

EMI: A Multi-Objective Evolutionary Massive Integration Method for Multi-Omics Data Analysis

Esmaeil Hasanzadeh¹, Mohammad Saniee Abadeh^{2*}

Faculty of Electrical and Computer Engineering, Tarbiat Modares University^{1,2}

School of Computer Science Institute for Research in Fundamental Sciences (IPM) Tehran, Iran²

Article History:

Received: 26 January 2026

Received in revised form: 15 March 2026

Accepted: 16 May 2026

Available online: 15 August 2026

Abstract

Multi-omics data analysis is about integrating multiple omics data and create strong and complex models to extract deep relations between an organism's genotype and phenotype. The aim of this study is to propose a novel multi-objective evolutionary massive integration method (EMI) for the classification problem of multi-omics data. In this method, first, we have created massive Support Vector Machine (SVM) based models with random features from each omics data type. Then, using an evolutionary algorithm, a combined model of numerous models is created to satisfy two objectives. The first objective of the proposed method is to improve accuracy of the classification problem and the second objective is to find informative biomarkers related to target phenotype. To validate the proposed method, we have gathered breast cancer multi-omics data from The Cancer Genome Atlas (TCGA). We have used estrogen receptor (ER), progesterone receptor (PR) and human epidermal growth factor receptor 2 (HER2) to define breast cancer subtypes. The goal of this study is to predict these subtypes. The proposed method outperforms state of the art articles in this area of research in terms of accuracy and area under curve. Another contribution of this paper is the introduction of some new biomarkers that related to breast cancer subtypes.

Keywords: multi-omics data analysis, breast cancer, cancer classification, evolutionary computation, massive integration

I. INTRODUCTION

Over the past decade, with the advent of High-throughput advanced technologies, a lot of researches have been done on the classification of cancer phenotypes using the omics data [1], [2], [3], [4], [5], [6] Initially, researchers used a single

type of these omics data, such as gene expression or DNA methylation for this goal. But later they found that by using multiple omics data simultaneously and constructing more complex classification models with data integration methods, they can reach a deeper understanding of genotypic relationships in the human body and achieve more accurate and stable results [7]. Data integration methods for multi-omics data analysis are divided into three categories: model-based integration (Late Integration), concatenation-based integration (Early Integration) and transform-based integration (Intermediate Integration) [8]. Model-based integration methods independently build models on each of omics data type and ultimately integrate these models using methods that the simplest of them is voting. Many researches in the cancer classification area used a model-based integration method for multi-omics data analysis [9], [10], [11], [12]. For example, Tao et al. [13] used different kernels of the SVM learning method to diagnosis breast cancer subtypes. They independently constructed models on CNV, Methylation and mRNA omics data types with different SVM kernels and aggregated these models using the MKL method. In a similar work, Yang et al. [14] used CNV, mRNA and miRNA omics data to detect breast cancer subtypes. They used the genetic algorithm to obtain the best coefficients in integrating models in the MKL method. The disadvantage of the model-based integration methods is that they do not consider the correlation between features in different

omics data type during the learning process. Some researchers also used concatenation-based integration to analyze multi-omics datasets [15], [16], [17]. In this approach, the datasets matrixes of the different data types are first combined and construct a larger matrix. Then the learning process is performed using the machine learning algorithms. Rakshit et al. [18] first linked methylation, mRNA and miRNA data and built a larger matrix to prognosis breast cancer subtypes. Then they used autoencoder deep learning dimensionality reduction algorithm and reduced the dimensions of this matrix. Afterwards, they trained learning algorithm on this new matrix. In another research, Song et al. [19] used concatenation-based integration to classify breast cancer grades using gene expression and DNA methylation datasets. Eventually transform-based integration methods, first

