2024 32nd Signal Processing and Communications Applications Conference (SIU) | 979-8-3503-8896-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/SIU61531.2024.10601041

Classification of Breast Cancer Molecular Subtypes with Grouping-Scoring-Modeling Approach that Incorporates Disease-Disease Association Information

Emma Qumsiyeh
Department of Computer Science and Information Technology
Al-Quds University
Jerusalem, Palestine

Burcu Bakir-Gungor
Department of Computer Engineering
Faculty of Engineering
Abdullah Gul University
Kayseri, Turkey

Malik Yousef
Department of Information Systems
Galilee Digital Health Research Center
Zefat Academic College
Zefat, Israel

***Abstract*—This study uses modern sequencing technology and large biological databases to investigate the molecular intricacies of complicated diseases like cancer. Using gene expression databases and biomarkers, the research aims to improve breast cancer molecular subtype identification for better patient outcomes. Using BRCA LumAB_Her2Basal dataset, this study compares an integrative machine learning-based strategy (GediNET) to traditional feature selection approaches across machine learning classifiers. GediNET excels at uncovering crucial disease-disease connections and potential biomarkers using the Grouping-Scoring-Modeling (GSM) approach, which favors gene groupings above individual genes. Our comparative analysis highlights GediNET's exceptional performance, notably in terms of accuracy and Area Under the Curve metrics, underscoring its effectiveness in uncovering the genetic intricacies of breast cancer. GediNET's promise to improve disease classification and biomarker identification by improving biological mechanism understanding goes beyond exceeding traditional approaches. The work shows that GediNET's integrative method can promote bioinformatics research by identifying the most informative genes associated with certain diseases, enabling focused and customized medicine.**

Keywords—bioinformatics, integrative approach, feature selection methods, grouping-scoring-modeling (g-s-m), disease-disease associations, biomarker discovery, machine learning.

I. Introduction

Modern sequencing technology has greatly increased our understanding of complex disease molecular pathways [1]. Analysis of gene expression datasets is crucial for uncovering gene groups that contribute to the development and progression of various disorders and understanding their molecular mechanisms [2]. In the past decade, the surge in large-scale datasets has led to the creation of extensive data repositories, enriching the field of biological research. These resources include miRTarBase [3], Gene Ontology (GO) [4], and the Gene Expression Omnibus (GEO) [5]. Additionally, the Cancer Genome Atlas (TCGA) serves as a crucial database for gene expression and RNA sequencing data [6], while KEGG stands out as a detailed knowledge base for biological pathways [7]. DisGeNET, with its extensive gene-disease-variant connections, is another important database [8].

Lifestyle, diet, carcinogen exposure, oncogene and tumor suppressor gene mutations all contribute to cancer [9], [10].Worldwide, 2.3 million women are diagnosed with breast cancer each year, making it the most common malignancy. Breast cancer categorization improves patient care by allowing more personalized treatment plans [11]. Research in this domain has resulted in identifying diverse biological markers derived from various data sources. For instance, the overexpression of HER2 (epidermal growth factor receptor II) is associated with uncontrolled cell growth, and patients with HER2-positive breast cancer (BRCA) often have a poorer prognosis compared to those without HER2 overexpression [12].

Human diseases usually include gene or protein abnormalities in molecular pathways, causing severe symptoms. Genes with similar diseases or phenotypes may have similarities [13]. This discovery has changed research methodologies from focusing on individual genes/features to understanding a group of genes utilizing integrated information. Deep analysis tools that incorporate biological knowledge provide more insight than clustering and machine learning [14]. The rise of high-throughput technology in biology has shifted research towards more integrative methodologies, replacing the previously dominant traditional machine learning and clustering techniques [15]. While constructing bioinformatics pipelines, some recent approaches incorporate prior knowledge rather than relying solely on statistical and machine learning algorithms that often overlook the biological context. Such integrative approaches represent notable progress in bioinformatics [16], [17], [18].

