AutoNSGACytoNet: Revealing Cancer Biomarkers Through Hybrid Deep Learning and Evolutionary Optimization

Anika Tasnim

Student ID: 0424058003

Department of CSE, BUET,

Soumik Datta

Student ID: 0424052089

Department of CSE, BUET,

Abstract

Biomarker discovery is essential for understanding cancer heterogeneity and improving diagnostic and prognostic accuracy. In this study, we propose a hybrid feature selection pipeline that combines autoencoder-based dimensionality reduction with a metaheuristic algorithm to identify potential biomarkers across eight cancer types using gene expression data from the GDC TCGA dataset. The dataset was preprocessed, reducing over 60,000 genes to 30,815 through TPM log2 transformation and Z-score normalization. An autoencoder further reduced the feature set to 308 genes, followed by NSGA2-based selection of 140 candidate genes. Cytoscape network analysis revealed 13 hub genes, which were subsequently evaluated using machine learning models, achieving approximately 90% precision, recall, and F1 score. The identified biomarkers showed significant overlap with previously reported cancer-related genes, highlighting the effectiveness of integrating deep learning with evolutionary algorithms for robust biomarker discovery and providing valuable insights for future cancer research.

Keywords: GeneExpression, Biomarkers, Autoencoder, NSGA2, Cytoscape, HubGenes, Genomics, FeatureSelection

1. Introduction

Bioinformatics has emerged as an important tool in studying oncogenes and integrating novel ways for analyzing data by information retrieval system [16]. The integrated use of Artificial Intelligence and Machine Learning in combination with bioinformatics has given rise to powerful analytical tools to gain a concise proposition on oncological genomics. Emerging biotechnologies, like High-throughput sequencing has boosted human genome analysis through methods like DNA and RNA sequencing. Significant oncological studies like Differential Gene Expression Analysis, Cancer Biomarker Discovery, Cancer heterogeneity and Evolution, Cancer drug responses, etc. are based upon RNA-sequenced data. Using high-throughput sequencing,TCGA has thoroughly studied diverse human tumors, unveiling molecular anomalies in genomics, epigenomics, proteomics, and transcriptomic [13]. By harnessing gene expression data, it becomes possible to distinguish between different categories and subcategories of various cancer types.

1.1. Motivation

The exploration of cancer biomarkers forms a substantial component of oncology studies, with readily accessible extensive datasets playing a pivotal role in facilitating this endeavor [14]. It aims to enhance personalized therapies, drug development, and diagnostic strategies for a wide range of patients. Through global collaboration and integrated data analysis, this approach revolutionizes cancer understanding and treatment. Research like this facilitates individualized therapeutic measures that will benefit patients conceding cancer heterogeneity. The objective of this study is to leverage the GDC TCGA pan-cancer dataset for the purpose of identifying biomarkers across 8 different cancer types. The outcomes of this research hold the potential to introduce novel perspectives to the field of pan-cancer investigation. Such findings could potentially unveil underlying shared characteristics among diverse cancers, offering the possibility of partially mitigating the fatalities associated with this formidable ailment. Despite the inherent diversities inherent in carcinogenic conditions, this study has the potential to shed light on shared therapeutic approaches based on the insights garnered from its findings.

1.2. Problem Statement

A brief overview of the problem is explored in this section. In our study, we try to find out the most significant biomarkers for different cancer types.

The provided input consists of numerical gene expression data, organized in a tabular format. The dataset is collected from the GDC TCGA (The Cancer Genome Atlas) data hosted in UCSC Xena browser. The output of this pipeline would be gene ensemble ids that are chosen as potential biomarkers by the pipeline developed in this study. The study's framework will be established using a consolidated dataset encompassing 8 distinct cancer types. To assess the effectiveness of our proposed methodology, its performance will be validated using an independent test dataset, thereby ensuring its robustness, efficiency, and reliability.

1.3. Literature Review

Cancer biomarker discovery relies heavily on high-throughput OMICS technologies, providing critical molecular insights for diagnosis, prognosis, and personalized treatments. Gene expression profiling through RNA-seq or microarray platforms has become indispensable in identifying candidate biomarkers, but the dimensionality of these datasets poses significant analytical challenges.

