Ahliya Journal of Allied Medico-Technology Science
[Published by Palestine Ahliya University]

Vol. 01 Issue 01 (2024)
ISSN: 3007-9489

Enhancing Breast Cancer Subtype Classification through GediNET: Integrating Disease- Disease Association Data with a Grouping-Scoring-Modeling Approach

Emma Qumsiyeh[1]

[1] Faculty of Engineering and Information Technology, Palestine Ahliya University (Palestine)

✉ e.qumsiyeh@paluniv.edu.ps

Received: 15/05/2024 Accepted: 01/06/2024 Published: 30/06/2024

Abstract: ***The development of sequencing technology and the increase in biological data repositories have allowed for a more thorough understanding of the complex molecular aspects of diseases like cancer. This paper evaluates GediNET, an integrative machine learning approach that employs a Grouping-Scoring-Modeling (GSM) approach to classify different molecular subtypes of breast cancer using the BRCA LumAB_Her2Basal dataset against different feature selection approaches and machine learning classifiers. GediNET distinguishes itself from traditional feature selection methods by analyzing groups of genes to identify relevant disease-disease associations and potential biomarkers. The results of our study show that GediNET performs better than traditional approaches in terms of accuracy and Area Under the Curve (AUC) metrics. This demonstrates that GediNET is effective in understanding the genetic intricacies of breast cancer. This approach improves the identification of molecular subtypes and promotes the development of targeted medicines and customized medicine.***

***Keywords:* GediNET; Breast Cancer Subtype Classification; Grouping-Scoring-Modeling Approach; Machine Learning; Disease-Disease Associations; Biomarker Discovery; Genomic Data Analysis.**

1. Introduction

The recent progress in sequencing technologies has fundamentally transformed our comprehension of the molecular pathways that underlie intricate diseases [1]. These technologies offer a vast amount of biological data, which allows for a thorough examination of gene expression databases [2]. This is crucial for identifying gene groups that impact disease progression and gaining insight into molecular pathways. The increase in data accessibility has resulted in the creation of comprehensive biological databases such as miRTarBase [3], Gene Ontology(GO) [4], the Gene Expression Omnibus (GEO) [5], The Cancer Genome Atlas (TCGA) [6], and DisGeNET [7]. These tools are essential for performing in-silico experiments and creating statistical machine-learning models for disease classification and biomarker development.

The identification of molecular subtypes is necessary for the development of tailored therapeutics for breast cancer, a multifaceted disease influenced by genetic and environmental factors [8]. Classification is crucial for improving patient care, as it enables the development of customized treatment plans for individuals. Continuing study in this area has resulted in identifying several biological indicators obtained from diverse data repositories. For instance, the excessive expression of HER2 (epidermal growth factor receptor II) is associated with the rapid and uncontrolled development of cells. Patients with HER2-positive breast cancer (BRCA) often have a worse prognosis than those whose tumors do not have an overexpression of HER2 [9]. This comprehensive comprehension aids in formulating targeted therapies that can alleviate the rapid advancement of the disease in affected persons.

Human disorders are frequently characterized by significant disruptions in genes or proteins within molecular pathways, resulting in many symptoms, some of which can be highly severe. Based on "guilt-by-association," interconnected genes are likely to have shared functions, which can be attributed to genetic or physical linkages. Therefore, genes associated with similar diseases or phenotypes are likely to have resemblances [10]. This comprehension has accelerated a notable transformation in research approaches, transitioning from concentrating on individual genes or characteristics to analyzing gene clusters within a framework incorporating extensive biological knowledge. These techniques improve the level of analysis beyond what can be achieved with classic clustering and machine-learning methods. Furthermore, the introduction of high-throughput technology in the field of biology has led to a shift towards more comprehensive research approaches, replacing traditional methods in machine learning and clustering [11].