convert data into an intermediate form such as a graph or network or a core matrix, and then combine these intermediate forms and train a machine learning model on it. Many papers used this approach for solving data integration problem [20], [21], [22], [23]. For example, Sharifi-Noghabi et al. [24] used copy number aberration, somatic mutation and gene expression. They used a deep learning algorithm for each input omics data to learn features for each data type. The learned attributes integrated into one representation graph by concatenation. The concatenated graph has been served as input for final classification task. Gade et al. [25] used mRNA and miRNA expression changes to improve clinical outcome prediction in prostate cancer. Since miRNAs are known regulators of mRNAs they used the correlations between them as well as the target prediction information to build a bipartite graph representing the relations between miRNAs and mRNAs. This graph was used to guide the feature selection in order to improve the prediction performance. They achieved to improve prediction performance as well as the stability of the feature selection. Each data integration approaches has advantages and disadvantages that can be found in Table 1. We can merge these approaches in a hybrid approach to have their advantages simultaneously. In this paper, we introduced a novel hybrid multi-objective evolutionary massive integration method to integrate the breast cancer multi omics data for classification task. In this method, we have used advantages of all concatenation-based integration, model-based integration and transform-based integration approaches to achieve better results in breast cancer subtypes prognosis.

Table 1 . Advantages and Disadvantages of Each Data Integration Approaches

Integration approach	advantages	Disadvantages
Model-based integration(MBI)	Predictive models can be consolidated from various multi-omics data types, and each data type can be gathered from a various set of patients with same phenotype.	It may be challenging to avoid overfitting.
Concatenation-based integration(CBI)	It is fairly easy to leverage various machine learning methods for analyzing continuous or categorical data once a large input matrix is formed.	It may be challenging to combine a large input matrix.
Transform-based integration(TBI)	Unique variables such as patient identifiers can be used to link multi-omics data types and integrate a variety of continuous or categorical data values.	It may be challenging to transform into intermediate forms.

The rest of this paper is organized as follows. In section 2, the used datasets are presented. Section 3 introduces

the proposed multi-objective evolutionary massive integration method for multi-omics data analysis. In section 4, the experimental results are reported and the effectiveness of our new proposed method is confirmed through experiments and finally section 5 concludes the paper.

II. DATASETS

In this research, breast cancer omics data were gathered from The Cancer Genome Atlas (TCGA), which is a landmark cancer genomics program, molecularly characterized over 20,000 primary cancers and matched normal samples spanning 33 cancer types. The mRNA, DNAm and CNV data were chosen to validate our proposed method, because these three types of omics data have been measured on most patient records simultaneously. In this research, we used the RNA sequencing level 3 data as mRNA data, DNAm 450k level 3 data as DNAm data, and the Affymetrix SNP 6.0 array data with GRCH 38 genome data as CNV data. The details of the data used in this paper are presented in Table 2.

For data preprocessing, first patients who do not have any of the three omics data types are eliminated. Then we have removed the features, which have missing values over 200 patients. For other features missing values estimated by other values of that feature for different patients. In mRNA dataset as for methylation data, we have only considered regulatory relationships between the gene transcription and the hyper-methylation or hypo-methylation of relevant promoter. Moreover, probes in methylation data were matched to genes using these relationships. CNV annotation tools in PennCNV were used to convert the ID of probes to gene symbols [26]. In addition, values of each gene were normalized between [0, 1]. Eventually, each omics data was represented as a two-dimensional matrix with each column representing one gene symbol and each row representing one patient with its own corresponding subtype. Then these three datasets are attached together based on the patient records and a bigger dataset is made. Breast cancer was categorized into five subtypes based on presence or absence of immunohistochemistry (IHC) markers which contain estrogen receptor (ER), progesterone receptor (PR) and human epidermal growth factor receptor 2 (HER2). These molecular subtypes comprise Luminal A, Luminal B, Basal like, HER2+, and Unclear [27]. The detailed definition of each subtype can be found in Table2. The “unclear” subtype denotes the patients that have missing data in ER, PR or HER2, and we have removed patients that have “unclear” breast cancer subtype.