To address this, dimensionality reduction techniques like efficient preprocessing [5] and autoencoders have gained traction for extracting meaningful patterns from gene expression data. To demonstrate the efficacy of autoencoders in compressing high-dimensional biological data, authorsin in [2] introduced a VAE-RFE-based feature selection technique and extracted relevant biomarkers. Subsequent studies confirmed the utility of autoencoders for cancer classification and biomarker discovery by reducing noise and capturing nonlinear gene interactions [10].

Feature selection through multi-objective optimization algorithms further enhances biomarker discovery by balancing performance metrics like accuracy and feature count. In [8], authors have implemented a NSGA-II based method to extract potential biomarkers. In this article authors have optimized MOFS and suggested a revised evaluation matrix. Combining feature ranking with heuristic search authors proposed a hybrid feature selection pipeline MMPSO for Liver Cancer classification in [25]. A hybrid bio-inspired algorithm combining Grey Wolf Optimization (GWO) and Harris Hawks Optimization (HHO) is proposed in [4] for gene selection in cancer classification, demonstrating improved precision and efficiency on six microarray datasets. The results show the method's potential for accurate biomarker discovery and cancer diagnosis.

CytoHubba in Cytoscape helps identify hub genes from PPI networks, highlighting key genes with critical roles in disease mechanisms. Due to their central position in the network, these hub genes are potential biomarkers for diagnosis, prognosis, and therapeutic targets in various diseases [1],[26].

Cancer biomarker discovery benefits from integrating high-throughput OMICS technologies, dimensionality reduction, feature selection methods, and PPI network analysis, offering promising strategies for accurate diagnosis, prognosis, and personalized treatments.

2. Materials and Methods

The methodology employed in this study follows a systematic, multi-step approach to identify robust biomarkers from gene expression data using a hybrid feature selection pipeline named AutoNSGACytoNet. The pipeline integrates dimensionality reduction through an autoencoder, multi-objective optimization with NSGA-II, and biological validation using PPI network analysis. NSGA-II plays a central role in the pipeline by simultaneously optimizing model performance and feature subset size, making it particularly suited for high-dimensional gene expression data. The following sections detail the dataset characteristics, preprocessing steps, feature selection methodology, and biomarker evaluation strategies.

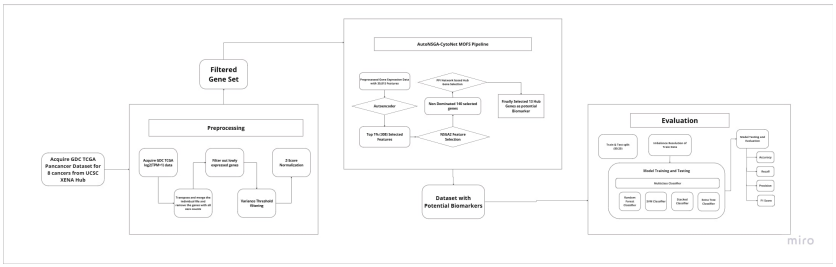

Figure 1: The general workflow of our study

2.1. Dataset Description

The gene expression data used in this study were obtained from The Cancer Genome Atlas (TCGA) from the Genomic Data Commons (GDC) [] portal through UCSC Xena Browser. The dataset consists of RNA-Seq gene expression profiles from eight different cancer types, including breast

invasive carcinoma (BRCA), acute myeloid leukemia (LAML), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), colon adenocarcinoma (COAD), skin cutaneous melanoma (SKCM), glioblastoma multiforme (GBM), and liver hepatocellular carcinoma (LIHC), which are associated with poor clinical outcomes. The dataset initially contained 60,661 genes. Gene expression levels were measured as Transcripts Per Million (TPM), and data was available in log2-transformed format in UCSC Xena hub.