The Grouping-Scoring-Modeling (GSM) approach, which was introduced by Yousef et al. [16], [18], represents a significant shift from traditional feature selection methods. Unlike traditional techniques that identify individual informative features, the GSM methodology groups these features into groups. These groups are then evaluated and scored, and a classification model is constructed based on these high-ranking groups. The strength of the GSM method resides in its flexibility. The GSM approach has been used in many computational tools. The original tool that evaluates groups of genes instead of individual genes was SVM-RCE-R tool [19], [20]. SVM-RCE (Support Vector Machines -Recursive Cluster Elimination) groups genes according to their gene expression values and assigns a score to each gene group using a machine learning algorithm. miRcorrNet [23], 3Mint [27], and miRModuleNet [28] identify groups from mRNA and miRNA expression datasets. maTE [21] utilizes microRNA target genes as groups. Gene Ontology [4] creates groups based on Gene Ontology information; CogNet [22] and PriPath [24] use KEGG

pathways for grouping; TextNetTopics [30] uses text topics as groups; miRdisNET [29] uses miRNA data and miRNA target genes as groups. A novel methodology known as miRGediNET [26] integrates information from three distinct biological databases to investigate the involvement of miRNAs in the advancement of diseases. microBiomeGSM [25] uses taxonomy information in disease-associated metagenomics datasets. GediNET [31] is based on the GSM approach to detect disease-disease associations (DDAs). This method categorizes genes into groups based on their associations with diseases using the DisGeNET database [8]. Subsequently, these groups are evaluated based on their importance in categorizing diseases, emphasizing the most noteworthy sets of genes. The top-ranked groups are then employed to train a machine-learning model. GediNET facilitates the identification of diseases that share common genetic markers. GediNETPro [32], an upgraded version, integrates cross-validation and clustering techniques like K-means to improve the identification of disease group connections. A new tool that uses the GSM method includes a statistical method for defining how similar two diseases are in terms of their semantic meaning [33].

In this study, we compared GediNET tool to standard feature selection approaches. Four classifiers and six feature selection methods were uniformly used to the BRCA-TCGA dataset. Our study explores this comparative landscape to illuminate GediNET's strengths, specifically its capacity to navigate and understand breast cancer data's complex genetics.

II. DATASET

We used the Breast Invasive Carcinoma (TCGA-BRCA) dataset [6] to test GediNET against other feature selection methods. The dataset was downloaded using Xena Public Data Hubs, which provided mRNA datasets mapped to GRCh38 [34]. PAM50 was used to identify BRCA's molecular intrinsic subtype groups. This assay classifies samples into molecular subtypes using 50 gene signatures [35]. Cancer samples were sorted by Luminal A, B, Her2-enriched, and Basal-like molecular subtypes. Two groups of samples were used: 124 (ER-negative, including Her2-enriched and Basal-like subtypes but not normal-like) and 248 (Luminal, including Luminal A and B subtypes). In this study, BRCA molecular subtype LumAB vs. Her2Basal was the main focus. LumAB samples are classified as positive, while Her2Basal samples are negative. Raw TCGA gene expression counts were retrieved and normalized using edgeR with the trimmed mean of M-values (TMM) [36].

III. METHODOLOGY

High-dimensional datasets affect classification methods. Thus, a feature selection strategy is needed to reduce feature size and simplify classification tasks [37]. GediNET, as a feature selection approach, differs from conventional strategies by focusing on the collective analysis of gene groups instead of isolating individual genes to specific diseases. This approach involves aggregating genes into groups based on their known associations with diseases. Once gene groups are established, the GediNET approach employs a scoring measurement to assess the classification significance of each group. The high-ranking gene groups, as determined by their classification significance, are then utilized to train a machine-learning model, which is Random Forest. A 100-fold Monte Carlo cross-validation (MCCV) procedure is used to compute different performance measurements. A subset of the data is selected as the training set for the MCCV methodology, whereas the remaining data is assigned to the test set. Subsequently, this process is randomly iterated, yielding new training and testing parts with each iteration. A training set comprising 90% of the data and a test set comprising 10% are utilized in our investigations.

This study compares the results of GediNET to traditional feature selection approaches using the BRCA LumAB_Her2Basal dataset. We used CMIM [38], mRmR [39], IG [40], SKB [41], FCBF [42], XGB [43]. An evaluation was conducted to assess the efficacy of various classification techniques, including RF, SVM, LogitBoost, Decision Tree, and AdaBoost. The default parameters are used to execute every algorithm. The feature selection methods were iteratively executed, and the outcomes were averaged and shared. GediNET selected 75 features on average in its analyses for the top 2 groups within 10 different datasets. Therefore, in these experiments and for the remaining feature selection methods, the number of features was set to 75.