Researchers are utilizing advanced techniques to provide a comprehensive framework for analyzing individual diseases by harnessing biological knowledge. When constructing bioinformatics pipelines, modern methods rely more on existing information rather than solely relying on statistics and machine learning algorithms. This is because these algorithms may overlook the crucial biological context. These integrative methods are essential progressions in the field, combining a thorough comprehension of biological processes with state-of-the-art analytical techniques to produce a more detailed understanding of gene expression data and the fundamental causes of diseases [12]. The integrative approaches not only improve our comprehension of disease pathophysiology but also facilitate the creation of specific treatments and diagnostic methods, thus advancing the limits of customized medicine [13].

The Grouping-Scoring-Modeling (GSM) framework, introduced by Yousef et al. [14], [15], represents a significant change in traditional approaches and feature selection methodology. The GSM framework differs from conventional models as it identifies individual genes by grouping features into distinct groups. The groupings of genes are evaluated and assigned scores, which are then used to create a categorization model based on the highest-ranking groups. The GSM is unique because it can incorporate computational, statistical, or domain-specific knowledge into the grouping process. This requires a lot of expertise and makes each application unique and powerful. GSM has been integrated into various computational tools, such as SVM-RCE-R [16], maTE [17], CogNet [22], miRcorrNet [18], PriPath [19], miRGediNET [20], and miRdisNET [21] and others.

The introduction of gene grouping in gene analysis originated from the SVM-RCE-R program [22], which utilizes Support Vector Machines for Recursive Cluster Elimination to categorize genes based on their expression values, awarding scores using a machine-learning method. miRcorrNet [18], 3Mint [23], and miRModuleNet [24] are tools that analyze and extract groups from datasets, including mRNA and miRNA information. Additional tools such as maTE [17] specifically target microRNA target genes. CogNet [25] and PriPath [19] employ KEGG pathways. Utilizing information from multiple biological databases, the miRGediNET [26] is a unique tool that explores the role of miRNAs in the development of disease. On the other hand, microBiomeGSM [27] utilizes taxonomic information from metagenomics datasets to classify diseases based on GSM principles.

GediNET [28], an integrative GSM approach, intends to discover disease-disease associations (DDAs) by grouping genes into groups according to disease knowledge obtained from the DisGeNET database [7]. Subsequently, these groupings are examined to identify the most noteworthy gene sets to classify diseases. GediNET utilizes the input from the leading groups to enhance the training of machine-learning models, allowing for the identification of common genetic markers among different diseases. The updated version of GediNET, named GediNETPro [29], utilizes Monte Carlo cross-validation and clustering techniques, such as K-means, to enhance the identification of disease groups. Another tool enhances this process by using a statistical method to measure the semantic similarities between diseases [30], enabling a detailed examination of disease clusters by applying Monte Carlo cross-validation and semantic evaluation.

This study thoroughly evaluates the GediNET using a Grouping-Scoring-Modeling (GSM) approach compared to traditional feature selection methods and several feature selection approaches. It utilizes various classifiers and feature selection strategies on the BRCA-TCGA dataset. Examining these comparative dynamics, our goal is to emphasize GediNET's proficiency in traversing and comprehending the intricate genomic landscape of breast cancer. This will highlight the significant differences in performance, analytical accuracy, and the level of biological understanding obtained from the dataset.

2. Dataset

In this study, we employed the Breast Invasive Carcinoma (TCGA-BRCA) dataset [6], a comprehensive collection of gene expression data designed for breast cancer research. The dataset was obtained from the Xena Public Data Hubs [31], which provide access to a diverse collection of mRNA datasets. These datasets enable researchers to investigate different aspects of cancer biology by analyzing gene expression profiles.

Our dataset primarily consists of tumor samples that are classified into four molecular subtypes of breast cancer: Luminal A (LumA), Luminal B (LumB), Her2-enriched, and Basal-like [8]. To optimize our study and improve the precision of our results, we have merged these findings into two primary categories for classification:

1. The Positive Group (LumA) consists of 302 samples belonging to the Luminal A subtype, which is characterized by a more favorable prognosis and lower grade in comparison to other subtypes.

2. The Negative Group (LumBHer2Basal) comprises 247 samples, including the Luminal B, Her2-enriched, and Basal-like subtypes. These subtypes are typically associated with a more aggressive behavior and worse results.