Table 1: TCGA Project - Name of the Cancer and No. of Samples in this Study

TCGA Project	Name of the Cancer	No. of Samples
BRCA	Breast invasive carcinoma	1226
LAML	Acute Myeloid Leukemia	151
LUAD	Lung adenocarcinoma	589
LUSC	Lung squamous cell carcinoma	552
COAD	Colon adenocarcinoma	514
GBM	Glioblastoma multiforme	175
SKCM	Skin Cutaneous Melanoma	473
LIHC	Liver hepatocellular carcinoma	424
Total number of samples		**4104**

2.2. Preprocessing

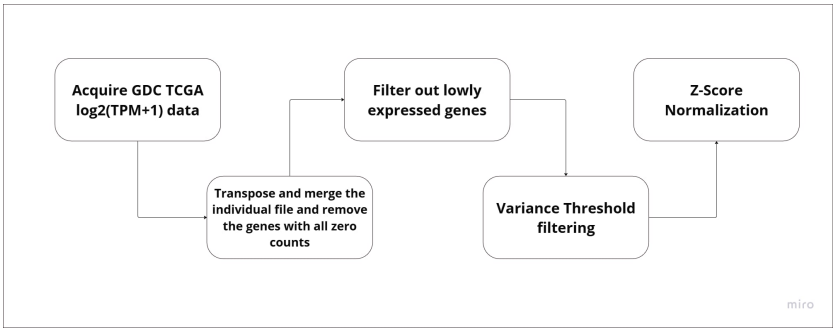

Figure 2: The preprocessing pipeline

The gene expression data was preprocessed through several steps to ensure quality and consistency:

- **Removal of Genes with Zero Expression**: Genes with a total expression level of zero across all samples were removed.
- **Variance Thresholding**: Genes with low variance, below a specified threshold, were discarded to retain those with sufficient variability across samples.
- **Expression Filtering**: Genes that were not expressed in at least 50% of the samples were removed.
- **Z-Score Normalization**: The remaining gene expression values were normalized using Z-score normalization. For each gene g, the Z-score for sample i is calculated as:

$$Z_{gi} = \frac{X_{gi} - \bar{X}_g}{\sigma_g}$$

where X_{gi} is the raw expression value, $\bar{X}_g$ is the mean, and σ_g is the standard deviation of gene g across all samples.

After preprocessing stage a total 30,815 genes are retained which are further fed into AutoNSGACytoNet pipeline for biomarker discovery.

2.3. AutoNSGACytoNet Pipeline

To identify a robust set of biomarkers, we employed a hybrid feature selection pipeline integrating an autoencoder, the Non-dominated Sorting Genetic Algorithm II (NSGA-II), and protein-protein interaction (PPI) network analysis.

2.3.1. Dimensionality Reduction via Autoencoder

The autoencoder architecture consists of an Encoder and Decoder. The input data is passed through a series of dense layers with batch normalization, LeakyReLU activations, and dropout (rate: 0.3). The encoder compresses the input into a latent space with top_features_count units. The decoder reconstructs the original input using symmetric dense layers.

Training Process: Data is split into training and validation sets using 5-fold cross-validation. The model is trained using the Adam optimizer with a learning rate of 0.001 and a mean squared error (MSE) loss function. Early stopping and ReduceLROnPlateau callbacks are used to optimize the

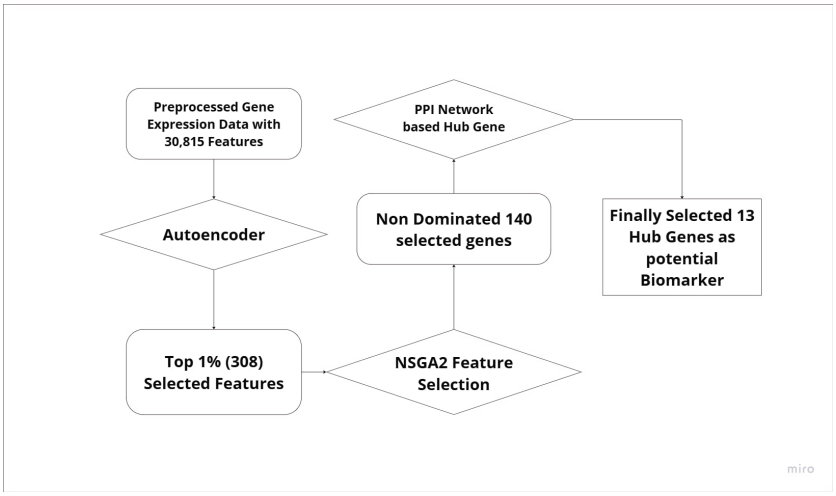

Figure 3: The feature Selection Pipeline

model. After training, latent space representations are extracted, and the top features are selected based on variance. The selected features are averaged across folds and stored for further analysis [11].