Table 2 . Definition of Breast Cancer Subtypes

Breast cancer subtypes	Definition
Luminal A	ER/PR+; HER2-
Luminal B	ER/PR+; HER2+
Basal like	ER/PR-; HER2-
HER2+	ER/PR-; HER2+
Unclear	Missing ER/PR/HER2

III. PROPOSED METHOD

According to the flowchart that is shown in Figure 1, the proposed method contains three principal stages. The first stage is massive model generation. The second is create an intermediate form of the original dataset and the third stage is a multi-objective evolutionary algorithm which attempts to combine those massive models to a final accurate model and find biomarkers that related to breast cancer subtypes.

A. Massive model generation

In this part of the proposed method, massive models are constructed using the SVM based learning algorithm with the RBF kernel. The main idea behind this work is to build numerous numbers of weak learners and then build a robust and strong model by combining them. For this purpose, 250,000 of SVM models are created. Afterwards, different omics data are combined and made a bigger dataset which we call it the combined dataset. Each model that is produced in this phase has been trained by 15 different features that have been randomly selected from the combined dataset. Among the produced models, those that have accuracy more than 70 percent are stored in a pool and other models are eliminated. Another important point is that during this process, another dataset is generated that contains the feature subset that have been employed to train each model in the pool. At the end of this step, 10676 models were stored in the models pool. By choosing 15 features that are selected from the features of the combined dataset, it is possible for the final algorithm to consider the correlation between the features of different omics in its modeling process. Indeed a type of concatenation based integration performed in this stage of the proposed method.

B. Create intermediate form of dataset

As mentioned, it is clear that each model has stored in the pool trained by what features. At this point, for each model, we filter the training dataset by the features that the model is trained by, and then obtain the prediction of that model for each patient and store it in a new matrix. As a result, a new intermediate form of dataset is created whose rows represent patients and its columns represent the prediction of each model in the model pool for that patient. Finally, last column of the origin matrix is appended to new data as the real class of patients. Then the test dataset, which is separated from the first stage, is given to this process and a new test data is created.

C. Multi-objective evolutionary algorithm

In this section, the proposed multi-objective genetic algorithm has been discussed. This algorithm, constructs an accurate model from the intermediate data made in the previous step to predict breast cancer subtypes. The chromosome encoding of the proposed genetic algorithm is a binary array that each element represents a model that has been saved in the pool. The value of one for each model indicates that this model is involved in constructing the final model. For the initial population, 200 chromosomes, each

containing 10676 elements (the number of models stored in the models pool) are produced. In the initialization step, 10 to 100 elements of each chromosome is equal to one. The fitness function of this algorithm is defined to satisfy two objectives. According to equation (1), the first objective is to improve the accuracy of the SVM based model and the second objective is to build the model using a combination of data that maximizes the overlaps between features (to find main biomarkers related to breast cancer subtypes). As a result, algorithm tries to increase the overlaps of features by increasing the total number of employed models. To mitigate this problem we have divided the overlaps of features in the fitness function to the number of models. The ratio of the weight of the first objective to the second objective is nine to one.

$$f = 0.9 * accuracy + 0.1 * (overlaps / number of models) \quad (1)$$

We have used the Roulette wheel method as the algorithm selection strategy. Elitism has been also used in the proposed genetic algorithm. Thus, 10 chromosomes with the most value of fitness function are directly selected for the next generation. Parent Difference-based crossover which is shown in Figure 2 is used as the crossover operator [28]. The proposed algorithm only applies crossover for those elements of two parents that have different values. The reason for using this crossover algorithm is that only a limited number of genes in each chromosome have value of one and thus most elements of two parent chromosomes are equal. The mutation operator randomly selects five features and for each feature, it changes its value by 50 percent probability. The proposed genetic algorithm with 100 chromosomes as the population size is run for 200 generations and the chromosomes in the last generation were returned as the best answers.

