scientific reports

Check for updates

OPEN

Adaptive multi-omics integration framework for breast cancer survival analysis

Esmaeil Hasanzadeh & Nasrollah Moghadam Charkari✉

Breast cancer remains a major global health issue, requiring novel strategies for prognostic evaluation and therapeutic decision-making. In this study, we leverage multi-omics data from The Cancer Genome Atlas to obtain deeper insights into breast cancer biology. By integrating genomics, transcriptomics, and epigenomics, we aim to identify complex molecular signatures that drive breast cancer progression and impact patient survival. To optimize the integration and feature selection process within the multi-omics dataset, we have employed genetic programming. Genetic programming helps us to optimize multi-omics integration, enabling the identification of robust biomarkers and more accurate survival analysis. The proposed framework consists of three key components: data preprocessing, adaptive integration and feature selection via genetic programming, and model development. The experimental results indicate that the integrated multi-omics approach yields a concordance index (C-index) of 78.31 during 5 fold cross-validation on the training set and 67.94 on the test set. In conclusion, our study demonstrates the potential of adaptive multi-omics integration in improving breast cancer survival analysis. It also highlights the importance of considering the complex interplay between different molecular layers. Furthermore, this framework provides a flexible and scalable approach that can be extended to other cancer types, offering valuable insights into oncological processes.

Keywords Multi-omics, Breast cancer, Survival analysis, Data integration, Genetic programming

Breast cancer is a multifaceted and heterogeneous condition that affects millions of women globally, profoundly influencing their quality of life, as well as that of their families and communities. Despite considerable advancements in molecular biology and genomics, breast cancer persists as a leading cause of cancer-related mortality among women, accounting for approximately one in six cancer deaths globally[1]. The development of effective treatment strategies for breast cancer is hindered by a complex interaction between genetic, epigenetic, and environmental factors that contribute to the disease[2]. This complexity is further compounded by the inherent heterogeneity of breast cancer, which encompasses a diverse array of molecular subtypes, each with distinct clinical and pathological characteristics[3]. To overcome these challenges, there is a growing need for innovative approaches that can integrate multiple types of molecular data to identify robust biomarkers and predict patient outcomes. The advent of high-throughput sequencing technologies has enabled the generation of large-scale molecular data, including genomics, transcriptomics, proteomics, and epigenomics[4]. These types of data provide a comprehensive view of the molecular landscape of breast cancer, while their integration and analysis remain a significant challenge[5]. Traditional single-omics approaches have been widely used to identify biomarkers and predict patient outcomes. However, they often fail to capture the complex interactions between different molecular layers[6]. The limitations of single-omics approaches are well-documented in various studies. For instance, genomic studies have identified numerous genetic mutations associated with breast cancer, but these mutations often do not give a complete picture of the disease[7]. Similarly, transcriptomic studies have identified gene expression signatures associated with breast cancer subtypes, but they fail to capture the full range of molecular heterogeneity within each subtype[8]. To address these limitations, multi-omics integration has emerged as a promising approach for breast cancer research. Multi-omics integration involves the combination of multiple types of molecular data to identify patterns and relationships that are not apparent from single-omics analyses[9].

Integrating multi-omics data, including genomics, transcriptomics, proteomics, and epigenomics, has become a crucial step in understanding the complex biology of diseases. There are three primary strategies for integrating multi-omics data: early integration, intermediate integration, and late integration. Early integration

Faculty of Electrical and Computer Engineering, Tarbiat Modares University, Tehran, Iran. ✉email: moghadam@modares.ac.ir

involves combining data from different omics levels at the beginning of the analysis pipeline. This approach can help identify correlations and relationships between different omics layers, but may also lead to information loss and biases. Intermediate integration, on the other hand, involves integrating data from different omics levels at the feature selection, feature extraction, or model development stages, allowing for more flexibility and control over the integration process. The third strategy, late integration, also known as "vertical integration," involves the analysis of each omics dataset separately, and further combining the results at the final stage of the analysis pipeline. This approach helps preserve the unique characteristics of each omics dataset, but may also lead to difficulties in identifying the relationships between different omics layers. The selection of an integration strategy depends on the research question, data characteristics, and analytical objectives. Accordingly, a comprehensive understanding of the strengths and weaknesses of each approach is essential for effective multi-omics data analysis[6,10,11].

Survival analysis is a crucial aspect of breast cancer research, as it enables the evaluation of the effectiveness of different treatment strategies and the identification of prognostic biomarkers. The C-index, also known as the concordance index, is a widely used metric for evaluating the performance of survival models. It measures the proportion of pairs of patients who are correctly ordered by the model, with higher values indicating better performance[12]. In the context of breast cancer, the C-index can be used to evaluate the ability of a model to predict patient outcomes, such as overall survival or disease-free survival[13].

The integration of multiple types of molecular data is a complex and challenging task that requires the development of sophisticated computational methods. As the volume and diversity of molecular data are growing continuously, there is an increasing need for innovative computational approaches that can effectively integrate and analyze these data, identify robust biomarkers, and provide new insights into the underlying biological mechanisms. However, the development of accurate survival models for breast cancer is challenging due to the complex interplay between genetic, epigenetic, and environmental factors that contribute to the disease. By integrating multiple types of molecular data and using advanced computational approaches, we aim to develop a novel survival model that can accurately predict patient outcomes and provide new insights into the molecular mechanisms driving breast cancer.

Genetic programming is a powerful algorithm that has been used for solving complex problems in various fields, such as bioinformatics and computational biology[14,15]. This approach leverages the use of evolutionary principles to search for the optimal solutions for a problem. It has been shown to be effective in identifying complex patterns and their relationships in large datasets[16]. In the context of multi-omics integration, genetic programming can be used to evolve optimal combinations of molecular features that are associated with breast cancer outcomes. It also helps identify robust biomarkers that can be used for patient stratification and treatment planning. By using genetic programming, we aim to develop a novel multi-omics integration framework that can effectively identify complex patterns and relationships in breast cancer data, and provide new insights into the molecular mechanisms underlying this disease.