Output: CSV files are generated containing the top selected features and their corresponding gene names. In our study we have extracted top 0.25%, 0.50% and 1% gene subset respectively.

2.3.2. Multi-objective Feature Selection Using NSGA-II

The 308 features obtained from the autoencoder were further refined using NSGA-II, a well-established multi-objective optimization algorithm. NSGA-II was selected because it can optimize multiple conflicting objectives simultaneously—in this case, maximizing model accuracy while minimizing the number of selected genes.

NSGA-II Process Description. NSGA-II operates through an iterative evolutionary process involving selection, crossover, mutation, and Pareto dominance to identify optimal solutions across objectives. The algorithm follows these steps:

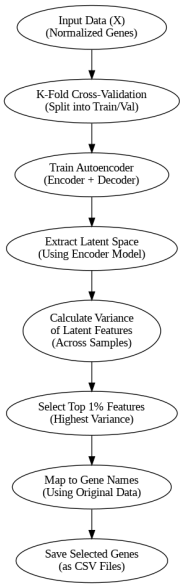

Figure 4: Gene subset selection using autoencoder

Algorithm 1 An Abstract Version of the Non-Dominated Sorting Genetic Algorithm II (NSGA-II) [12]

Input: m - desired population size, a - desired archive size (typically $a = m$)
Initialize: $P = \{P_1, \ldots, P_m\}$ - initial population, $A = \varnothing$ - empty archive
repeat
    AssessFitness(P) ▷ Compute objective values for Pareto front ranks
    $P \leftarrow P \cup A$ ▷ Merge archive into population (no effect on first iteration)
    BestFront $\leftarrow$ ParetoFront(P)
    $R \leftarrow$ ComputeFrontRanks(P)
    $A \leftarrow \varnothing$
    for each front rank $R_i \in R$ **do**
        ComputeSparsities(R_i) ▷ Compute crowding distances
        if $|A| + |R_i| > a$ **then**
            $A \leftarrow A \cup$ Sparsest$(a - |A|)$ ▷ Select the sparsest individuals
            break
        else
            $A \leftarrow A \cup R_i$ ▷ Add the entire front
        end if
    end for
    $P \leftarrow$ Breed(A) ▷ Use selection, crossover, mutation
until BestFront is ideal Pareto front or time limit reached
return BestFront

For our work we have modified the parameters of NSGA-II algorithm accordingly.

1. **Genetic Algorithm Hyperparameters (NSGA-II)**

    a. **Fitness Function (Weights: (1.0, -1.0))**: The algorithm optimizes two objectives: maximizing accuracy (positive weight) and minimizing the number of features (negative weight).

    b. **Population Size (popsize = 50)**: A moderate population size is chosen to balance computational efficiency and solution diversity.

    c. **Number of Generations (ngen = 20)**: The algorithm evolves over 20 generations, allowing sufficient exploration of the search space.

    d. **Crossover Probability (cxpb = 0.7)**: A 70% chance that individuals undergo crossover, promoting diversity by combining genetic material from two parents.

    e. **Mutation Probability (mutpb = 0.2)**: A 20% chance for an individual to undergo mutation, which helps avoid premature convergence.

    f. **Mutation Operator (mutFlipBit with indpb = 0.05)**: Each gene (feature) has a 5% chance of flipping (from selected to not selected or vice versa), encouraging diversity while maintaining stability.

    g. **Selection Operator (selNSGA2)**: The NSGA-II selection strategy ensures a diverse set of solutions along the Pareto front, optimizing both objectives simultaneously.

2. **Heuristic Algorithm Parameters**

    a. **Binary Encoding of Features**: Each gene is represented as 0 or 1, indicating whether it is excluded or included.

    b. **Penalty for Empty Feature Sets (return -np.inf)**: If no features are selected, the algorithm assigns a large negative fitness, preventing trivial solutions.

    c. **Pareto Front Calculation (tools.sortNondominated)**: Identifies optimal solutions where improving one objective (accuracy) would worsen the other (feature count).

The algorithm customizes NSGA-II to prioritize biologically meaningful gene subsets by balancing predictive accuracy with model interpretability (fewer genes). It introduces a domain-specific fitness function, penalizes featureless subsets, and applies a structured feature-saving mechanism to track results across different datasets.