IV. RESULTS

TABLE 1. AVERAGE CUMULATIVE PERFORMANCE OF GEDINET ACROSS 100 ITERATIONS FOR THE TOP 6 GENE GROUPS.

#Groups	#Genes	Sensitivity	Specificity	Precision	Accuracy	AUC	F-measure
1	98	0.99	0.96	0.98	0.98	0.99	0.98
2	148	0.99	0.96	0.98	0.98	0.99	0.99
3	197	0.99	0.96	0.98	0.98	0.99	0.99
4	245	0.99	0.96	0.98	0.98	0.99	0.99
5	267	0.99	0.97	0.98	0.98	0.99	0.99
6	287	0.99	0.96	0.98	0.98	0.99	0.99

Table 1 illustrates GediNET's performance across the top 6 gene groups in the BRCA LumAB_Her2Basal dataset. The reported results refer to the mean outcomes from 100 random MCCV iterations, illustrating the performance indices for the cumulative groups of highest-ranking genes. In this evaluation, each row encapsulates the collective performance metrics for the top-ranked gene groups. For instance, in Table 1, the row that belongs to #Groups = 1 shows the performance indicators employing solely the first group, which, on average, encompasses around 99 genes. This first group exhibits remarkable competence, attaining an average AUC of 99%, which signifies its potent discriminative capacity. The row for #Groups = 2 details the performance metrics when the genes from the top-ranked group are combined with those from the second-ranked group, thereby enhancing the dataset for analysis.

TABLE 2. RESULTS OBTAINED WITH ROBUSTRANKAGGREG IN GEDINET .

Group	p-value	List of genes
Breast carcinoma	2.14394E-95	GCLC, CFTR, …
Malignant neoplasm of prostate	2.35477E-94	KRIT1, BAD, …
Malignant neoplasm of breast	3.36329E-88	NFYA, STPG1, …
Leukemia	3.49327E-85	CD99, CASP10, …

Table 2 presents an example of the top-ranked groups of GediNET, along with their assigned P-values obtained through robust rank aggregation [44]. Results prompts a biological inquiry into the links between the highly ranked diseases (such as BREAST CARCINOMA, MALIGNANT NEOPLASM OF

PROSTATE, MALIGNANT NEOPLASM OF BREAST, etc.) and breast cancer molecular subtype identification problem.

A performance comparison of several feature selection (FS) models, namely XGB, IG, SKB, CMIM, FCBF, MRMR, and GediNET, was conducted using a Random Forest (RF) classifier, evaluating across metrics such as Accuracy, Sensitivity, Specificity, F-measure, Precision, and Area Under Curve (AUC). The results are presented in Figure 1.

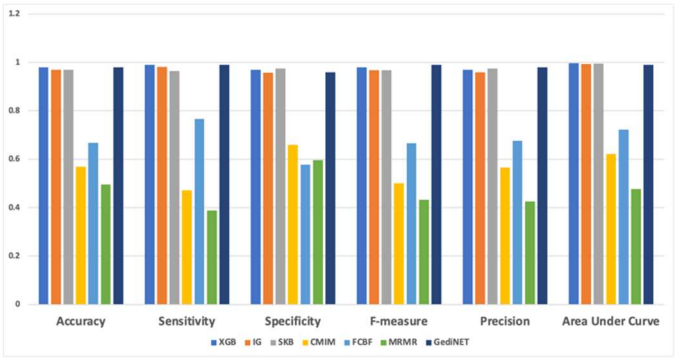

Fig. 1. Comparative Analysis of Feature Selection Models Using Random Forest Classifier in the BRCA LumAB_Her2Basal Dataset.