Our main contribution to this study is the development of an adaptive multi-omics integration framework that utilizes genetic programming to evolve optimal integration of omics data. Unlike traditional multi-omics integration approaches that rely on fixed integration methods, the proposed method employs genetic programming to adaptively select the most informative features from each omics dataset at each integration level. This approach optimizes feature selection and integration, leading to more accurate and robust biomarker discovery. By using genetic programming to evolve the integration process, it would be possible to identify the most relevant features and relationships between different omics datasets and gain a deeper insight into the complex molecular mechanisms driving breast cancer. This adaptive integration approach has the potential to revolutionize the field of multi-omics integration, and provides new insights into the diagnosis, prognosis, and treatment of complex diseases like breast cancer.

Related works

The integration of multi-omics data represents a pivotal advancement in cancer research, offering an intricate understanding of the complex biological processes underlying various cancer types, including breast cancer. This burgeoning field has seen a variety of innovative methodologies that facilitate the synthesis of diverse omic datasets, thereby enhancing our understanding of cancer mechanisms and paving the way for tailored therapeutic approaches. In the realm of breast cancer, the classification of subtypes has greatly benefited from the application of advanced computational techniques, particularly deep learning[17,18]. For example, Lin et al. developed DeepMO, a sophisticated deep neural network that amalgamates mRNA expression, DNA methylation, and copy number variation (CNV) data to classify breast cancer subtypes, achieving a notable binary classification accuracy of 78.2%[19]. In a similar vein, Choi et al. introduced moBRCA-net, which employs a self-attention mechanism to integrate gene expression, DNA methylation, and microRNA expression[20]. Islam et al. proposed a deep neural network framework that autonomously extracts features from raw copy number alteration and gene expression data, further contributing to this problem[21]. Beyond subtype classification, the prognostic applications of multi-omics integration have shown significant promise in enhancing predictions of patient survival. Fan et al. applied machine learning techniques to integrate multi-omics data, utilizing differential network analysis methods to identify genes associated with breast cancer prognosis and providing insights into the molecular mechanisms underlying the disease[22]. Poirion et al. presented DeepProg, which combines deep-learning and machine-learning techniques that robustly predicts survival subtypes across liver and breast cancer datasets, with a concordance index (C-index) ranging from 0.68 to 0.80[23]. Moreover, Bichindaritz et al. demonstrated the utility of combining genomic mRNA and DNA methylation data for improved survival prediction in breast cancer, employing an adaptive multi-task learning approach that integrates Cox and ordinal loss for survival analysis[24]. Recent advancements in adaptive multi-omics integration methods highlight their transformative potential. Dong et al. introduced MOGLAM, an end-to-end interpretable method that uses a

dynamic graph convolutional network with feature selection to generate high-quality omic-specific embeddings and identify important biomarkers. A multi-omics attention mechanism further enhances the integration of different omics data, demonstrating enhanced performance relative to other existing methods[25]. Similarly, Cheng et al. developed MoAGL-SA, which employs graph learning and self-attention to improve cancer subtype classification. This method creates patient relationship graphs from omics datasets and utilizes three-layer graph convolutional networks to extract specific embeddings, adaptively weighting them for integration. MoAGL-SA demonstrates superior performance compared to various widely-used algorithms across multiple datasets[26]. The scope of multi-omics integration extends beyond breast cancer to other malignancies such as liver cancer, colon adenocarcinoma, esophageal squamous cell carcinoma, and muscle-invasive bladder cancer, where it continues to yield promising outcomes. For instance, Chaudhary et al. developed a deep learning model that integrates multi-omics data for liver cancer survival prediction[27]. Lv et al. and Yu et al. respectively proposed deep learning algorithms for survival prediction in colon adenocarcinoma and esophageal squamous cell carcinoma[28,29]. Zhang et al. introduced a robust prognostic subtyping framework for muscle-invasive bladder cancer based on multi-omics data integration[30]. In contrast to deep learning approaches, some studies have adopted classical statistical models. For example, Liu et al. introduced SKI-Cox and LASSO-Cox, which enhance prognosis prediction in glioblastoma and lung adenocarcinoma by incorporating inter-omics relationships into the Cox regression model[31]. Another alternative to neural integration is latent factor modeling. MOFA and its scalable variant MOFA + perform Bayesian group factor analysis to learn a shared low-dimensional representation across omics datasets. These models infer latent factors that capture key sources of variability, using sparsity-promoting priors to distinguish shared from modality-specific signals. This enables interpretability by linking latent factors to specific molecular features. Compared to neural networks that learn fused representations, MOFA explicitly models factor structure and often requires less data to train[32]. Collectively, these studies underscore the transformative potential of multi-omics integration in oncology, significantly advancing our understanding of cancer biology and facilitating the development of effective prognostic models across various cancer types. This not only improves patient outcomes but also refines existing treatment strategies, marking a significant stride toward personalized medicine. As observed, most research in this domain has been grounded in deep-learning methodologies. While these techniques demonstrate high accuracy, they face a significant challenge regarding the interpretability of their results. In the medical field, particularly in oncology, deep learning models often cannot independently identify the specific features or genes that contribute to their predictive accuracy. This limitation highlights the need for approaches that enhance interpretability, ensuring that the insights gained from these models can be effectively translated into clinical practice.

Material and methods

The framework suggested in this study, known as the Adaptive Multi-Omics Integration Framework, utilizes evolutionary genetic programming technique to identify the most effective strategies for integrating multi-omics data. Additionally, it seeks to determine the optimal features for integration at each level, enhancing the overall analysis and interpretation of complex biological information. This innovative approach represents a novel design of the genetic programming algorithm, allowing it to discover the most effective sequence for integrating omics data, the most suitable feature selection methods at each integration level, and the optimal count of features to be selected at each stage. The suggested framework comprises two primary components. In this section, we will discuss the details of both stages: the pre-processing stage and the suggested genetic programming approach.

Dataset

In this study, we utilized the breast cancer dataset from The Cancer Genome Atlas (TCGA), a pioneering cancer genomics project that has analyzed over 20,000 primary cancer samples along with corresponding normal samples across 33 distinct cancer categories.[33]. The TCGA breast cancer cohort provides high-quality, standardized multi-omics data with comprehensive clinical annotations, making it ideal for developing and validating integration methodologies.