The analytical approach began by extracting the raw gene expression numbers from the TCGA database. Afterward, the data was preprocessed by normalizing the counts using the TMM method from the edgeR package [32]. Normalization is a critical approach that accounts for the impacts of RNA composition. It allows for more precise comparisons between samples by stabilizing variance and improving the discovery of genes that are expressed differently.

3. Methodology

Handling with datasets that have a large number of dimensions poses a considerable difficulty in classification tasks because of the intricate nature and size of the data. To properly handle this, it is crucial to have a robust feature selection method that reduces the number of features and simplifies the classification process [14].

Our study utilizes GediNET [28], a novel feature selection strategy that diverges from traditional approaches by prioritizing the collective analysis of gene groups rather than individual genes associated with specific disorders. This methodology categorizes genes according to their associations with different diseases, considering the intricate biological context in which these genes operate. Subsequently, these groupings are evaluated and assigned scores to ascertain their significance in disease categorization. GediNET utilizes these scores to identify the most prominent gene groups, which are subsequently employed to train a machine-learning model, specifically a Random Forest classifier. We use a 100-fold Monte Carlo cross-validation (MCCV) technique [33] to accurately assess performance measures. In this method, 90% of the data is allocated for training, and the remaining 10% is used for testing. The subsets are randomly selected in each iteration to ensure thorough coverage and unbiased evaluation.

This study aims to evaluate the performance of GediNET in comparison to traditional feature selection methods on the BRCA LumA_LumBHer2Basal dataset. The comparative analysis incorporates various well-established feature selection techniques, including conditional mutual information maximization (CMIM) [34], minimum redundancy maximum relevance (mRmR) [35], information gain (IG) [36], SelectKBest (SKB) [37], Fast Correlation Based Filter (FCBF) [38], and extreme gradient boosting (XGB) [39]. The effectiveness of several classifiers, such as Random Forest (RF) [40], Support Vector Machine (SVM) [41], LogitBoost [42], Decision Tree [39], and AdaBoost [43], was evaluated using these feature selection techniques. Uniform settings were used for all algorithms to ensure uniformity. Each feature selection approach, including GediNET, was responsible for picking a set of 75 features that would be best for analysis. This decision was based on the findings obtained from the top two groups in ten different datasets. This standardized approach effectively assesses GediNET's capacity to improve feature selection and classification accuracy in breast cancer subtypes.

4. Results

Table 1 displays the results obtained by GediNET on the BRCA LumA_LumBHer2Basal dataset. It gives a comprehensive examination of the performance changes when additional gene groups are included in the study. At the initial phase, when the number of groups is set to 1, the performance measures depend exclusively on the genes in the first group. The initial setup of this model provides a summary of its capability to categorize different forms of breast cancer with an accuracy of 90.1%, a sensitivity of 94.1%, and a specificity of 81.7%. Significantly, the Area Under Curve (AUC), which is a vital metric for evaluating the model's capacity to differentiate between classes, exhibits a great first performance of 96.1%.

When transitioning to #Groups = 2, which involves analyzing the top two ranked groups of genes together, there is a noticeable enhancement in most measures. The accuracy exhibits a modest increase to 90.6%, the sensitivity experiences a marginal improvement to 94.4%, and the specificity demonstrates a rise to 82.6%. These results indicate an enhanced performance resulting from the inclusion of additional genetic information. This pattern highlights the extra value provided by each gene group, demonstrating an improved ability to predict and a strong resilience of the model.

Following this trend, the performance metrics for each consecutive group (ranging from #Groups = 3 to #Groups = 10) exhibit a consistent and stable level, with only minimal variations. When the number of groups is set to 6, the accuracy reaches its highest point at 91.0%, the sensitivity reaches its highest point at 95.0%, and the AUC remains consistently high at 96.4%. The results suggest that the initial gene groups contribute significantly to the predictive capabilities of the model, but further additions improve its accuracy and resilience.