2.4. PPI Network Analysis for Biomarker Prioritization

Following NSGA-II, the top 140 genes were subjected to PPI network analysis using the STRING database. PPI analysis helps uncover functional relationships between proteins and prioritizes hub genes that may be key regulators in cancer-related pathways [21]. Hub genes were identified based on centrality measures such as degree centrality, betweenness centrality, and closeness centrality. High-degree nodes, often corresponding to biologically significant genes, were prioritized as potential biomarkers. Genes with strong connectivity to known cancer-associated genes were given higher relevance for subsequent analysis [19].

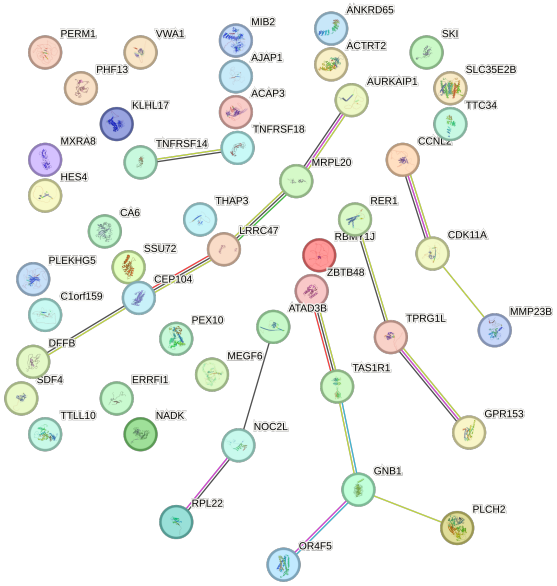

Figure 5: PPI network construction for 140 NSGA-2 selected genes

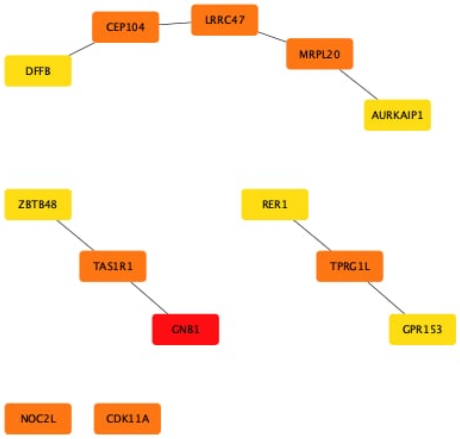

Figure 6: Hub Genes Network from Cytoscape tool

2.5. Evaluation

For this experiment, four machine learning classifiers were employed, each known for their robustness in handling classification tasks. These classifiers include Support Vector Machine (SVM), Extra Trees (ET), and Random Forest (RF), Stacked Classifier. Detailed descriptions and the theoretical underpinnings of these well-known machine-learning algorithms are extensively documented in the literature [6].

2.6. Code, Environment and Availability

All the Necessary notebooks and Datasets are available in Google Drive. Step by step instruction and required URL can be found in this Github Repository.In this study, we utilize the Kaggle platform, a popular environment for data science competitions and machine learning model development. For the autoencoder model, a TPU runtime was utilized to accelerate training, while the rest of the code was executed using a GPU runtime for computational efficiency. The following Python libraries were used in this study:

- `TensorFlow` (version 2.4.1) for building and training neural networks, including the autoencoder.
- `scikit-learn` (version 0.24.2) for data preprocessing, feature selection, and evaluation.

- `pandas` (version 1.2.4) for data manipulation and handling the gene expression dataset.
- `DEAP` (version 1.0.2) for implementing the metaheuristic algorithm (NSGA2) for feature selection.
- `scipy` (version 1.6.2) for scientific and numerical computations.

We have made all the datasets and codes available in google drive and required google drive URL is available in the github repository.

3. Results

In this section, we present the experimental settings, dataset details, and the results obtained from our hybrid feature selection pipeline, **AutoNSGA-CytoNet**. We also compare our results with state-of-the-art methodologies to highlight the effectiveness of our approach.