I. EXPERIMENTAL RESULTS

The experimental results related to the execution of the proposed method are presented in two subsections. First, the performance-test will be shown and second, the analysis of selected genes will be presented.

A. Performance test

To evaluate the performance of the proposed method, the best model that has been found in the last generation of the genetic algorithm, has been used to predict the real cancer subtype of the patients in the test dataset. Figure 3 shows the confusion matrix of prediction of the best model on the test dataset. Table 3 shows a comparison between this paper and another research work [13] that has investigated this dataset.

Table 3 . Proposed EMI and Other Studies Comparison

Ref.	Accuracy on train	Accuracy on test
Tao et al. [13]	83.68	79.23
EMI	91.78	88.12

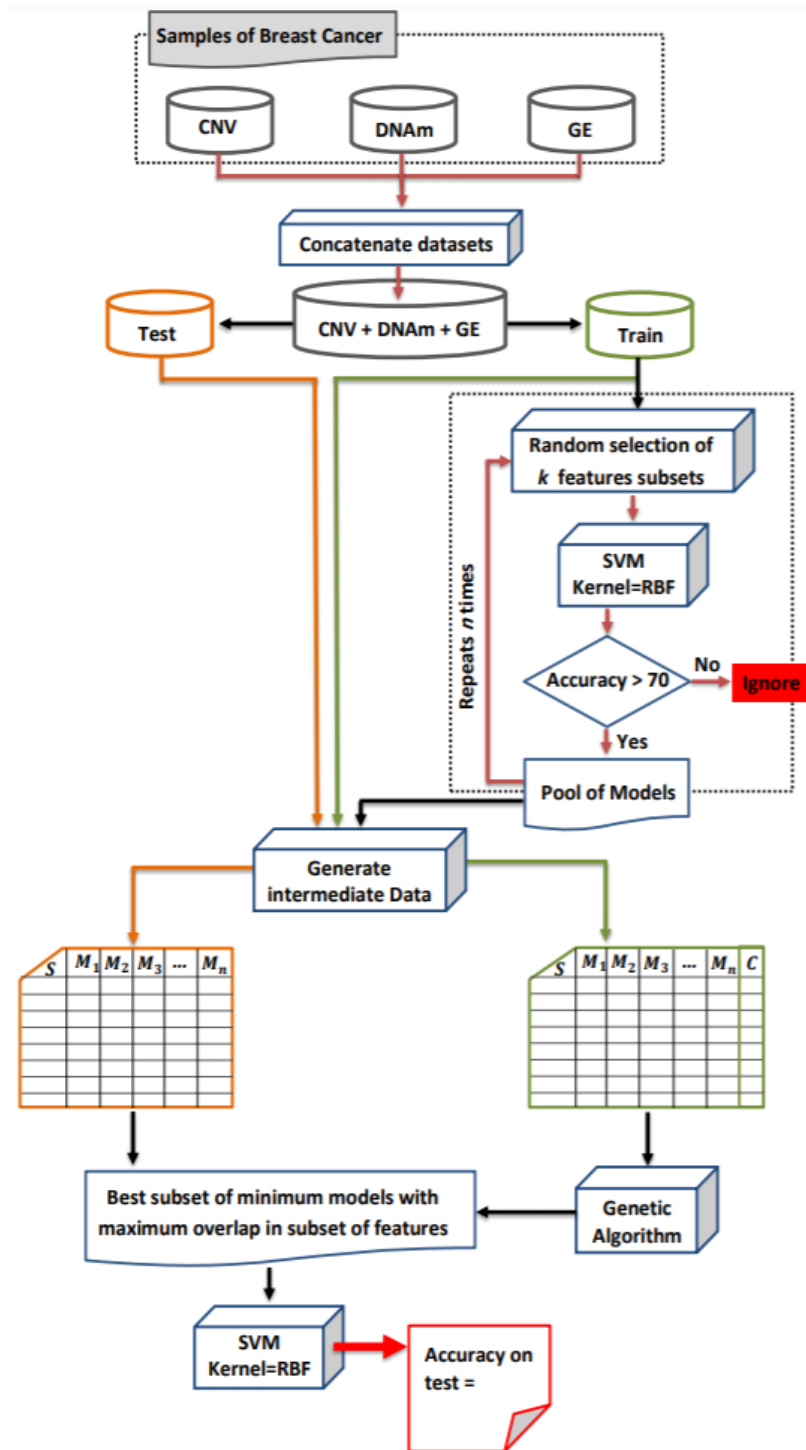


Figure 1. Flowchart of the proposed method

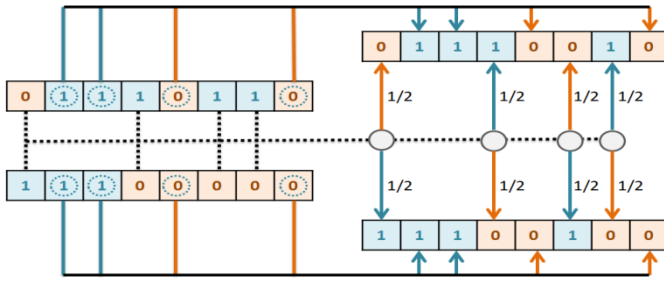


Figure 2. Parent difference-based crossover [28].

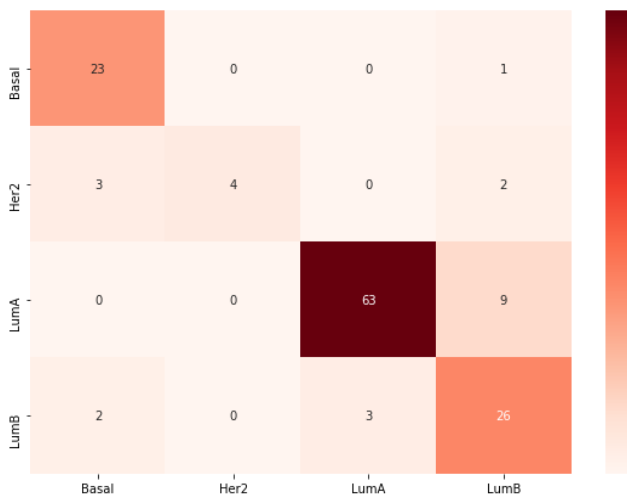


Figure 3. Confusion matrix of prediction on test dataset

B. Analysis of selected biomarkers

Table 4 shows the 20 best biomarkers that have been found by EMI.

Table 4 . Best Biomarkers Related to Breast Cancer Subtypes Found By EMI

Gene expression biomarkers	DNA Methylation biomarkers	CNV biomarkers
ZNF623, PRKAR1B, FAM189A2, RND1, PPP1R14C, TACR1, SGOL1, SPC25, SLC22A24, CHADL, TXNDC15, ASAP1	ATP8B4, FBXL17, C2orf55, TOX, VEZF1	BRMS1, RPL10L, PYROXD1

II. DISCUSSION

Most of the recent research in the field of genomic data analysis has used only one type of genomic data, such as gene expression data. For this reason, they usually ended up with simple models that either suffer from low predictive accuracy or have some kind of instability problem (the final extracted model of the algorithm is different in each run).

Researchers then found that by combining different omics data and using them simultaneously, they could both achieve better results in predicting a particular phenotype and increase the stability of their algorithm. In other words, combining these types of data and use the whole multi-omics data as a single consolidated dataset, the hidden biological ordering that exists in the genotype of the body is revealed and the algorithm can achieve more precise and stable models.