The consistent AUC values observed across all groups, predominantly around the 96% range, indicate that GediNET can discriminate regardless of the number of gene groups utilized. The findings demonstrate a model that initially exhibits excellent efficacy and after that achieves incremental yet significant enhancements by incorporating supplementary gene clusters. This pattern indicates that although the main gene groups are strong predictors, the supplementary groups aid in capturing more specific details, which may be associated with less dominant but medically significant genetic expressions. This offers a more comprehensive understanding of the genomic landscape in different breast cancer subtypes.

Table 1: Performance Metrics of GediNET Across Top 10 Gene Groups in the BRCA LumA_LumBHer2Basal Dataset Over 100 Monte Carlo Cross-Validation Iterations.

#Groups	Accuracy	Sensitivity	Specificity	F-measure	Area Under Curve	Cohen's kappa	Precision
1	0.901	0.941	0.817	0.928	0.961	0.769	0.917
2	0.906	0.944	0.826	0.931	0.964	0.781	0.921
3	0.907	0.947	0.824	0.933	0.965	0.784	0.92
4	0.907	0.944	0.829	0.932	0.964	0.783	0.922
5	0.905	0.946	0.818	0.931	0.964	0.777	0.918
6	0.91	0.95	0.828	0.935	0.964	0.791	0.922
7	0.907	0.946	0.825	0.932	0.964	0.782	0.92
8	0.906	0.949	0.817	0.932	0.963	0.781	0.917
9	0.905	0.947	0.817	0.931	0.962	0.778	0.918
10	0.909	0.948	0.828	0.934	0.961	0.788	0.922

Table 2 shows GediNET's effective identification of relevant gene groups linked to different forms of cancer. Each association is validated by robust rank aggregation [44] p-values, showing a high level of statistical significance. The analysis demonstrates a significant genetic link to Glioblastoma Multiforme, as indicated by a very low p-value of 1.5994E-87. This relationship involves TNMD, CFH, GCLC, CFTR, and KRIT1 genes. These genes have the potential to play crucial roles in the pathogenesis of the disease, indicating prospective targets for new therapeutics. The gene group consisting of GCLC, CFTR, KRIT1, CD99, and MAD1L1, which is linked to the Malignant Neoplasm of the Stomach, shows a p-value of 2.77741E-86, indicating a strong association with the development of stomach cancer.

Additionally, Adenomatous Polyposis Coli and Leukemia are associated with gene sets that consist of TNMD, MAD1L1, TAC1, PAFAH1B1, and CFTR, BAD, CD99, CASP10, respectively, both demonstrating remarkably low p-values (1.05524E-85 and 3.49327E-85). These findings confirm the effectiveness of GediNET in revealing significant connections between diseases and genes and provide opportunities for further investigation into the functions of these genes in the genesis and progression of cancer. An in-depth analysis is essential for furthering our comprehension of oncogenic pathways and improving the precision of cancer therapy techniques.

Table 2: The results obtained using the RobustRankAggreg algorithm in the GediNET Tool

Group	p-value	List of genes
GLIOBLASTOMA MULTIFORME	1.5994E-87	TNMD, CFH, GCLC, CFTR, KRIT1, …
MALIGNANT NEOPLASM OF STOMACH	2.77741E-86	GCLC, CFTR, KRIT1, CD99, MAD1L1, …
ADENOMATOUS POLYPOSIS COLI	1.05524E-85	TNMD, MAD1L1, TAC1, PAFAH1B1
LEUKEMIA	3.49327E-85	CFTR, BAD, CD99, CASP10, …

Table 3 displays a thorough examination of different techniques for selecting features, assessed using many performance indicators, in the specific context of categorizing breast cancer subtypes. The compared approaches consist of Extreme Gradient Boosting (XGB), Information Gain (IG), SelectKBest (SKB), Conditional Mutual Information Maximization (CMIM), Fast Correlation Based Filter (FCBF), Minimum Redundancy Maximum Relevance (mRmR), and GediNET. The evaluated metrics are Accuracy, Sensitivity, Specificity, F-measure, Precision, and Area Under the Curve (AUC).