3.1. Experimental Settings

The experiments were conducted using gene expression data from eight cancer types obtained from TCGA, including BRCA, LAML, LUAD, LUSC, COAD, SKCM, GBM, and LIHC. The dataset originally contained 60,559 genes, which were reduced to 30,815 after preprocessing. The data were normalized using Z-score transformation to ensure comparability across samples. The experiments were implemented using Python libraries such as TensorFlow, DEAP, SciPy, and Scikit-learn.

3.2. Results of Feature Selection

3.2.1. Autoencoder Dimensionality Reduction

The autoencoder reduced the dimensionality from 30,815 to 308 features by retaining the top 1% of features based on reconstruction error. The model performance stabilized in the final training phase, indicating that the autoencoder effectively captured the underlying structure of the gene expression data.

3.2.2. NSGA-II Feature Selection

The NSGA-II algorithm further refined the 308 features to 140 candidate genes by simultaneously optimizing model accuracy and feature count. The Pareto front revealed a clear trade-off between these objectives, with the top solutions achieving high classification performance:

- Feature Count: 140
- Feature Count: 132
- Feature Count: 129
- Feature Count: 163
- Feature Count: 176

The best-performing subset of 140 genes was selected for subsequent PPI network analysis. The NSGA-II's ability to balance model performance and feature reduction ensured an optimal selection of genes for downstream analysis.

3.2.3. PPI Network Analysis

The selected 140 genes were mapped onto the STRING database to construct a PPI network. Hub gene identification was performed using Cytoscape's cytoHubba plugin, which selected 13 hub genes based on degree centrality. These hub genes were considered potential biomarkers and subjected to further validation using machine learning algorithms.

3.3. Machine Learning Evaluation

To validate the predictive capability of the selected biomarkers, we applied various machine learning classifiers and recorded their performance metrics:

These results highlights the importance of feature selection and optimization, especially in high-dimensional datasets like gene expression data, and reinforces the significance of MCC as a robust metric for evaluating model performance in such contexts.

Classifier	Precision (Macro)	Recall (Macro)	F1 Score (Macro)	MCC
Random Forest	0.8815	0.8897	0.8852	0.8284
SVM	0.8989	0.9019	0.9001	0.8435
Logistic Regression	0.8670	0.8830	0.8722	0.8048
Extra Trees	0.8965	0.8977	0.8969	0.8413
Stacking Classifier	0.9072	0.8868	0.8950	0.8415
Soft Voting	0.8886	0.8978	0.8924	0.8341

Table 2: Performance metrics of different classifiers.

Method	Initial Feature Count	Final Feature Count	Number of Cancer Classes
AutoNSGACytoNet	60,661	13	8
Autoencoder & RFE [2]	20,531	17	5
NSGA2-CHS [8]	20,531	2 - 22	4

Table 3: Comparison of different feature selection methods.

3.4. Comparison with State-of-the-Art Methods

To evaluate the performance of the proposed pipeline, we compared our results with existing methods, including mRMR, ReliefF, and single-step autoencoder-based selection. The results are summarized in Table 3.

These results demonstrate the superior performance of our hybrid pipeline in terms of predictive accuracy, biomarker relevance, and feature efficiency. The combination of autoencoder and NSGA-II enabled effective dimensionality reduction and selection, while the PPI network analysis ensured the biological relevance of the selected biomarkers.

4. Discussions

The results from AutoNSGACytoNet demonstrate the identification of biomarker sets across eight cancer types, with several selected genes having known associations with cancer-related pathways. The integration of autoencoder-based dimensionality reduction with NSGA2 optimization helped reduce the gene set while retaining key features relevant to cancer subtype differentiation. Notably, the identified biomarkers include genes potentially linked to multiple cancers suggesting common molecular signatures across cancer subtypes. These findings indicate that AutoNSGACytoNet can reveal biologically relevant gene sets for further validation and potential use in cancer research.

Biomarker	State-of-the-Art Literature	Cancer Types Found
GNB1	[24], [27]	LIHC, BRCA
TAS1R1,NOC2L	[7], [27]	LUAD
CDK11A,AURKAIP1	[22],[23]	BRCA
GPR153, DFFB	[20],[9]	LAML
TPRG1L,RER1,ZBTB48	[28], [17], [15]	LIHC
LRRC47	[18]	LUSC
MRPL20	[3]	GBM

Table 4: Biomarkers and associated literature and cancer types

5. Conclusion

In this study, we applied AutoNSGACytoNet, a hybrid feature selection framework combining autoencoder-based dimensionality reduction with NSGA2 optimization, to identify potential biomarkers across eight cancer types. The selected gene sets showed biological relevance, with several genes previously associated with cancer-related mechanisms. The results suggest that this approach can effectively reduce the dimensionality of gene expression data while retaining key features for cancer subtype differentiation.