Most recent research in the field of multi-omics data analysis has utilized either a concatenation-based integration approach, or a model-based integration approach or a transform-based integration approach. However, in the proposed method it is tried to use all three integration approaches in different stages of the algorithm to take advantage of each of these three approaches. Nevertheless, we acknowledge that our approach has some limitations. For example, we did not consider the impact of age on breast cancer subtypes. Miller et al. [29] has reported that breast cancer is more aggressive in younger women. Therefore, we could achieve better results by combining patient clinical data with patient genomic data.

III. CONCLUSION

In this paper, we proposed a multi-objective evolutionary massive integration method called EMI for multi-omics data analysis. The proposed method obtains the accuracy of 91.78 % on train and accuracy of 88.12 % on test data in breast cancer subtype prediction problem, which demonstrate the excellent performance of the proposed method. Moreover, EMI was capable of introducing some new biomarkers that related to breast cancer subtypes.

REFERENCES

- [1] R. Nourian, S. A. Motamedi, and M. Pourfard, "BHBA-GRNet: Cancer detection through improved gene expression profiling using Binary Honey Badger Algorithm and Gene Residual-based Network," *Comput. Biol. Med.*, vol. 184, p. 109348, Jan. 2025, doi: 10.1016/J.COMPBIOMED.2024.109348.
- [2] M. Hum and A. S. G. Lee, "DNA methylation in breast cancer: early detection and biomarker discovery through current and emerging approaches," *J. Transl. Med.* 2025 231, vol. 23, no. 1, pp. 1–23, Apr. 2025, doi: 10.1186/S12967-025-06495-2.
- [3] A. Das, M. N. Mohanty, P. K. Mallick, P. Tiwari, K. Muhammad, and H. Zhu, "Breast cancer detection using an ensemble deep learning method," *Biomed. Signal Process. Control*, vol. 70, p. 103009, Sep. 2021, doi: 10.1016/J.BSPC.2021.103009.