Figure 1 highlights the accuracy metric, reflecting the model's overall correctness, where GediNET and XGB stand out with the highest accuracies of 0.98 and 0.9796, respectively. IG and SKB also show high accuracy. In contrast, MRMR and CMIM have significantly lower accuracies, with MRMR being the lowest at 0.496. XGB, with its near-perfect AUC of 0.99, showcases an exceptional capacity to differentiate between classes, a critical aspect of model performance. SKB stands out for its specificity of approximately 0.97 and precision of 0.97. Although not topping the charts, IG still maintains a strong presence with an accuracy of 0.96 and an F-measure of 0.96. In contrast, MRMR and CMIM appear to grapple with the dataset when paired with an RF classifier. The FCBF method demonstrates moderate performance, as indicated by an accuracy of 0.66 and an AUC of 0.72. This suggests that FCBF has the potential to be a balanced feature selection method, particularly in situations where prioritizing extreme values in specific metrics is not necessary.

Figure 1 shows that GediNET and XGB are the top-performing feature selection models when paired with an RF classifier across all metrics.

TABLE 3. PERFORMANCE METRICS OF XGB FEATURE SELECTION ACROSS DIFFERENT MACHINE LEARNING MODELS.

FS Model	ML Model	Accuracy	Sensitivity	Specificity	F-measure	Precision	AUC
XGB	Adaboost	0.96	0.96	0.95	0.96	0.95	0.99
XGB	DT	0.94	0.94	0.94	0.94	0.94	0.94
XGB	LogitBoost	0.96	0.97	0.96	0.96	0.96	0.99
XGB	RF	0.97	0.99	0.97	0.97	0.96	0.99

Table 3 compares XGB feature selection among machine-learning models on the BRCA LumAB_Her2Basal dataset. Table 3 illustrates that the XGB feature selection strategy with several machine learning models performs well. In combination with a Random Forest (RF) classifier, the highest sensitivity and best AUC ties indicate good prediction performance. All models perform well, with AUC values often reaching 0.99, although the RF classifier works better with XGB for feature selection, showing a synergistic impact. Decision Trees (DT) perform slightly worse across the board, showing that classifier choice can affect feature effectiveness.

Moreover, we used the genes associated with the top-ranked group from Table 2 to analyze the common features between the top features of the different FS methods and GediNET. The results are presented in Table 4.

TABLE 4. COMPARISON OF THE TOP-RANKED GROUP'S FEATURES FROM GEDINET WITH THE 10 MOST INFORMATIVE FEATURES IDENTIFIED BY IG FOR THE BRCA LUMAB_HER2BASAL DATASET.

FS/Classifier	Top- ten Features Identified by FS Methods	Features Identified by the first top-ranked group by RRA in GediNET	Common Features between the Top Features of the FS Method and
IG/ Adaboost	PTGER3, TTK, PHGDH, A2ML1, ESR1, TTLL4, FAM241A, PGR GASK1B-AS1, SUSD3	FGR,CFH, GCLC,NFYA, SEMA3F,ESR1, CA12, CFTR, ANKIB1,KRIT1, RAD52,LRRC51, RNASEH1,AMER1 IL6ST, MAPT NOVA1, SLC39A6 BAD, LAP3, TFPI RBM5, SLC7A2, POLDIP2, CD38	PTGER3 PHGDH A2ML1 ESR1 TTLL4 FAM241A PGR
IG / LogitBoost	JOSD1, PTGER3 SYBU LINC01116, FAM241A, PHGDH, ESR1, CCDC170, A2ML1, NAT1, PGR		PTGER3 PHGDH ESR1, PGR
IG / RF	ESR1, CCDC170, GATA3, CA12, TBC1D9, AGR3, SLC39A6, NAT1, MAPT, C5AR2		ESR1 CA12

Table 4 compares the most relevant features discovered by IG feature selection and different classifiers with their correlation with the top-ranked group features determined by Robust Rank Aggregate in GediNET for BRCA LumAB_Her2Basal dataset. The table presents specific features identified by each feature selection method when combined with various classifiers and compares them with the features identified by GediNET's reliable ranking strategy.

The feature selection approach identifies PTGER3, PHGDH, A2ML1, ESR1, TTLL4, FAM241A, and PGR as highly ranked features for the IG/Adaboost combination. These features are also identified by the first top-ranked group in GediNET. Furthermore, the combination of IG/LogitBoost analysis indicates that PTGER3, PHGDH, ESR1, and PGR are shared between the selected features and the highest-ranked group in GediNET, hence highlighting the significance of these features. The IG/RF approach also resembles GediNET in recognizing ESR1 and CA12 as notable characteristics. The consistent presence of these features in multiple methods indicates a strong connection with the BRCA LumAB_Her2Basal dataset. This table represents an advancement in comprehending the intricate biological connections and possible biomarkers associated with BRCA molecular subtype identification.

