea, H., Ortiz, R., Orellana, P. C., et al. (2020). Caveolin-1 suppresses tumor formation through the inhibition of the unfolded protein response. *Cell. Death Dis.* 11 (8), 648. doi:10.1038/s41419-020-02792-4
- Dudek, H., Datta, S. R., Franke, T. F., Birnbaum, M. J., Yao, R., Cooper, G. M., et al. (1997). Regulation of neuronal survival by the serine-threonine protein kinase akt. *Science* 275 (5300), 661–665. doi:10.1126/science.275.5300.661
- Fang, O. H., Mustapha, N., and Sulaiman, M. N. (2014). An integrative gene selection with association analysis for microarray data classification. *Intell. Data Anal.* 18 (4), 739–758. doi:10.3233/IDA-140666
- Franz, M., Lopes, C. T., Huck, G., Dong, Y., Sumer, O., and Bader, G. D. (2016). Cytoscape.js: a graph theory library for visualisation and analysis. *Bioinformatics* 32 (2), 309–311. doi:10.1093/bioinformatics/btv557
- Gharaibeh, L., Elmadany, N., Alwosaibai, K., and Alshaer, W. (2020). Notch1 in cancer therapy: possible clinical implications and challenges. *Mol. Pharmacol.* 98 (5), 559–576. doi:10.1124/molpharm.120.000006
- GSEA (2023). MSigDB Browse human gene sets. Available at: <https://www.gsea-msigdb.org/gsea/msigdb/genesets.jsp?collection=GO> (Accessed November 26, 2022).
- Guide to GO evidence codes (2022). Gene ontology resource. <http://geneontology.org/docs/guide-go-evidence-codes/> (Accessed December 27, 2022).
- Guyon, I. (2006). *Feature extraction. Foundations and applications. Papers from NIPS 2003 workshop on feature extraction, Whistler, BC, Canada, December 11–13, 2003. With CD-ROM*. Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-540-35488-8
- Hall, P. A., Todd, C. B., Hyland, P. L., McDade, S. S., Grabsch, H., Dattani, M., et al. (2005). The septin-binding protein anillin is overexpressed in diverse human tumors. *Clin. Cancer Res. Official J. Am. Assoc. Cancer Res.* 11 (19), 6780–6786. doi:10.1158/1078-0432.CCR-05-0997
- Inza, I., Larrañaga, P., Blanco, R., and Cerrolaza, A. J. (2004). Filter versus wrapper gene selection approaches in DNA microarray domains. *Artif. Intell. Med.* 31 (2), 91–103. doi:10.1016/j.artmed.2004.01.007
- Jabeer, A., Temiz, M., Bakir-Gungor, B., and Yousef, M. (2023). miRdisNET: discovering microRNA biomarkers that are associated with diseases utilizing biological knowledge-based machine learning. *Front. Genet.* 13, 1076554. doi:10.3389/fgene.2022.1076554
- Jensen, R., and Shen, Q. (2008). *Computational intelligence and feature selection: Rough and fuzzy approaches*. Hoboken, New Jersey: Wiley. doi:10.1002/9780470377888
- Jiang, T., Zhang, G., and Lou, Z. (2020). Role of the sterol regulatory element binding protein pathway in tumorigenesis. *Front. Oncol.* 10, 1788. doi:10.3389/fonc.2020.01788
- Kanehisa, M., and Goto, S. (2000). Kegg: kyoto Encyclopedia of genes and genomes. *Nucleic Acids Res.* 28 (1), 27–30. doi:10.1093/nar/28.1.27
- König, I. R., Fuchs, O., Hansen, G., von Mutius, E., and Kopp, M. V. (2017). What is precision medicine? *Eur. Respir. J.* 50 (4), 1700391. doi:10.1183/13993003.00391-2017
- Lei, S. (2012). 'A feature selection method based on information gain and genetic algorithm', in Proceedings of the 2012 International Conference on Computer Science and Electronics Engineering. Hangzhou, China, March 2012. 355–358. doi:10.1109/ICCSEE.2012.97
- Levy, S. E., and Boone, B. E. (2019). Next-generation sequencing strategies. *Cold Spring Harb. Perspect. Med.* 9 (7), a025791. doi:10.1101/cshperspect.a025791
- Li, B., Zhu, H. B., Song, G. D., Cheng, J. H., Li, C. Z., Zhang, Y. Z., et al. (2019). Regulating the CCNB1 gene can affect cell proliferation and apoptosis in pituitary adenomas and activate epithelial-to-mesenchymal transition. *Oncol. Lett.* 18 (5), 4651–4658. doi:10.3892/ol.2019.10847