Concerning XGBoost and GediNET, both approaches demonstrate exceptional performance in all measures, with XGB getting almost the greatest scores in Accuracy, Sensitivity, Specificity, and F-measure, closely followed or matched by GediNET. Both approaches demonstrate exceptional proficiency in accurately distinguishing between cancer subtypes, as seen by their high AUC ratings (0.99). These methods are particularly effective in managing the complex data structures commonly encountered in genomic data. This is likely because they have algorithmic solid underpinnings that take use of both the relevance of individual features and the relationships between groups.

Regarding Information Gain (IG) and SelectKBest (SKB), although these approaches exhibit strong performance and consistent results, they are somewhat surpassed by XGB and GediNET. Their performance demonstrates a commendable equilibrium between sensitivity and specificity, showcasing a robust capacity to detect pertinent characteristics that help to precise classification of cancer subtypes.

Concerning Conditional Mutual Information Maximization (CMIM) and Fast Correlation Based Filter (FCBF), these approaches exhibit intermediate performance with lower scores in comparison to XGB, IG, SKB, and GediNET, specifically in terms of Accuracy and F-measure. The Minimum Redundancy Maximum Relevance (mRmR) method demonstrates the poorest performance compared to the other methods examined, with notably lower scores in all parameters. This suggests that mRmR may have difficulties processing the complex and diverse data commonly encountered in gene expression data for different forms of cancer. Its method to decrease repetition may excessively simplify the range of features, perhaps leaving out vital information required for precise categorization.

To summarize, the data presented in Table 3 highlights the significance of selecting a suitable feature selection technique that aligns with the unique attributes of the dataset and the intended objectives of the research. The exceptional efficacy of XGB and GediNET implies that techniques that successfully integrate intricate relationships between characteristics and utilize machine learning algorithms capable of handling extensive datasets with sophisticated patterns are especially well-suited for genomic data processing. This understanding is crucial for advancing diagnostic tools and customized medicine techniques in oncology since the accurate categorization of cancer subtypes is imperative for efficient treatment strategizing.

Table 3: Evaluating the Performance of Different Feature Selection Methods

Feature Selection Method	Accuracy	Sensitivity	Specificity	F-measure	Precision	Area Under Curve
XGB	0.97	0.98	0.96	0.98	0.97	0.99
IG	0.95	0.97	0.95	0.96	0.96	0.98
SKB	0.93	0.95	0.93	0.94	0.94	0.96
CMIM	0.7	0.73	0.71	0.72	0.71	0.75
FCBF	0.65	0.67	0.64	0.66	0.65	0.68
mRmR	0.5	0.52	0.49	0.5	0.51	0.53
GediNET	0.98	0.99	0.97	0.98	0.98	0.99

Table 4: Evaluation of Machine Learning Methods Paired with Feature Selection Techniques: A Performance Comparison Using Accuracy, Sensitivity, Specificity, F-measure, Precision, and AUC Metrics.

ML method	FS method	Accuracy	Sensitivity	Specificity	F-measure	Precision	Area Under Curve
RF	FCBF	0.66	0.76	0.57	0.66	0.67	0.72
LogitBoost	FCBF	0.59	0.85	0.34	0.66	0.55	0.70
LogitBoost	IG	0.96	0.97	0.947	0.95	0.94	0.99
Adaboost	MRMR	0.51	0.32	0.68	0.46	0.44	0.50
RF	CMIM	0.56	0.47	0.66	0.50	0.56	0.62
RF	GediNET	0.98	0.99	0.96	0.99	0.98	0.99

Table 4 provides the performance of various machine learning (ML) methods paired with different feature selection (FS) techniques, measured across several metrics on the BRCA LumA_LumBHer2Basal Dataset.

Concerning Random Forest with FCBF, this combination has moderate performance with an accuracy of 0.66. The sensitivity is relatively high at 0.76, suggesting it can identify the positive cases well. However, its specificity is low at 0.57, indicating a weakness in correctly identifying negative cases. The F-measure and precision are consistent with accuracy, and the AUC of 0.72 suggests moderate discriminative ability.