DISCUSSION AND CONCLUSION

Comparing GediNET to standard feature selection and machine learning methods shows considerable advances in comprehending complicated diseases like breast cancer. GediNET's unique Grouping-Scoring-Modeling (GSM) technique finds relevant disease-disease associations and biomarkers effectively. Its capacity to classify gene groups, unlike traditional methods, confirms this.

GediNET's resilience and accuracy were demonstrated in this BRCA-TCGA dataset research using several classifiers and feature selection methods. Its use of domain-specific knowledge and machine learning to identify essential genetic markers gave it an edge over previous methods.

Additionally, this work emphasizes the need of combining biological and computational methodologies. This synthesis improves disease classification and biomarker discovery. GediNET's main goal is not to compete with other feature selection methods. As shown by its ranks across all measures, GediNET and XGB perform better using a Random Forest classifier. Our research is not limited by this achievement. GediNET is important because it can precisely identify the most informative genes significantly associated with specific diseases. This ability is crucial because it allows scientists to study the complex molecular causes of many diseases and better understand their mechanisms. Therefore, GediNET's true worth is its tremendous contribution to biological understanding.

REFERENCES

[1] O. Akintunde, T. Tucker, and V. J. Carabetta, "The evolution of next-generation sequencing technologies," *ArXiv*, p. arXiv:2305.08724v1, May 2023.

[2] K. Batko and A. Ślęzak, "The use of Big Data Analytics in healthcare," *J Big Data*, vol. 9, no. 1, p. 3, 2022, doi: 10.1186/s40537-021-00553-4.

[3] miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database | Nucleic Acids Research | Oxford Academic." Accessed: Nov. 30, 2021.

[4] "Gene Ontology: tool for the unification of biology | Nature Genetics." Accessed: Nov. 30, 2021. [Online]. Available: https://www.nature.com/articles/ng0500_25/

[5] T. Barrett *et al.*, "NCBI GEO: archive for functional genomics data sets--update," *Nucleic Acids Res*, vol. 41, no. Database issue, pp. D991-995, Jan. 2013, doi: 10.1093/nar/gks1193.

[6] K. Tomczak, P. Czerwińska, and M. Wiznerowicz, "The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge," *Contemp Oncol (Pozn)*, vol. 19, no. 1A, pp. A68-77, 2015, doi: 10.5114/wo.2014.47136.

[7] "KEGG: Kyoto Encyclopedia of Genes and Genomes | Nucleic Acids Research | Oxford Academic." Accessed: Jul. 04, 2023. [Online].

[8] J. Piñero *et al.*, "DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants," *Nucleic Acids Res*, vol. 45, no. D1, pp. D833–D839, Jan. 2017, doi: 10.1093/nar/gkw943.

[9] B. Chen, X. Shang, M. Li, J. Wang, and F.-X. Wu, "Identifying Individual-Cancer-Related Genes by Rebalancing the Training Samples," *IEEE Transactions on NanoBioscience*, vol. 15, pp. 1–1, Apr. 2016, doi: 10.1109/TNB.2016.2553119.

[10] M. G. Kann, "Advances in translational bioinformatics: computational approaches for the hunting of disease genes," *Briefings in Bioinformatics*, vol. 11, no. 1, pp. 96–110, Jan. 2010, doi: 10.1093/bib/bbp048.

[11] S. Łukasiewicz, M. Czeczelewski, A. Forma, J. Baj, R. Sitarz, and A. Stanisławek, "Breast Cancer—Epidemiology, Risk Factors, Classification, Prognostic Markers, and Current Treatment Strategies—An Updated Review," *Cancers (Basel)*, vol. 13, no. 17, p. 4287, Aug. 2021, doi: 10.3390/cancers13174287.

[12] J.-C. Neel and J.-J. Lebrun, "Activin and TGFβ regulate expression of the microRNA-181 family to promote cell migration and invasion in breast cancer cells," *Cell Signal*, vol. 25, no. 7, pp. 1556–1566, Jul. 2013, doi: 10.1016/j.cellsig.2013.03.013.