We specifically selected three complementary omics modalities based on their biological interconnectedness and relevance to cancer progression. MicroRNA expression data was chosen because miRNAs are critical post-transcriptional regulators that control gene expression and play essential roles in cancer progression, metastasis, and therapeutic response[34]. miRNA dysregulation is particularly relevant in breast cancer, where specific miRNA signatures have been associated with prognosis and treatment outcomes[35]. Gene expression data is included as transcriptomic data captures the functional output of the genome and reflects cellular responses to environmental conditions and disease states. Gene expression profiles are fundamental for understanding cancer biology and identifying therapeutic targets[36]. DNA methylation data is selected because epigenetic modifications, particularly CpG methylation, represent Inherited modifications that occur without changing the underlying DNA sequence but significantly impact gene expression. DNA methylation patterns are extensively altered in cancer and provide complementary information to genetic and transcriptomic data[37].

The rationale for this specific combination is based on the regulatory hierarchy where DNA methylation influences gene expression, gene expression affects protein production, and miRNAs provide post-transcriptional control. While genetic variant data (SNPs/CNVs) are more widely available and could be used, we focused on functional molecular layers that directly reflect cellular state and regulatory interactions. This approach captures the flow of biological information from epigenetic regulation (methylation) through transcriptional control (gene expression) to post-transcriptional regulation (miRNA), providing a comprehensive view of molecular dysregulation in breast cancer.

Preprocessing pipeline

To prepare the dataset for modeling, we performed a set of pre-processing steps designed to ensure data quality, reduce dimensionality, and improve model performance. As illustrated in Fig. 1, the pre-processing component encompasses several steps. First, the dataset is randomly split into training and testing subsets with a ratio of 70:30 to ensure proper evaluation of the model's performance. Next, the samples are filtered to only include those with complete data across all modalities, ensuring comprehensive datasets. All features were normalized using min–max scaling implemented through scikit-learn's MinMaxScaler, which transforms each feature to the [0,1] range using the formula: (x—min(x)) / (max(x)—min(x)). This normalization approach was applied uniformly across all omics modalities (gene expression, miRNA expression, and DNA methylation) to ensure equal contribution potential during the integration process and prevent any single modality from dominating due to scale differences. To maintain the integrity of the dataset, the missing values are estimated using averaging methods that consider other feature values and samples[38]. This method was chosen for its computational efficiency and simplicity. In this approach, each missing value in a feature was replaced with the arithmetic mean of the observed values for that same feature. While we acknowledge that this method can reduce the natural variance of the feature, it was considered a suitable approach given the low percentage of missing data, as it preserves the feature's central tendency without the need for more complex modeling assumptions. To manage the high dimensionality of the omics data, we employed a feature selection strategy. This method is a univariate filter based on feature variance, designed to remove uninformative features and reduce the computational burden for the subsequent, more complex selection process. We selected the top 2000 features with the highest variance from each omics modality. The rationale for this 2000-feature cutoff was determined empirically. We tested several cutoffs (e.g., 1000, 2000, 5000, and all features) and found that using the top 2000 features provided the optimal trade-off between the final model's predictive performance and the computational time required for the genetic programming (GP) to converge. This pre-selection directly influences the genetic programming by constraining its search space. By focusing the GP on a pre-filtered subset of high-variance features, we significantly enhance its efficiency and reduce the risk of overfitting to noise. This hierarchical approach assumes that features with higher variance are more likely to be biologically relevant. While there is a risk of excluding low-variance features that could be informative in a multivariate context, this strategy allows the GP to more effectively explore the combinatorial interactions within a high-priority feature set.

Adaptive integration using genetic programming

Genetic Programming (GP) is a type of evolutionary algorithm that is particularly well-suited for solving complex optimization problems. GP is a population-based approach that uses the concepts of natural selection and genetics to identify optimal solutions. The GP involves a series of iterations, called generations, where a population of candidate solutions is evolved using a fitness function that evaluates the quality of each individual. The fitness function determines the likelihood of each chromosome being selected to reproduce and create offspring for the next generation. The GP operators, including mutation, crossover, and selection, are used to create new individuals and introduce genetic variation into the population. In GP, the chromosomes are represented as trees. A key feature of GP is its use of a tree-based structure of chromosomes to represent solutions, which is ideal for modeling the hierarchical nature of multi-omics integration. GP also uses a

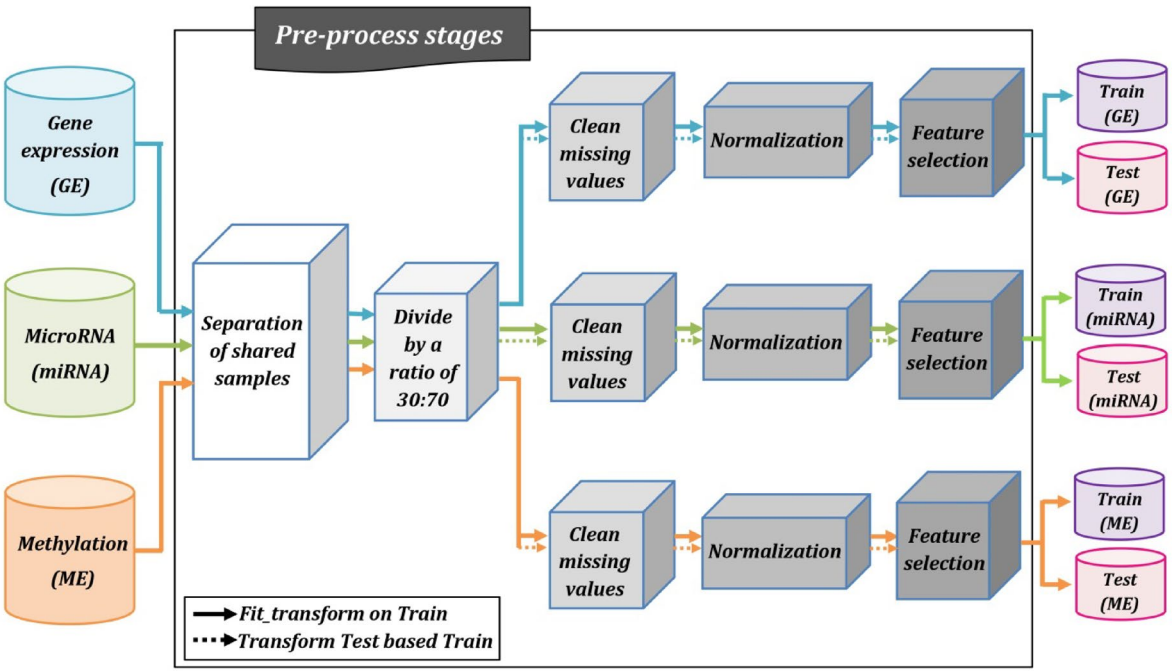

Fig. 1. Flowchart of the pre-processing stage in the proposed method.