- Li, H., Cao, Y., Li, S., Zhao, J., and Sun, Y. (2020). XGBoost model and its application to personal credit evaluation. *IEEE Intell. Syst.* 35 (3), 52–61. doi:10.1109/MIS.2020.2972533
- Li, Y., Casey, S. C., and Felsner, D. W. (2014). Inactivation of MYC reverses tumorigenesis. *J. Intern. Med.* 276 (1), 52–60. doi:10.1111/joim.12237
- Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J. P., and Tamayo, P. (2015). The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell. Syst.* 1 (6), 417–425. doi:10.1016/j.cels.2015.12.004
- Lin, Z., and Fang, D. (2013). The roles of SIRT1 in cancer. *Genes. & Cancer* 4 (3–4), 97–104. doi:10.1177/1947601912475079
- Liu, H., Li, D., Zhou, L., Kan, S., Zhou, K., et al. (2020). LMNA functions as an oncogene in hepatocellular carcinoma by regulating the proliferation and migration ability. *J. Cell. Mol. Med.* 24 (20), 12008–12019. doi:10.1111/jcmm.15829
- Liu, H., and Motoda, H. (1998). *Feature selection for knowledge discovery and data mining*. Boston, MA: springer US. doi:10.1007/978-1-4615-5689-3
- Lorente, G., Syriani, E., and Morales, M. (2014). Actin filaments at the leading edge of cancer cells are characterized by a high mobile fraction and turnover regulation by profilin I. *PLOS ONE* 9 (1), e85817. doi:10.1371/journal.pone.0085817
- Lou, Y., Jiang, Y., Liang, Z., Liu, B., Li, T., and Zhang, D. (2021). Role of RhoC in cancer cell migration. *Cancer Cell. Int.* 21, 527. doi:10.1186/s12935-021-02234-x

- Magliozzi, R., Low, T. Y., Weijts, B. G. M. W., Cheng, T., Spanjaard, E., Mohammed, S., et al. (2013). Control of epithelial cell migration and invasion by the IKK β - and CK1 α -mediated degradation of RAPGEF2. *Dev. Cell.* 27 (5), 574–585. doi:10.1016/j.devcel.2013.10.023
- MedlinePlus Genetics (2022). MedlinePlus genetics. Available at: <https://medlineplus.gov/genetics/gene/rpe65/> (Accessed September 7, 2022).
- Milella, M., Falcone, I., Conciatori, F., Cesta Incani, U., Del Curatolo, A., Inzerilli, N., et al. (2015). Pten: multiple functions in human malignant tumors. *Front. Oncol.* 5, 24. doi:10.3389/fonc.2015.00024
- Nair, S. S., and Kumar, R. (2012). Chromatin remodeling in cancer: a gateway to regulate gene transcription. *Mol. Oncol.* 6 (6), 611–619. doi:10.1016/j.molonc.2012.09.005
- Olakowski, M., Tyszkiewicz, T., Jarzab, M., Król, R., Oczko-Wojciechowska, M., Kowalska, M., et al. (2009). NBL1 and anillin (ANLN) genes over-expression in pancreatic carcinoma. *Folia Histochem. Cytobiol.* 47 (2), 249–255. doi:10.2478/v10042-009-0031-1
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830. doi:10.48550/arXiv.1201.0490
- Perscheid, C., Grasnack, B., and Uflacker, M. (2019). Integrative gene selection on gene expression data: providing biological context to traditional approaches. *J. Integr. Bioinforma.* 16 (1), 20180064. doi:10.1515/jib-2018-0064
- Perscheid, C. (2021). Integrative biomarker detection on high-dimensional gene expression data sets: a survey on prior knowledge approaches. *Briefings Bioinforma.* 22 (3), bbaa151. doi:10.1093/bib/bbaa151
- Pijuan, J., Barceló, C., Moreno, D. F., Maiques, O., Sisó, P., Marti, R. M., et al. (2019). *In vitro* cell migration, invasion, and adhesion assays: from cell imaging to data analysis. *Front. Cell. Dev. Biol.* 7, 107. doi:10.3389/fcell.2019.00107
- Piñero, J., Queralt-Rosinach, N., Bravo, À., Deu-Pons, J., Bauer-Mehren, A., Baron, M., et al. (2015). DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. *J. Biol. Databases Curation* 2015, bav028. doi:10.1093/database/bav028
- Qi, J., and Tang, J. (2007) 'Integrating gene ontology into discriminative powers of genes for feature selection in microarray data', in Proceedings of the 2007 ACM Symposium on Applied Computing (SAC). Seoul, Korea, March-2007, doi:10.1145/1244002.1244101