[13] J. Gillis and P. Pavlidis, "'Guilt by Association' Is the Exception Rather Than the Rule in Gene Networks," *PLOS Computational Biology*, vol. 8, no. 3, p. e1002444, Mar. 2012, doi: 10.1371/journal.pcbi.1002444.

[14] C. Deng, X. Ji, C. Rainey, J. Zhang, and W. Lu, "Integrating Machine Learning with Human Knowledge," *iScience*, vol. 23, no. 11, p. 101656, Nov. 2020, doi: 10.1016/j

[15] B. Mirza, W. Wang, J. Wang, H. Choi, N. C. Chung, and P. Ping, "Machine Learning and Integrative Analysis of Biomedical Big Data," *Genes*, vol. 10, no. 2, p. 87, Jan. 2019.

[16] M. Yousef, A. Kumar, and B. Bakir-Gungor, "Application of Biological Domain Knowledge Based Feature Selection on Gene Expression Data," *Entropy (Basel)*, vol. 23, no. 1, p. E2, Dec. 2020, doi: 10.3390/e23010002.

[17] C. Perscheid, "Integrative biomarker detection on high-dimensional gene expression data sets: a survey on prior knowledge approaches," *Briefings in Bioinformatics*, vol. 22, no. 3, p. bbaa151, May 2021, doi: 10.1093/bib/bbaa151.

[18] C. Kuzudisli, B. Bakir-Gungor, N. Bulut, B. Qaqish, and M. Yousef, "Review of feature selection approaches based on grouping of features," *PeerJ*, vol. 11, p. e15666, Jul. 2023, doi: 10.7717/peerj.15666.

[19] M. Yousef, A. Jabeer, and B. Bakir-Gungor, "SVM-RCE-R-OPT: Optimization of Scoring Function for SVM-RCE-R," in *Database and Expert Systems Applications - DEXA 2021 Workshops*.

[20] M. Yousef, B. Bakir-Gungor, A. Jabeer, G. Goy, R. Qureshi, and L. C. Showe, "Recursive Cluster Elimination based Rank Function (SVM-RCE-R) implemented in KNIME," *F1000Res*, vol. 9, p. 1255, Jan. 2021, doi: 10.12688/f1000research.26880.2.

[21] M. Yousef, L. Abdallah, and J. Allmer, "maTE: discovering expressed interactions between microRNAs and their targets," *Bioinformatics*, vol. 35, no. 20, pp. 4020–4028, Oct. 2019, doi: 10.1093/bioinformatics/btz204.

[22] M. Yousef, E. Ülgen, and O. Uğur Sezerman, "CogNet: classification of gene expression data based on ranked active-subnetwork-oriented KEGG pathway enrichment analysis," *PeerJ Comput Sci*, vol. 7, p. e336, 2021, doi: 10.7717/peerj-cs.336.

[23] M. Yousef, G. Goy, R. Mitra, C. M. Eischen, A. Jabeer, and B. Bakir-Gungor, "miRcorrNet: machine learning-based integration of miRNA and mRNA expression profiles, combined with feature grouping and ranking," *PeerJ*, vol. 9, p. e11458, 2021, doi: 10.7717/peerj.11458.

[24] M. Yousef, F. Ozdemir, A. Jaaber, J. Allmer, and B. Bakir-Gungor, "PriPath: Identifying Dysregulated Pathways from Differential Gene Expression via Grouping, Scoring and Modeling with an Embedded Machine Learning Approach," In Review, preprint, Apr. 2022. doi: 10.21203/rs.3.rs-1449467/v1.

[25] B. Bakir-Gungor, M. Temiz, A. Jabeer, D. Wu, and M. Yousef, "microBiomeGSM: the identification of taxonomic biomarkers from metagenomic data using grouping, scoring and modeling (G-S-M) approach," *Front Microbiol*, vol. 14, p. 1264941, Nov. 2023.

[26] E. Qumsiyeh, Z. Salah, and M. Yousef, "miRGediNET: A comprehensive examination of common genes in miRNA-Target interactions and disease associations: Insights from a grouping-scoring-modeling approach," *Heliyon*, vol. 9, no. 12, p. e22666, Dec. 2023, doi: 10.1016/j.heliyon.2023.e22666.