termination criterion to determine when to stop the evolution process. One of the primary strengths of GP is its capability to automatically discover complex relationships and patterns in data, but it can also be computationally expensive and require significant tuning of its parameters[39]. Our GP framework is designed to explore a flexible spectrum of integration strategies, ranging from early to intermediate integration. In multi-omics analysis, early integration typically involves concatenating all features from all omics types into a single matrix before performing feature selection and model training. Late integration, in contrast, involves building separate models for each omic and combining their predictions. Our approach focuses on discovering optimal intermediate integration strategies, where subsets of omics data are integrated in a hierarchical, tree-like fashion. The GP autonomously designs this integration tree, determining which omics to combine first and how to progressively reduce feature dimensionality at each step. The flowchart illustrating the suggested genetic programming in this study is presented in Fig. 2.

To evolve these integration strategies, chromosome encoding is a critical issue. In our framework, each chromosome represents a complete integration pipeline. This is encoded by defining the integration tree's structure and the operations at each node. Specifically, four genes define each non-leaf node in the tree: integration tree depth, number of children per node, the feature selection method, and the count of selected features, as illustrated in Fig. 3.

- *Integration Tree Depth* This gene dictates the maximum allowable depth of the subtree rooted at the current node, with alleles from the set {1, 2, 3, 4}. This directly controls the complexity and hierarchical nature of the integration. For example, a tree with a maximum depth of 1 forces a flat structure akin to early integration, where all selected child nodes are combined in a single step. Conversely, a depth of 4 allows for a deep, multi-level hierarchy, representing a complex intermediate integration strategy. The GP therefore learns not just which omics to combine, but the very modality (early vs. intermediate) of integration.
- *Number of Children Per Node* This gene controls the branching factor (alleles from {1, 2, 3, 4}), influencing the intricacy of the tree.
- *Feature Selection Method* This gene determines which of eight distinct algorithms is used to select features at a node: variance, Pearson correlation, random forest, gradient boosting, AdaBoost, ExtraTrees, mutual information, or f-regression[40,41].
- *Count of Selected Features* This specifies the number of features to be retained by the feature selection method at that node.

Several key aspects of the chromosome's construction and representation are noteworthy as follows;

- Leaf nodes have a different structure, characterized by a single value that specifies the type of omics data.

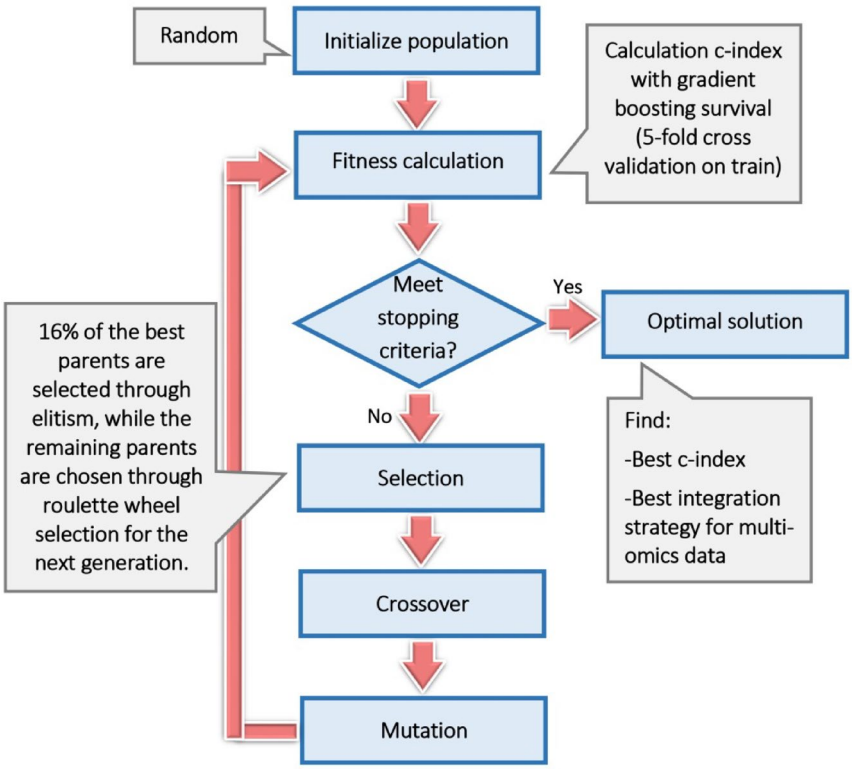

Fig. 2. Flowchart of the suggested genetic programming.

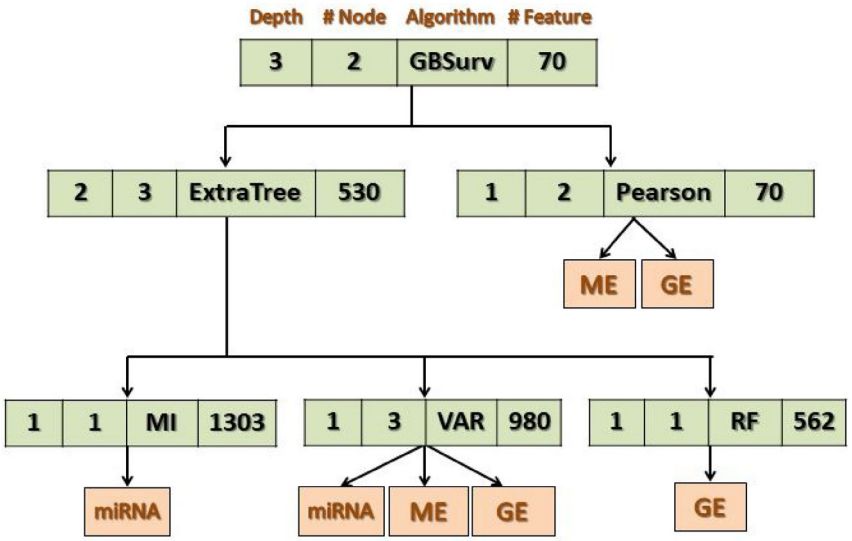

Fig. 3. Chromosome encoding of the suggested genetic programming: the tree structure is encoded in a chromosome, where each node is represented by four cells. The first cell indicates the depth of the sub-tree, the second cell specifies the number of children for each node, the third cell denotes the feature selection method employed, and the fourth cell represents the count of features to be selected using the designated feature selection method.