Concerning LogitBoost with FCBF, the accuracy is low at 0.59. Still, it has a high sensitivity of 0.85, which means it identifies most positive cases, albeit at the cost of a low specificity of 0.34, indicating many false positives. The F-measure and precision are not in line with the high sensitivity, and the AUC of 0.70 is modest.

Pairing LogitBoost with IG, this method shows excellent performance across all metrics with an accuracy of 0.96. High sensitivity at 0.97 and a specificity of 0.947 indicate a strong balance in identifying positive and negative cases. The F-measure and precision are also high, reflecting a balanced precision-recall trade-off. An AUC of 0.99 shows outstanding discriminative ability.

The Adaboost with MRMR combination demonstrates underwhelming performance with only a 0.51 accuracy and a sensitivity of 0.32, which is relatively low, indicating difficulty in correctly identifying true positives. The specificity is somewhat better at 0.68, but the ability to distinguish between classes is no better than a random chance, as shown by the AUC of 0.50.

Similarly, the Random Forest paired with CMIM also shows lackluster results, with an accuracy of 0.56 and a low sensitivity of 0.47. Though it performs slightly better in specificity at 0.66, the F-measure and precision are moderate, and the AUC of 0.62 points to a relatively weak discriminative power. Both methods suggest significant room for improvement in effectively classifying the given dataset.

This combination of Random Forest with GediNET is highly effective, with an excellent accuracy of 0.98. It achieves almost perfect sensitivity at 0.99 and strong specificity at 0.96, indicating it can distinguish very effectively between positive and negative cases. The F-measure and precision are at 0.99 and 0.98, respectively, supporting the high accuracy. The AUC is also at an impressive 0.99, indicating superior discriminative ability.

Table 4 shows that the Random Forest with GediNET and LogitBoost with IG combinations show exceptionally high performance on the *BRCA LumAB_Her2Basal* dataset. In contrast, the Adaboost with MRMR and the LogitBoost with FCBF methods show considerable weaknesses, reflected in their lower metrics across the board. These results indicate the importance of the right pairing between feature selection methods and machine learning algorithms to achieve optimal performance.

Table 5: AUC Performance metrics of Adaboost, Decision Tree, Logitboost, and Random Forest Models Across Various Feature Selection Methods for the BRCA LumAB_Her2Basal Dataset.

Model	SKB	IG	XGB	FCBF	MRMR	CMIM
Adaboost	0.99	0.99	0.99	0.68	0.50	0.58
DT	0.96	0.93	0.944	0.49	0.50	0.50
Logitboost	0.99	0.99	0.99	0.70	0.50	0.58
RF	0.99	0.99	0.99	0.72	0.47	0.62

Table 5 compares the AUC performance metric for several models using different feature selection methods on the BRCA LumAB_Her2Basal dataset. Adaboost performs exceptionally well with SKB, IG, and XGB feature selection methods, achieving AUCs close to 0.99, indicating excellent discriminative ability. However, its performance drops significantly with FCBF, MRMR, and CMIM, with AUCs ranging from moderate to no better than random chance.

Decision Tree (DT) shows good discriminative power with SKB and XGB, with AUCs just above 0.94, but its effectiveness diminishes with IG and further drops with FCBF, MRMR, and CMIM, where AUCs are close to 0.50, suggesting poor performance.

Logitboost exhibits high AUCs with SKB, IG, and XGB, similar to Adaboost, reflecting strong classification capabilities. However, like Adaboost, it sees a decline in performance with FCBF and even lower AUCs with MRMR and CMIM.

Random Forest (RF) achieves outstanding AUC with XGB at nearly 1.0 and strong results with SKB and IG. Its AUC with FCBF is moderate and underperforms with MRMR and CMIM, though it still performs better than DT and Adaboost with these feature selection methods.

We can conclude that XGB, as a feature selection method, consistently leads to the highest AUC across all machine learning models, indicating a robust synergy.