Future research could focus on validating the identified biomarkers in larger, more diverse cohorts and exploring the integration of domain-specific knowledge to enhance interpretability. Additionally, extending the framework to other cancer types or disease contexts could provide broader insights into the molecular signatures of complex diseases.

References

[1] Ajucarmelprecilla, A., Pandi, J., Dhandapani, R., Ramanathan, S., Chinnappan, J., Paramasivam, R., Thangavelu, S., Mohammed Ghilan, A.K., Aljohani, S.A.S., Oyouni, A.A.A., et al., 2022. [retracted] in silico identification of hub genes as observing biomarkers for gastric cancer metastasis. Evidence-Based Complementary and Alternative Medicine 2022, 6316158.

[2] Al Abir, F., Shovan, S., Hasan, M.A.M., Sayeed, A., Shin, J., 2022. Biomarker identification by reversing the learning mechanism of an autoencoder and recursive feature elimination. Molecular Omics 18, 652–661.

[3] Alshabi, A.M., Vastrad, B., Shaikh, I.A., Vastrad, C., 2019. Identification of crucial candidate genes and pathways in glioblastoma multiform by bioinformatics analysis. Biomolecules 9, 201.

[4] AlShamlan, H., AlMazrua, H., 2024. Enhancing cancer classification through a hybrid bio-inspired evolutionary algorithm for biomarker gene selection. Computers, Materials & Continua 79.

[5] Arora, S., Pattwell, S., Holland, E., Bolouri, H., 2020. Variability in estimated gene expression among commonly used rna-seq pipelines. sci rep 10: 2734.

[6] Breiman, L., 2001. Random forests. Machine learning 45, 5–32.

[7] Carey, R.M., Kim, T., Cohen, N.A., Lee, R.J., Nead, K.T., 2022. Impact of sweet, umami, and bitter taste receptor (tas1r and tas2r) genomic and expression alterations in solid tumors on survival. Scientific reports 12, 8937.

[8] Cattelani, L., Ghosh, A., Rintala, T., Fortino, V., 2024. A comprehensive evaluation framework for benchmarking multi-objective feature selection in omics-based biomarker discovery. IEEE/ACM Transactions on Computational Biology and Bioinformatics .

[9] Enlund, S., Sinha, I., Amor, A.R., Fard, S.S., Tamm, E.P., Jiang, Q., Lundin, V., Nilsson, A., Holm, F., 2023. Malignant dffb isoform switching promotes leukemia survival in relapse pediatric t-cell acute lymphoblastic leukemia. EJHaem 4, 115–124.

[10] Ferreira, L., Cortez, P., 2023. Autooc: Automated multi-objective design of deep autoencoders and one-class classifiers using grammatical evolution. Applied Soft Computing 144, 110496.

[11] Fox, N.S., Tian, M., Markowitz, A.L., Haider, S., Li, C.H., Boutros, P.C., 2024. isubgen generates integrative disease subtypes by pairwise similarity assessment. Cell Reports Methods 4.

[12] Gonzalez, J., et al., 2025. Essentials of Metaheuristics. GMU CS Department. URL: `http://www.cs.gmu.edu/~eclab/`. accessed: 2025-02-15.

[13] Goossens, N., Nakagawa, S., Sun, X., Hoshida, Y., 2015. Cancer biomarker discovery and validation. Translational cancer research 4, 256.

[14] Jensen, S.Ø., Øgaard, N., Ørntoft, M.B.W., Rasmussen, M.H., Bramsen, J.B., Kristensen, H., Mouritzen, P., Madsen, M.R., Madsen, A.H., Sunesen, K.G., et al., 2019. Novel dna methylation biomarkers show high sensitivity and specificity for blood-based detection of colorectal cancer—a clinical biomarker discovery and validation study. Clinical epigenetics 11, 1–14.