- A fundamental property of the tree is that each intermediate node and root at depth n must have at least one child node at depth n-1, ensuring a hierarchical structure.
- Additionally, no intermediate node can exceed the depth of its parent node, maintaining a consistent depth hierarchy.
- Furthermore, to prevent redundancy, no two leaves under the same parent node can share identical omics values, ensuring diversity in the represented data.
- The number of selected features for each node must be less than the total number of features entering that node from its branches.

To calculate the fitness of each chromosome, we start from the leaf nodes and move toward the root node, following a bottom-up approach to integrate features from child nodes. At each node, we concatenate the features selected by its child nodes and apply the designated feature selection algorithm to select the specified count of features. Ultimately, at the root node, the final integrated dataset is obtained, comprising the selected features. We then train a gradient-boosting survival model using the scikit-survival library on this dataset[42]. The fivefold cross-validation c-index metric is used as the fitness score, providing a robust evaluation of the model's performance.

To evolve the population, we use a combination of elitism and roulette wheel selection. Elitism ensures that the best-performing individuals from one generation are carried over to the next, preserving high-quality solutions. Crossover and mutation operators introduce genetic diversity. Given the variability in the depth of chromosome trees, it is essential to establish two distinct types of crossover operators, as illustrated in Fig. 4. The selected parent trees may either possess the same depth or are different in depth. When both parents have equal depths, the offspring produced will inherit this depth. In this crossover process, the largest subtree from either tree is exchanged, leading to the formation of two new chromosome trees derived from the original parents. Conversely, when the depths of the selected parents are not aligned, a subtree from the larger tree is replaced with the smaller tree, provided that the subtree has a matching depth. In this scenario, one offspring will mirror the depth of the smaller parent tree, while the other will reflect the depth of the larger parent. This approach effectively generates new offspring from the initial parent trees, accommodating the differences in their structural depths.

To execute the mutation operator, a mutation rate must first be established, which is a small value. Then, by traversing the tree from the root, a random number is generated for each node. If this random number is less than the mutation rate, that node undergoes mutation; if it is greater than the mutation rate, the node remains unchanged. For intermediate nodes and the root node, the mutation involves the complete regeneration of all subtrees of that node, and no further mutation occurs for the regenerated subtrees. However, for leaf nodes, the mutation process involves randomly selecting different omics in place of the previous omics. If there are no alternative options for the omics, no mutation occurs.

Statistical measurements

The concordance index (C-index) is widely recognized as a statistical metric used to evaluate the predictive accuracy of risk models, particularly in survival analysis and time-to-event data. The C-index quantifies the

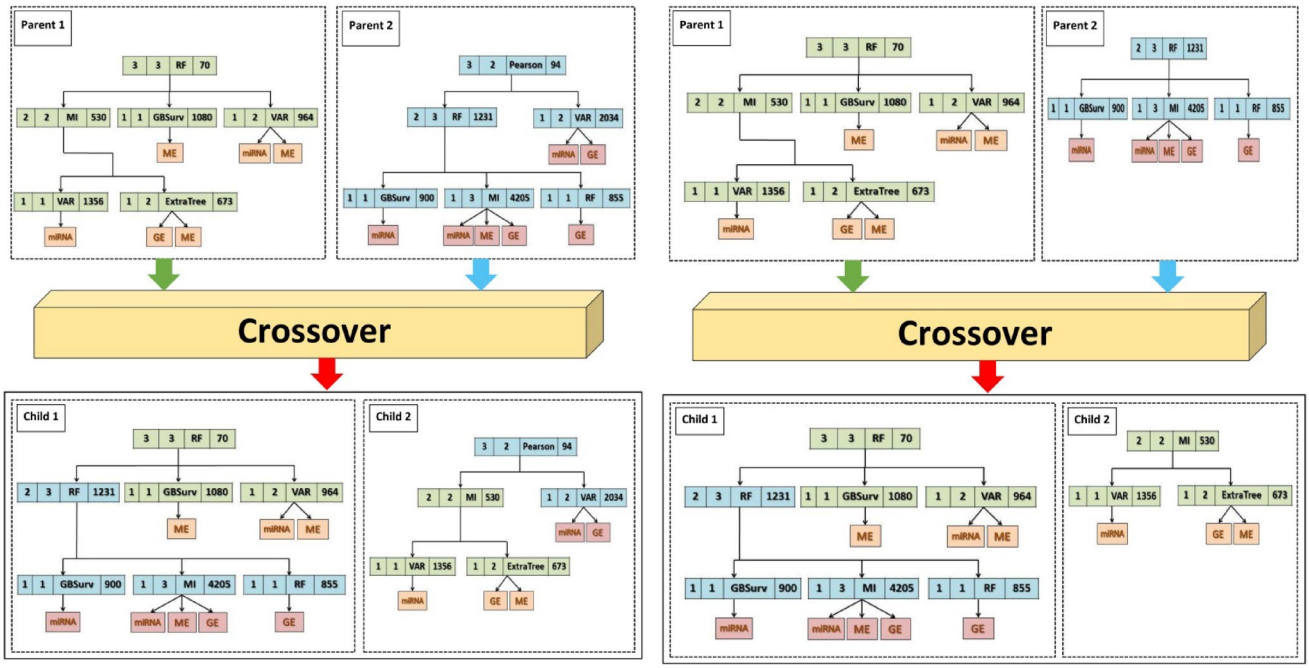

Fig. 4. Crossover of two chromosome with (**a**) identical and (**b**) different depths.