- [4] J. X. Ren, L. Chen, W. Guo, K. Y. Feng, Y.-D. Cai, and T. Huang, "Patterns of Gene Expression Profiles Associated with Colorectal Cancer in Colorectal Mucosa by Using Machine Learning Methods," *Comb. Chem. High Throughput Screen.*, vol. 27, no. 19, pp. 2921–2934, Nov. 2023, doi: 10.2174/0113862073266300231026103844/CITE/REFWORKS.
- [5] M. Ghosh, S. Begum, R. Sarkar, D. Chakraborty, and U. Maulik, "Recursive Memetic Algorithm for gene selection in microarray data," *Expert Syst. Appl.*, vol. 116, pp. 172–185, 2019, doi: 10.1016/j.eswa.2018.06.057.
- [6] M. Ghosh, S. Adhikary, K. K. Ghosh, A. Sardar, S. Begum, and R. Sarkar, "Genetic algorithm based cancerous gene identification from microarray data using ensemble of filter methods," *Med. Biol. Eng. Comput.*, vol. 57, no. 1, pp. 159–176, Jan. 2019, doi: 10.1007/S11517-018-1874-4.
- [7] S. Huang, K. Chaudhary, and L. X. Garmire, "More is better: Recent progress in multi-omics data integration methods," *Front. Genet.*, vol. 8, no. JUN, pp. 1–12, 2017, doi: 10.3389/fgene.2017.00084.
- [8] Z. Momeni, E. Hassanzadeh, M. Saniee Abadeh, and R. Bellazzi, "A survey on single and multi omics data mining methods in cancer data classification," *J. Biomed. Inform.*, vol. 107, p. 103466, Jul. 2020, doi: 10.1016/j.jbi.2020.103466.
- [9] F. Carrillo-Perez, J. C. Morales, D. Castillo-Secilla, O. Gevaert, I. Rojas, and L. J. Herrera, "Machine-Learning-Based Late Fusion on Multi-Omics and Multi-Scale Data for Non-Small-Cell Lung Cancer Diagnosis," *J. Pers. Med.*, vol. 12, no. 4, p. 601, Apr. 2022, doi: 10.3390/JPM12040601/S1.
- [10] M. Di Filippo et al., "INTEGRATE: Model-based multi-omics data integration to characterize multi-level metabolic regulation," *PLOS Comput. Biol.*, vol. 18, no. 2, p. e1009337, Feb. 2022, doi: 10.1371/JOURNAL.PCBI.1009337.
- [11] D. Kim et al., "Knowledge boosting: A graph-based integration approach with multi-omics data and genomic knowledge for cancer clinical outcome prediction," *J. Am. Med. Informatics Assoc.*, vol. 22, no. 1, pp. 109–120, 2015, doi: 10.1136/amiajnl-2013-002481.
- [12] S. Ma, J. Ren, and D. Fenyő, "Breast Cancer Prognostics Using Multi-Omics Data," *AMIA Summits Transl. Sci. Proc.*, vol. 2016, p. 52, 2016, Accessed: Sep. 20, 2024. [Online]. Available: /pmc/articles/PMC5001766/
- [13] M. Tao et al., "Classifying Breast Cancer Subtypes Using Multiple Kernel Learning Based on Omics Data," *Genes (Basel)*, vol. 10, no. 3, p. 200, 2019, doi: 10.3390/genes10030200.
- [14] H. Yang, H. Cao, T. He, T. Wang, and Y. Cui, "Multilevel heterogeneous omics data integration with kernel fusion," *Brief. Bioinform.*, vol. 00, no. April, pp. 1–15, 2018, doi: 10.1093/bib/bby115.
- [15] J. A. THOMPSON and C. J. MARSIT, "A METHYLATION-TO-EXPRESSION FEATURE MODEL FOR GENERATING ACCURATE PROGNOSTIC RISK SCORES AND IDENTIFYING DISEASE TARGETS IN CLEAR CELL KIDNEY CANCER," in *Biocomputing 2017, WORLD SCIENTIFIC*, Jan. 2017, pp. 509–520. doi: 10.1142/9789813207813_0047.
- [16] A. Fu, H. R. Chang, and Z. F. Zhang, "Integrated multiomic predictors for ovarian cancer survival," *Carcinogenesis*, vol. 39, no. 7, pp. 860–868, 2018, doi: 10.1093/carcin/bgy055.
- [17] Y. El-Manzalawy, T. Y. Hsieh, M. Shivakumar, D. Kim, and V. Honavar, "Min-redundancy and max-relevance multi-view feature selection for predicting ovarian cancer survival using multi-omics data 06 Biological Sciences 0604 Genetics," *BMC Med. Genomics*, vol. 11, no. Suppl 3, 2018, doi: 10.1186/s12920-018-0388-0.
- [18] S. Rakshit, I. Saha, S. S. Chakraborty, and D. Plewczynski, "Deep Learning for Integrated Analysis of Breast Cancer Subtype Specific Multi-omics Data," *IEEE Reg. 10 Annu. Int. Conf. Proceedings/TENCON*, vol. 2018-October, pp. 1917–1922, Feb. 2019, doi: 10.1109/TENCON.2018.8650144.
- [19] T. Song, Y. Wang, W. Du, S. Cao, Y. Tian, and Y. Liang, "The method for breast cancer grade prediction and pathway analysis based on improved multiple kernel learning," *J. Bioinform. Comput. Biol.*, vol. 15, no. 1, pp. 1–21, 2017, doi: 10.1142/S0219720016500372.
- [20] E. Muller, I. Shiryan, and E. Borenstein, "Multi-omic integration of microbiome data for identifying disease-associated modules," *Nat. Commun.* 2024 151, vol. 15, no. 1, pp. 1–13, Mar. 2024, doi: 10.1038/s41467-024-46888-3.
- [21] W. Lan, H. Liao, Q. Chen, L. Zhu, Y. Pan, and Y. P. P. Chen, "DeepKEGG: a multi-omics data integration framework with biological insights for cancer recurrence prediction and biomarker discovery," *Brief. Bioinform.*, vol. 25, no. 3, Mar. 2024, doi: 10.1093/BIB/BBAE185.
- [22] Y. L. Bernal Rubio et al., "Whole-Genome Multi-omic Study of Survival in Patients with Glioblastoma Multiforme," *G3*, #58; Genes|Genomes|Genetics, vol. 8, no. 11, pp. 3627–3636, 2019, doi: 10.1534/g3.118.200391.