SKB and IG also have strong AUC performance, particularly with Adaboost, Logitboost, and RF models. Conversely, FCBF, MRMR, and CMIM feature selection methods show significantly weaker discriminative power when paired with these ML models, with MRMR consistently at the lower end of the spectrum.

5. Discussion

This study comprehensively evaluates the effectiveness of the GediNET technique, which uses the Grouping-Scoring-Modeling (GSM) methodology, in categorizing different subtypes of breast cancer within the BRCA LumA_LumBHer2Basal dataset. GediNET distinguishes itself from conventional feature selection approaches by strategically emphasizing gene groupings rather than individual genes. This emphasis enables exploring intricate connections between diseases and discovering vital biomarkers necessary for developing precise treatments and individualized medicine.

GediNET's exceptional performance in terms of accuracy and the Area Under the Curve (AUC), as emphasized in our analysis, demonstrates its strength in managing the complex genomic data related to breast cancer. This approach utilizes pre-existing biological knowledge to categorize genes, resulting in improved data processing and increased biological significance of the predictions. In the context of breast cancer, the influence of molecular subtypes on treatment and prognosis is particularly crucial.

Although GediNET surpasses standard feature selection methods, it is important to acknowledge that incorporating these advanced approaches with conventional techniques can yield a more comprehensive study. For example, whereas approaches like as XGB and IG also demonstrated impressive performance, GediNET stands out due to its distinctive capability to include and analyze intricate biological data sets, providing a more profound understanding of genetic relationships and pathway involvements.

The practical consequences of these discoveries are significant. GediNET enhances the precision of breast cancer subtype classification, hence facilitating more accurate diagnostics. This is crucial for customizing treatment methods to suit the specific needs of each patient. This technique of precision medicine holds the potential to improve therapeutic results by focusing on treatments that have a high probability of being successful, based on the genetic characteristics of the tumor.

6. Conclusion

To summarize, this study emphasizes the considerable capacity of the GediNET approach to transform the area of cancer genomics by accurately categorizing different subtypes of breast cancer. Using sophisticated machine learning techniques and extensive biological expertise, GediNET surpasses conventional approaches and facilitates the development of enhanced, precise, and individualized cancer therapies. GediNET, a tool in the field of bioinformatics, will have a crucial role in converting intricate genomic data into practical clinical insights. This will ultimately improve patient outcomes in the field of oncology.

This work contributes to the continuing discussion in computational biology, highlighting the crucial importance of creative, analytical frameworks in the age of large-scale data and customized medicine. Therefore, this study's results can shape future research and clinical practices, significantly impacting the management and treatment of breast cancer.

Additional study is recommended to broaden the usefulness of the GediNET tool with other types of cancer and other disorders. The tool's predicted accuracy may be further validated and improved by including a more comprehensive range of genetic data sources. Furthermore, investigating the incorporation of GediNET with other artificial intelligence (AI) tools may result in the development of more resilient systems for disease diagnosis and prognosis.

References:

[1] H. P. J. Buermans and J. T. den Dunnen, "Next generation sequencing technology: Advances and applications," *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, vol. 1842, no. 10, pp. 1932–1941, Oct. 2014, doi: 10.1016/j.bbadis.2014.06.015.

[2] G. Parmigiani, E. S. Garrett, R. A. Irizarry, and S. L. Zeger, "The Analysis of Gene Expression Data: An Overview of Methods and Software," in *The Analysis of Gene Expression Data: Methods and Software*, G. Parmigiani, E. S. Garrett, R. A. Irizarry, and S. L. Zeger, Eds., New York, NY: Springer, 2003, pp. 1–45. doi: 10.1007/0-387-21679-0_1.

[3] "miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database | Nucleic Acids Research | Oxford Academic." Accessed: Nov. 30,2021. [Online]. Available: https://academic.oup.com/nar/article/44/D1/D239/2503072

[4] The Gene Ontology Consortium *et al.*, "The Gene Ontology knowledgebase in 2023," *Genetics*, vol. 224, no. 1, p. iyad031, May 2023, doi: 10.1093/genetics/iyad031.