degree of concordance between predicted and observed outcomes, providing a robust measure of model performance in distinguishing between different risk levels. Mathematically, the C-index is characterized as the ratio of all viable patient pairs in which the predictions and outcomes are concordant. Specifically, for a pair of patients, if the patient with the higher predicted risk experiences the event of interest (e.g., death, disease progression) before the patient with the lower predicted risk, this pair is considered concordant. Conversely, if the patient with the lower predicted risk experiences the event first, the pair is discordant. Pairs in which either patient does not experience the event within the study period are considered censored; these pairs are excluded from the calculation of the C-index because they do not provide sufficient information to assess the relative risk between the patients. The formula for the C-index is expressed as follows:

$$C - \text{index} = \frac{1}{|k|} \sum_{i=1}^{m} \sum_{j:t_i<t_j} I\left(F\left(x_i\right) < F\left(x_j\right)\right) \tag{1}$$

Here, |k| represents the total number of comparable pairs, t_i and t_j are the times of events for patients i and j, respectively, F(x) denotes the risk function evaluated at patient covariates x, and I is the indicator function that equals 1 when the condition is true and 0 otherwise. This condition checks whether the risk score $F(x_i)$ for patient i is less than the risk score $F(x_j)$ for patient j, under the circumstance where the event time t_i is less than t_j. The C-index ranges from 0.5 to 1.0, where a C-index of 0.5 shows no discriminative capability (equivalent to random chance), while a C-index of 1.0 signifies perfect discrimination. A higher C-index value indicates better model performance in predicting the order of events, making it a valuable tool for assessing the effectiveness of the model developed in this research. The C-index values gained from our experiments are reported in the Results section, providing a quantitative measure of the predictive accuracy of the proposed framework[43].

Experimental results

Genetic programming model configuration and tuning

The performance of the GP is highly dependent on its parameter configuration. To ensure the reliability of our results and to find an optimal balance between the final model's predictive accuracy and computational cost, we conducted a series of preliminary experiments to tune key hyperparameters. The final parameters used for the experiment, along with the rationale and tested ranges, are detailed in Table 1.

This systematic tuning process allowed us to establish a robust and justified configuration for the GP. This provides confidence that the discovered integration strategies are the output of a methodologically sound optimization process rather than being artifacts of an arbitrary parameter choice.

Parameter	Tested range/options	Selected value	Rationale
Encoding	N/A (core design)	Custom tree structure (Fig. 3)	Designed to represent complex, hierarchical multi-omics integration strategies
#Generation	{50, 100, 150}	100	Empirically determined to be sufficient for the fitness score to converge to a stable solution
Population size	{50, 100}	50	Provided the best trade-off between solution diversity and manageable computational load per generation
Selection method	{Roulette wheel, tournament}	Roulette wheel	A standard probabilistic method that gives fitter individuals a higher chance of being selected
Cross-over operator	N/A (core design)	Custom depth-aware operator (Fig. 4)	Specifically designed to handle the variable-depth tree structures in the population
Mutation rate	{0.1, 0.2, 0.3, 0.4}	0/3	A relatively high rate was chosen to encourage thorough exploration of the vast and complex solution space defined by the integration strategies
Elitism	{4, 8, 12} individuals	8 chromosomes in each generation (16%)	Preserves the best-performing solutions in each generation, ensuring that high-quality strategies are not lost. The rate of 16% was found to be optimal
Random Injection	{4, 8, 12} individuals	8 chromosomes in each generation	In each generation, a set of new random chromosomes was injected to maintain population diversity and help prevent premature convergence to a local optimum
Fitness function	N/A (core design)	Gradient boosting survival with fivefold cross-validation	Provides a robust measure of predictive performance for survival data

Table 1. Parameter configurations of the suggested genetic programming.

Root node of solution (chromosome)	5-Fold C-index (training set)	C-index (testing set)	Omics in final features
[4,3,"GBSurv",18]	**78.31**	66.42	miRNA, methylation, gene expression
[2,1,"GBSurv",36]	77.34	65.27	Methylation, miRNA
[2,1,"GBSurv",14]	76.92	**67.94**	miRNA
[1,2,"GBSurv",71]	75.90	66.14	Gene expression, methylation
[2,3,"FIAdaBoost",41]	74.92	66.42	miRNA, Methylation, gene expression

Table 2. The performance comparison of the solutions generated by the suggested GP. The bold values represent the best c-index achieved for the training and testing datasets.

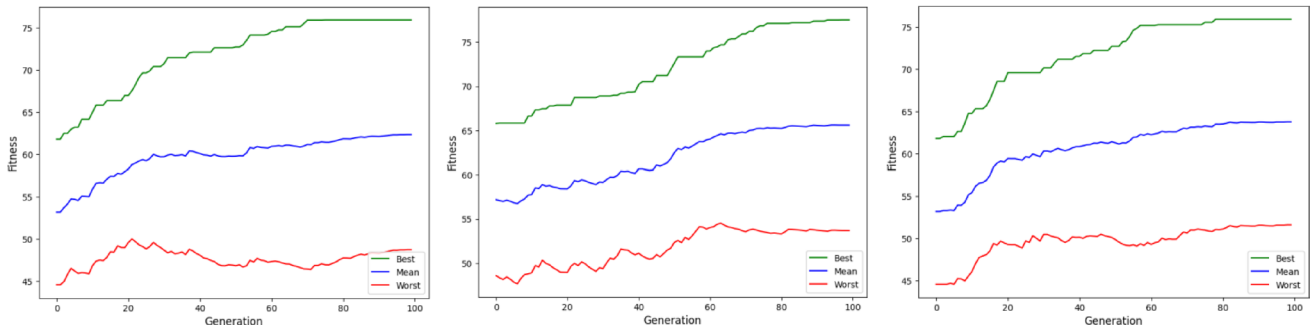

Fig. 5. Progress of the three best runs of the proposed genetic programming algorithm.

Performance evaluation

In this section, we evaluate the performance of the adaptive framework developed through our genetic programming approach. This framework has demonstrated high predictive accuracy, as measured by the concordance index (C-index). The best solution identified by the proposed model achieved a C-index of 78.31 on the training set, indicating proper predictive accuracy and a strong fit to the training data. The accuracy rate reflects the effectiveness of the model in ranking patients according to their risk of experiencing an event, where higher C-index values indicate better discrimination between patients who experienced events and healthy ones. In the testing phase, the model recorded a C-index of 66.42, validating its robustness and its ability to generalize well to unseen data. This performance is significant, demonstrating the model's utility in real-world settings where it can effectively predict outcomes on new patient data, outside of the controlled conditions of the training dataset.