- Qumsiyeh, E., Showe, L., and Yousef, M. (2022). GediNET for discovering gene associations across diseases using knowledge based machine learning approach. *Sci. Rep.* 12 (1), 19955. doi:10.1038/s41598-022-24421-0
- Randa, N. C. (2019). Identification of PSEN1 and PSEN2 gene variants and clinical findings with the literature. *Int. J. Neurodegener. Disord.* 8. doi:10.23937/IJND-2017/1710007
- Sa-nguanraksa, D., and O-charoenrat, P. (2012). The role of vascular endothelial growth factor A polymorphisms in breast cancer. *Int. J. Mol. Sci.* 13 (11), 14845–14864. doi:10.3390/ijms131114845
- Senliol, B., Gulgezen, G., Yu, L., and Cataltepe, Z. (2008) 'Fast correlation based filter (FCBF) with a different search strategy', Proceedings of the 2008 23rd International Symposium on Computer and Information Sciences, Istanbul, Turkey, October 2008, 1–4. doi:10.1109/ISCIS.2008.4717949
- Shet, A. S. (2008). Characterizing blood microparticles: technical aspects and challenges. *Vasc. Health Risk Manag.* 4 (4), 769–774. doi:10.2147/vhrm.s955
- Sigismund, S., Avanzato, D., and Lanzetti, L. (2018). Emerging functions of the EGFR in cancer. *Mol. Oncol.* 12 (1), 3–20. doi:10.1002/1878-0261.12155
- Snaebjornsson, M. T., Janaki-Raman, S., and Schulze, A. (2020). Greasing the wheels of the cancer machine: the role of lipid metabolism in cancer. *Cell. Metab.* 31 (1), 62–76. doi:10.1016/j.cmet.2019.11.010
- Sodi, A., Banfi, S., Testa, F., Corte, M. D., Passerini, I., Pelo, E., et al. (2021). RPE65-associated inherited retinal diseases: consensus recommendations for eligibility to gene therapy. 11.
- Söylemez, Ü. G., Yousef, M., and Bakir-Gungor, B. (2023). AMP-GSM: prediction of antimicrobial peptides via a grouping-scoring-modeling approach. *Appl. Sci.* 13 (8), 5106. doi:10.3390/app13085106
- Srinivasa, K. G., Siddesh, G. M., and Manisekhar, S. R. (2020). *Statistical modelling and machine learning principles for bioinformatics techniques, tools, and applications*. Singapore: Springer. doi:10.1007/978-981-15-2445-5
- Tadist, K., Najah, S., Nikolov, N. S., Mrabti, F., and Zahi, A. (2019). Feature selection methods and genomic big data: a systematic review. *J. Big Data* 6 (1), 79. doi:10.1186/s40537-019-0241-0
- The Gene Ontology Consortium (2015). Gene ontology Consortium: going forward. *Nucleic Acids Res.* 43, D1049–D1056. doi:10.1093/nar/gku1179
- The Gene Ontology Consortium (2019). The gene ontology resource: 20 years and still GOing strong. *Nucleic Acids Res.* 47 (D1), D330–D338. doi:10.1093/nar/gky1055
- The Human Protein Atlas (2021). The human protein Atlas. Available at: <https://www.proteinatlas.org/ENSG00000177606-JUN/pathology> (Accessed September 7, 2022).
- The UniProt Consortium (2017). UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* 45 (D1), D158–D169. doi:10.1093/nar/gkw1099
- Tomczak, K., Czerwińska, P., and Wiznerowicz, M. (2015). The cancer genome Atlas (TCGA): an immeasurable source of knowledge. *Współczesna Onkol.* 1A, 68–77. doi:10.5114/wo.2014.47136
- Tuan, N. M., and Lee, C. H. (2020). Role of anillin in tumour: from a prognostic biomarker to a novel target. *Cancers* 12 (6), 1600. doi:10.3390/cancers12061600
- Uhlén, M., Fagerberg, L., Hallström, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., et al. (2015). Proteomics. Tissue-based map of the human proteome. *Sci. (New York, N.Y.)* 347 (6220), 1260419. doi:10.1126/science.1260419
- Wang, A., An, N., Chen, G., Li, L., and Alterovitz, G. (2015). Accelerating wrapper-based feature selection with K-nearest-neighbor. *Knowledge-Based Syst.* 83, 81–91. doi:10.1016/j.knsys.2015.03.009
- Wang, Q., He, G., Hou, M., Chen, L., Chen, S., Xu, A., et al. (2018). Cell cycle regulation by alternative polyadenylation of CCND1. *Sci. Rep.* 8 (1), 6824. doi:10.1038/s41598-018-25141-0