- [23] L. Zhang et al., "Deep learning-based multi-omics data integration reveals two prognostic subtypes in high-risk neuroblastoma," *Front. Genet.*, vol. 9, no. OCT, p. 416381, Oct. 2018, doi: 10.3389/FGENE.2018.00477/BIBTEX.
- [24] H. Sharifi-Noghabi, O. Zolotareva, C. C. Collins, and M. Ester, "MOLI: multi-omics late integration with deep neural networks for drug response prediction," *Bioinformatics*, vol. 35, no. 14, pp. i501–i509, Jul. 2019, doi: 10.1093/BIOINFORMATICS/BTZ318.
- [25] S. Gade et al., "Graph based fusion of miRNA and mRNA expression data improves clinical outcome prediction in prostate cancer," *BMC Bioinformatics*, vol. 12, no. 1, 2011, doi: 10.1186/1471-2105-12-488.
- [26] K. Wang et al., "PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data," *Genome Res.*, vol. 17, no. 11, pp. 1665–1674, Nov. 2007, doi: 10.1101/gr.6861907.
- [27] A. A. Onitilo, J. M. Engel, R. T. Greenlee, and B. N. Mukesh, "Breast cancer subtypes based on ER/PR and Her2 expression: Comparison of clinicopathologic features and survival," *Clin. Med. Res.*, vol. 7, no. 1–2, pp. 4–13, 2009, doi: 10.3121/cmr.2008.825.
- [28] Z. Momeni and M. S. Abadeh, "Mapreduce-based parallel genetic algorithm for CpG-site selection in age prediction," *Genes (Basel)*, vol. 10, no. 12, p. 969, Dec. 2019, doi: 10.3390/genes10120969.
- [29] K. D. Miller et al., "Cancer treatment and survivorship statistics, 2016.," *CA. Cancer J. Clin.*, vol. 66, no. 4, pp. 271–89, 2016, doi: 10.3322/caac.21349.

Email: e_hasanzadeh@modares.ac.ir



Mohammad Saniee Abadeh received the B.S. degree in Computer Engineering from Isfahan University of Technology, Isfahan, Iran, in 2001, the M.S. degree in Artificial Intelligence from Iran University of Science and Technology in 2003, and the Ph.D. degree in Artificial Intelligence from the Department of Computer

Engineering at Sharif University of Technology in 2008. He is currently a Faculty Member of the Faculty of Electrical and Computer Engineering at Tarbiat Modares University. His research interests include data mining, knowledge discovery, computational intelligence, bio-inspired computing, evolutionary algorithms, fuzzy genetic systems, and memetic algorithms.

Email: saniee@modares.ac.ir

How to cite: E. Hasanzadeh, M. Saniee Abadeh, **EMI: A Multi-Objective Evolutionary Massive Integration Method for Multi-Omics Data Analysis**, *Journal of Distributed Computing and Systems (JDACS)*, Vol 9, No 1, Pages 89-95, 2026.



Esmail Hasanzadeh received the B.Sc. degree from Iran University of Science and Technology, Tehran, Iran, in 2017, and his M.Sc. degree from Tarbiat Modares University, Tehran, Iran, in 2020. He is currently pursuing the Ph.D. degree in computer engineering at Tarbiat Modares University, Tehran, Iran.