To provide a comprehensive overview of the performance of the various solutions generated by our genetic programming method, we present a comparison of the top five solutions in Table 2. This comparison includes their C-index results, the genes identified in each solution, and the root node of the chromosome.

Figure 5 illustrates the convergence of the three best runs of the proposed genetic programming algorithm toward the optimal solution. The green line represents the C-index value of the best-performing chromosome in

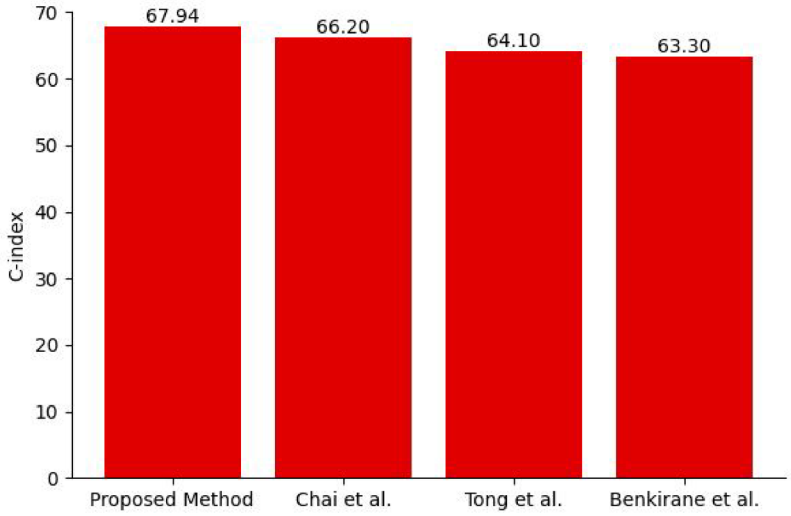

Fig. 6. Comparison of C-index performance among the proposed framework, Chai et al., Tong et al., and Benkirane et al.

Dataset	Data source	Sample size (n)	Available omics for validation	C-index (%)
PCAWG	UCSC Xena browser	89	miRNA, gene expression	60.65
GSE19783	GEO	78	miRNA only	62.09

Table 3. Summary of external validation datasets and performance.

each generation, showcasing the progressive improvement toward optimality. The red line denotes the C-index value of the worst-performing chromosome in each generation, highlighting the diversity within the population. The blue line depicts the average C-index of all chromosomes in each generation, providing an overall measure of population performance.

Currently, deep learning methods are extensively employed in this field due to their remarkable capacity to capture intricate patterns within data. To further demonstrate the superior performance of our framework, we benchmarked it against several state-of-the-art studies, including those conducted by Chai et al.[44], Tong et al.[45], and Benkirane et al.[46], which are recognized as prominent works in the literature. These studies utilized various deep learning-based approaches. Figure 6 presents a comprehensive performance comparison between our proposed method and these approaches, showcasing the advantages of our framework.

External validation of the prognostic signature

To evaluate the generalizability of the proposed model and validate the robustness of the identified 18-feature biomarker signature, we performed external validation using two independent breast cancer cohorts. This step addresses the limitation of single-cohort analyses and tests model performance on unseen data from distinct patient populations and technical platforms. The first cohort is a subset of the Pan-Cancer Analysis of Whole Genomes (PCAWG) project, obtained from the UCSC Xena Browser[47]. From this pan-cancer dataset, 89 breast cancer samples have been selected with available miRNA, gene expression and survival data. The second cohort, GSE19783 from the Gene Expression Omnibus (GEO), consist of 78 tamoxifen-treated, estrogen receptor-positive (ER+) breast cancer patients, with miRNA expression and survival data[48]. A summary of these datasets is provided in Table 3. The validation procedure was carefully designed to account for differences in data availability across cohorts and comprised three main steps:

1. *Feature Intersection* For each external dataset, we identified the subset of features present in our 18-biomarker signatures.
2. *Model Re-training* A survival model was trained exclusively on the original TCGA training dataset, restricted to the features available in the corresponding external cohort. This ensured that no external data were used during training.
3. *Performance Evaluation* The re-trained model was applied to the external dataset, and predictive performance was quantified using the concordance index (C-index).

The results are summarized in Table 3. The C-index values in both external cohorts are lower than those reported for the original model (Table 3). Two primary factors likely contribute to this reduced performance:

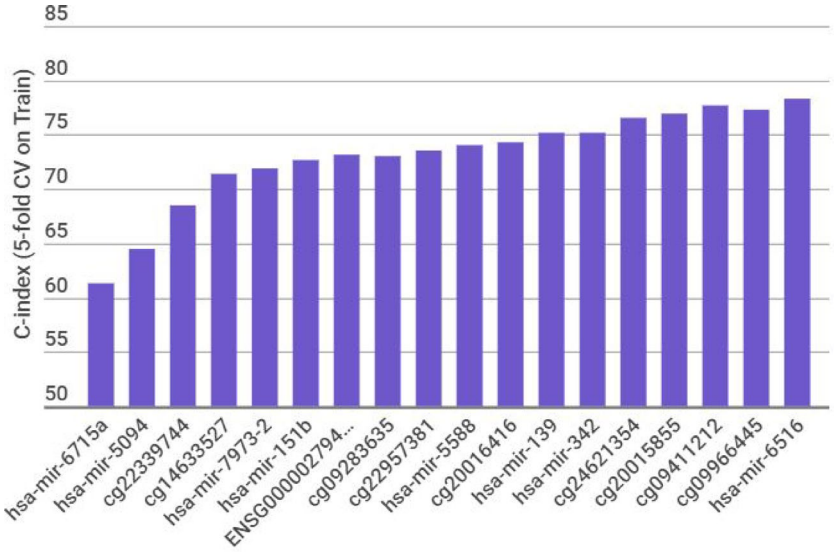

Fig. 7. Progress of gene selection utilizing the SFS algorithm.SFS algorithm.