[27] M. Unlu Yazici, J. S. Marron, B. Bakir-Gungor, F. Zou, and M. Yousef, "Invention of 3Mint for feature grouping and scoring in multi-omics," *Frontiers in Genetics*, vol. 14, 2023.

[28] M. Yousef, G. Goy, and B. Bakir-Gungor, "miRModuleNet: Detecting miRNA-mRNA Regulatory Modules," *Front Genet*, vol. 13, p. 767455, 2022, doi: 10.3389/fgene.2022.767455.

[29] A. Jabeer, M. Temiz, B. Bakir-Gungor, and M. Yousef, "miRdisNET: Discovering microRNA biomarkers that are associated with diseases utilizing biological knowledge-based machine learning," *Frontiers in Genetics*, vol. 13, 2023, Accessed: Jul. 07, 2023.

[30] M. Yousef and D. Voskergian, "TextNetTopics: Text Classification Based Word Grouping as Topics and Topics' Scoring," *Frontiers in Genetics*, vol. 13, 2022.

[31] E. Qumsiyeh, L. Showe, and M. Yousef, "GediNET for discovering gene associations across diseases using knowledge based machine learning approach," *Sci Rep*, vol. 12, no. 1, Art. no. 1, Nov. 2022, doi: 10.1038/s41598-022-24421-0.

[32] E. Qumsiyeh, M. Yazıcı, and M. Yousef, "GediNETPro: Discovering Patterns of Disease Groups," in *Proceedings of the 16th International Joint Conference on Biomedical Engineering Systems and Technologies - BIOINFORMATICS*, SciTePress, 2023, pp. 195–203. doi: 10.5220/0011690800003414.

[33] E. Qumsiyeh, M. Yousef, Z. Salah, and R. Jayousi, "Detecting Semantic Similarity of Diseases based Machine Learning," in *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Dec. 2023, pp. 3118–3124. doi: 10.1109/BIBM58861.2023.10385728.

[34] M. J. Goldman *et al.*, "Visualizing and interpreting cancer genomics data via the Xena platform," *Nat Biotechnol*, vol. 38, no. 6, pp. 675–678, Jun. 2020

[35] J. S. Parker *et al.*, "Supervised risk predictor of breast cancer based on intrinsic subtypes," *J Clin Oncol*, vol. 27, no. 8, pp. 1160–1167, Mar. 2009, doi: 10.1200/JCO.2008.18.1370.

[36] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data," *Bioinformatics*, vol. 26, no. 1, pp. 139–140, Jan. 2010, doi: 10.1093/bioinformatics/btp616.

[37] A. Kumar, A. Kaur, P. Singh, M. Driss, and W. Boulila, "Efficient Multiclass Classification Using Feature Selection in High-Dimensional Datasets," *Electronics*, vol. 12, no. 10, Art. no. 10, Jan. 2023, doi: 10.3390/electronics12102290.

[38] G. Brown, A. Pocock, M.-J. Zhao, and M. Luján, "Conditional Likelihood Maximisation: A Unifying Framework for Information Theoretic Feature Selection," *Journal of Machine Learning Research*, vol. 13, no. 2, pp. 27–66, 2012.

[39] Hanchuan Peng, Fuhui Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.

[40] J. T. Kent, "Information gain and a general measure of correlation," *Biometrika*, vol. 70, no. 1, pp. 163–173, 1983, doi: 10.1093/biomet/70.1.163.

[41] T. Desyani, A. Saifudin, and Y. Yulianti, "Feature Selection Based on Naive Bayes for Caesarean Section Prediction," *IOP Conf. Ser.: Mater. Sci. Eng.*, vol. 879, no. 1, p. 012091, Jul. 2020, doi: 10.1088/1757-899X/879/1/012091.

[42] L. Yu and H. Liu, "Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution," *Proceedings of the Twentieth International Conference on Machine Learning*, vol. Washington DC, 2003.

[43] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco California USA: ACM, Aug. 2016, pp. 785–794.

[44] R. Kolde, S. Laur, P. Adler, and J. Vilo, "Robust rank aggregation for gene list integration and meta-analysis," *Bioinformatics*, vol. 28, no. 4, pp. 573–580, Feb. 2012.