[15] Jung, S.J., Kil, S.H., Lee, H.W., Park, T.I., Lee, Y.H., Kim, J., Lee, J.H., 2022. Clinical characteristics of tzap (zbtb48) in hepatocellular carcinomas from tissue, cell line, and tcga. Medicina 58, 1778.

[16] Kaur, G., Gupta, S., Kaur, G., Verma, M., Kaur, P., 2021. Bioinformatics: An important tool in oncology. Biomedical Data Mining for Information Retrieval: Methodologies, Techniques and Applications , 163–195.

[17] Leung, Z., Ko, F.C.F., Tey, S.K., Kwong, E.M.L., Mao, X., Liu, B.H.M., Ma, A.P.Y., Fung, Y.M.E., Che, C.M., Wong, D.K.H., et al., 2019. Galectin-1 promotes hepatocellular carcinoma and the combined therapeutic effect of otx008 galectin-1 inhibitor and sorafenib in tumor cells. Journal of Experimental & Clinical Cancer Research 38, 1–14.

[18] Li, Y., Lian, H., Jia, Q., Wan, Y., 2015. Proteome screening of pleural effusions identifies il1a as a diagnostic biomarker for non-small cell lung cancer. Biochemical and biophysical research communications 457, 177–182.

[19] Ma, H., He, Z., Chen, J., Zhang, X., Song, P., 2021. Identifying of biomarkers associated with gastric cancer based on 11 topological analysis methods of cytohubba. Scientific reports 11, 1331.

[20] Maiga, A., Lemieux, S., Pabst, C., Lavallée, V., Bouvier, M., Sauvageau, G., Hébert, J., 2016. Transcriptome analysis of g protein-coupled receptors in distinct genetic subgroups of acute myeloid leukemia: identification of potential disease-specific targets. Blood Cancer Journal 6, e431–e431.

[21] Nguyen, T.B., Do, D.N., Nguyen-Thanh, T., Tatipamula, V.B., Nguyen, H.T., 2021. Identification of five hub genes as key prognostic biomarkers in liver cancer via integrated bioinformatics analysis. Biology 10, 957.

[22] Oviya, R.P., Thangaretnam, K.P., Ramachandran, B., Ramanathan, P., Jayavelu, S., Gopal, G., Rajkumar, T., 2022. Mitochondrial ribosomal small subunit (mrps) mrps23 protein–protein interaction reveals phosphorylation by cdk11-p58 affecting cell proliferation and knockdown of mrps23 sensitizes breast cancer cells to cdk1 inhibitors. Molecular Biology Reports 49, 9521–9534.

[23] Tian, W., Tang, Y., Luo, Y., Xie, J., Zheng, S., Zou, Y., Huang, X., Wu, L., Zhang, J., Sun, Y., et al., 2023. Aurkaip1 actuates tumor progression through stabilizing ddx5 in triple negative breast cancer. Cell Death & Disease 14, 790.

[24] Usman, M., Hameed, Y., 2023. Gnb1, a novel diagnostic and prognostic potential biomarker of head and neck and liver hepatocellular carcinoma. Journal of Cancer Research and Therapeutics .

[25] Wang, Y., Gao, X., Ru, X., Sun, P., Wang, J., 2022. A hybrid feature selection algorithm and its application in bioinformatics. PeerJ Computer Science 8, e933.

[26] Zhou, L., Tang, H., Wang, F., Chen, L., Ou, S., Wu, T., Xu, J., Guo, K., 2018. Bioinformatics analyses of significant genes, related pathways and candidate prognostic biomarkers in glioblastoma. Molecular medicine reports 18, 4185–4196.

[27] Zou, H., Chen, P., Li, Z., Yan, T., Cui, D., Gong, L., Fang, J., Ren, Y., Chen, M., Yu, J., et al., 2024. Identification of gnb1 as a downstream effector of the circrna-0133711/mir-145-5p axis involved in breast cancer proliferation and metastasis. Oncologie .

[28] Zou, L., Yang, Y., Zhou, B., Li, W., Liu, K., Li, G., Miao, H., Song, X., Yang, J., Geng, Y., et al., 2022. trf-3013b inhibits gallbladder cancer proliferation by targeting tprg1l. Cellular & Molecular Biology Letters 27, 99.