- Wheeler, D. L., Iida, M., and Dunn, E. F. (2009). The role of src in solid tumors. *Oncol.* 14 (7), 667–678. doi:10.1634/theoncologist.2009-0009
- Yoon, J., Blumer, A., and Lee, K. (2006). An algorithm for modularity analysis of directed and weighted biological networks based on edge-betweenness centrality. *Bioinforma. Oxf. Engl.* 22 (24), 3106–3108. doi:10.1093/bioinformatics/btl533
- You, W., Yang, Z., and Ji, G. (2014). PLS-based recursive feature elimination for high-dimensional small sample. *Knowledge-Based Syst.* 55, 15–28. doi:10.1016/j.knsys.2013.10.004
- Yousef, M., Abdallah, L., and Allmer, J. (2019). maTE: discovering expressed interactions between microRNAs and their targets. *Bioinformatics* 35 (20), 4020–4028. doi:10.1093/bioinformatics/btz204
- Yousef, M., Bakir-Gungor, B., Jabeer, A., Goy, G., Qureshi, R., and C Showe, L. (2020). 'Recursive cluster elimination based rank function (SVM-RCE-R) implemented in KNIME. *F1000Research* 9, 1255. doi:10.12688/f1000research.26880.2
- Yousef, M., Goy, G., and Bakir-Gungor, B. (2022b). miRModuleNet: detecting miRNA-mRNA regulatory modules. *Front. Genet.* 13, 767455. doi:10.3389/fgene.2022.767455
- Yousef, M., Goy, G., Mitra, R., Eischen, C. M., Jabeer, A., and Bakir-Gungor, B. (2021a). miRCorrNet: machine learning-based integration of miRNA and mRNA expression profiles, combined with feature grouping and ranking. *PeerJ* 9, e11458. doi:10.7717/peerj.11458
- Yousef, M., Jabeer, A., and Bakir-Gungor, B. (2021b). SVM-RCE-R-OPT: Optimization of scoring function for SVM-RCE-R. doi:10.1007/978-3-030-87101-7_21
- Yousef, M., Kumar, A., and Bakir-Gungor, B. (2021c). Application of biological domain knowledge based feature selection on gene expression data. *Entropy* 23 (1), 2. doi:10.3390/e23010002
- Yousef, M., Ozdemir, F., Jaaber, A., Allmer, J., and Bakir-Gungor, B. (2022a). PriPath: identifying dysregulated pathways from differential gene Expression via grouping, Scoring and Modeling with an embedded machine learning approach. *BMC Bioinforma.* 24 (1), 60. doi:10.21203/rs.3.rs-1449467/v1
- Yousef, M., Ozdemir, F., Jaber, A., Allmer, J., and Bakir-Gungor, B. (2023). PriPath: identifying dysregulated pathways from differential gene expression via grouping, scoring, and modeling with an embedded feature selection approach. *BMC Bioinforma.* 24 (1), 60. doi:10.1186/s12859-023-05187-2
- Yousef, M., Sayıcı, A., and Bakir-Gungor, B. (2021d) Integrating gene ontology based grouping and ranking into the machine learning algorithm for gene expression data analysis, mapping intimacies. Proceedings of the Database and Expert Systems Applications - DEXA 2021 Workshops, BIOKDD, IWCFS, MLKgraphs, AI-CARES, ProTime, AISys 2021, Virtual Event. September 2021, London, New York. doi:10.1007/978-3-030-87101-7_20
- Yousef, M., Ülgen, E., and Uğur Sezerman, O. (2021e). CogNet: classification of gene expression data based on ranked active-subnetwork-oriented KEGG pathway enrichment analysis. *PeerJ Comput. Sci.* 7, e336. doi:10.7717/peerj-cs.336
- Yousef, M., and Voskergian, D. (2022). TextNetTopics: text classification based word grouping as topics and topics' scoring. *Front. Genet.* 13, 893378. doi:10.3389/fgene.2022.893378
- Zaman, S., Chobrutskiy, B. I., Sikaria, D., and Blanck, G. (2019). MAPT (Tau) expression is a biomarker for an increased rate of survival for low-grade glioma. *Oncol. Rep.* 41 (2), 1359–1366. doi:10.3892/or.2018.6896
- Zhang, S., Xiong, X., and Sun, Y. (2020). Functional characterization of SOX2 as an anticancer target. *Signal Transduct. Target. Ther.* 5 (1), 135. doi:10.1038/s41392-020-00242-3
- Zhao, Z., Zou, S., Guan, X., Wang, M., Jiang, Z., Liu, Z., et al. (2018). Apolipoprotein E overexpression is associated with tumor progression and poor survival in colorectal cancer. *Front. Genet.* 9, 650. doi:10.3389/fgene.2018.00650