[5] E. Clough and T. Barrett, "The Gene Expression Omnibus Database," *Methods Mol Biol*, vol. 1418, pp. 93–110, Jan. 2016, doi: 10.1007/978-1-4939-3578-9_5.

[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] 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.

[8] K. S. Johnson, E. F. Conant, and M. S. Soo, “Molecular Subtypes of Breast Cancer: A Review for Breast Radiologists,” *Journal of Breast Imaging*, vol. 3, no. 1, pp. 12–24, Jan. 2021, doi: 10.1093/jbi/wbaa110.

[9] 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.

[10] M. Oti and H. Brunner, “The modular nature of genetic diseases,” *Clinical Genetics*, vol. 71, no. 1, pp. 1–11, 2007, doi: 10.1111/j.1399-0004.2006.00708.x.

[11] 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, doi: 10.3390/genes10020087.

[12] F. Curion and F. J. Theis, “Machine learning integrative approaches to advance computational immunology,” *Genome Medicine*, vol. 16, 2024, doi: 10.1186/s13073-024-01350-3.

[13] A. Holzinger, R. Goebel, V. Palade, and M. Ferri, “Towards Integrative Machine Learning and Knowledge Extraction,” in *Towards Integrative Machine Learning and Knowledge Extraction*, A. Holzinger, R. Goebel, M. Ferri, and V. Palade, Eds., Cham: Springer International Publishing, 2017, pp. 1–12. doi: 10.1007/978-3-319-69775-8_1.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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, May 2021, doi: 10.7717/peerj.11458.

[19] 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.

[20] 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.

[21] 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. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fgene.2022.1076554

[22] 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, 2020, doi: 10.12688/f1000research.26880.2.

[23] 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, Accessed: Feb. 12, 2024. [Online]. Available: https://www.frontiersin.org/journals/genetics/articles/10.3389/fgene.2023.1093326

[24] 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.

[25] 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.

[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] 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, doi: 10.3389/fmicb.2023.1264941.

[28] 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.

[29] 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.

[30] 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.

[31] 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, doi: 10.1038/s41587-020-0546-8.

[32] 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.

[33] Q.-S. Xu and Y.-Z. Liang, "Monte Carlo cross validation," *Chemometrics and Intelligent Laboratory Systems*, vol. 56, no. 1, pp. 1–11, Apr. 2001, doi: 10.1016/S0169-7439(00)00122-2.

[34] 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, doi: 10.1109/TPAMI.2005.159.

[35] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," *J. Bioinform. Comput. Biol.*, vol. 03, no. 02, pp. 185–205, Apr. 2005, doi: 10.1142/S0219720005001004.

[36] 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.

[37] 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.

[38] 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.

[39] G. Ke *et al.*, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2017. Accessed: Oct. 25, 2023. [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html

[40] T. K. Ho, "Random decision forests," in *Proceedings of 3rd International Conference on Document Analysis and Recognition*, Aug. 1995, pp. 278–282 vol.1. doi: 10.1109/ICDAR.1995.598994.

[41] M. Yousef, M. Ketany, L. Manevitz, L. C. Showe, and M. K. Showe, "Classification and biomarker identification using gene network modules and support vector machines," *BMC Bioinformatics*, vol. 10, no. 1, p. 337, Oct. 2009, doi: 10.1186/1471-2105-10-337.

[42] M. H. Kamarudin, C. Maple, T. Watson, and N. S. Safa, "A LogitBoost-Based Algorithm for Detecting Known and Unknown Web Attacks," *IEEE Access*, vol. 5, pp. 26190–26200, 2017, doi: 10.1109/ACCESS.2017.2766844.

[43] R. Wang, "AdaBoost for Feature Selection, Classification and Its Relation with SVM, A Review," *Physics Procedia*, vol. 25, pp. 800–807, Jan. 2012, doi: 10.1016/j.phpro.2012.03.160.

[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, doi: 10.1093/bioinformatics/btr709.