- *Incomplete feature overlap* Not all predictors in the original signature have been measured in the external cohorts. In each validation set, only a subset of the original features could be used. Retraining the model on this restricted feature set can decrease prognostic accuracy.
- *Dataset shift* The patient populations and data acquisition methods in the external cohorts differ from those in the training cohort. Dataset shift, defined as a mismatch between the training and testing data distributions, can degrade model performance. For example, differences in patient demographics, tumor biology, or assay platforms may alter feature–outcome relationships.

In summary, external validation on PCAWG and GSE19783 confirms that the prognostic signature retains predictive ability, but with a modest decline in performance relative to internal validation. This decline is consistent with the impact of incomplete feature overlap and cohort heterogeneity.

Analysis of the selected genes

To initiate biomarker refinement, we began with the top-performing genetic programming (GP) solution identified in Sect. "Performance evaluation". This model—Root Node [4,3,"GBSurv",18]—achieved a fivefold C-index of 78.31% on the training set and effectively integrated predictive features from miRNA, DNA methylation, and gene expression data. It provided a comprehensive and biologically diverse pool of candidate biomarkers for further selection and interpretation.

To identify the most predictive biomarkers for breast cancer survival, we have implemented a meticulous analytical strategy utilizing the stepwise forward selection (SFS) technique. This method systematically incorporates genes into the predictive model, prioritizing those that significantly improve the concordance index (C-index), a measure that quantifies the predictive accuracy of the model concerning survival outcomes. The SFS algorithm operates by adding genes in a step-wise procedure that when combined with those previously selected ones, yields the highest improvement in the C-index. This iterative selection process ensures that each gene included in the model contributes meaningfully to the overall predictive capability, thereby enhancing the precision and reliability of the model. Figure 7 illustrates the selection process using the TCGA breast cancer dataset. It visually depicts the incremental enhancements in the C-index with the addition of each gene, demonstrating how each selected gene contributes to refining the survival prediction model. This figure not only validates the effectiveness of the SFS technique in identifying robust biomarkers but also highlights the dynamic interplay between different genetic factors in influencing breast cancer prognosis. Through this rigorous selection method, our analysis not only advances the understanding of genetic influences on breast cancer outcomes but also makes a foundation for developing more targeted therapeutic strategies, ultimately aiming to improve patient survival rates based on personalized genetic profiles.

From our analysis, we identified 18 key features that serve as critical factors for analyzing survival in breast cancer samples, as shown in Fig. 7. This selection includes 8 microRNAs, 9 CpG sites, and 1 gene. While some of these biomarkers are newly identified, others have been previously recognized in the literature as key indicators of breast cancer prognosis. This reinforces the reliability of our proposed method in identifying significant features. The following biomarkers mentioned in the literature exemplify this reliability:

- hsa-mir-6715a: identified in several studies as a significant predictor of survival in breast cancer patients[49,50] and warrants further investigation as a key factor in breast cancer survival.
- hsa-mir-5094: Emerging as a significant factor associated with survival in breast cancer, hsa-mir-5094 has been identified as the most highly expressed novel miRNA induced by X-ray irradiation in HeLa cells, targeting STAT5b. Some researches indicate that the upregulation of miR-5094 suppresses STAT5b expression,

leading to decreased transcription of downstream genes involved in cell proliferation and survival[51]. Also shown that BRCA patients with elevated STAT5b expression tends to have better overall survival, relapse free survival, MDFS and survival after disease progression. The level of STAT5b expression may serve as a prognostic indicator, particularly in individuals with progesterone receptor (PR) positivity, HER2 negativity, and a wild-type TP53 gene[52]. These findings imply that hsa-mir-5094 play a critical role in modulating tumor progression, radiation response, and overall prognosis in breast cancer.

- hsa-mir-6516: hsa-mir-6516 has emerged as a significant microRNA implicated in various cancer types, including prostate, renal, and esophageal cancers. Studies have shown that miR-6516 plays a crucial role in regulating tumor cell proliferation, migration, and apoptosis by targeting specific genes such as PDK1 and ODC1[53–55]. Notably, its expression is altered in conditions like muscle atrophy, suggesting potential links to aging and survival mechanisms[56]. hsa-miR-6516 has also been shown to negatively regulates GPX4, a key inhibitor of ferroptosis, by being sponged by lncRNA ADAMTS9-AS1[57]. In breast cancer, especially triple-negative and ER + subtypes, GPX4 overexpression contributes to therapy resistance by preventing ferroptosis. Therefore, miR-6516 may enhance breast cancer cell sensitivity to treatments like gefitinib or CDK4/6 inhibitors by promoting ferroptosis via GPX4 suppression[58,59].
- hsa-mir-342: Extensively studied as a critical biomarker in breast cancer, particularly about patient survival and treatment responses. Research indicates that high expression levels of miR-342 are associated with improved survival outcomes in estrogen receptor-positive patients and those responding to tamoxifen therapy[60]. Additionally, miR-342 has been identified as a regulator of the HER2 signaling pathway, influencing tumor growth and cellular stress responses[61]. Its downregulation has been linked to tamoxifen resistance, underscoring its potential role in treatment efficacy[62]. Furthermore, miR-342 is involved in the regulation of BRCA1 and BRCA2, where it can negatively impact BRCA1 expression, potentially contributing to the pathogenesis of breast cancer, especially in BRCA1-mutant cases[63]. Its involvement in metabolic regulation and cell motility highlights its multifaceted role in breast cancer progression[64].
- hsa-mir-139: This miRNA has emerged as a significant biomarker in breast cancer, with extensive research highlighting its role in tumor progression and patient outcomes. Studies have shown that miR-139-5p is frequently downregulated in invasive breast tumors, correlating with aggressive disease characteristics and worse survival outcomes[65]. Its expression is linked to critical pathways involved in metastasis and cancer stem cell properties, suggesting it may serve as a prognostic marker for invasive breast cancer[66]. Moreover, miR-139-5p has been identified as a potent modulator of radiotherapy response, enhancing treatment efficacy and potentially improving patient survival[67,68]. Collectively, these findings underscore the relevance of hsa-miR-139 in breast cancer prognosis, reinforcing the confidence in our analysis of its role in survival outcomes.
- hsa-miR-151b: This miRNA has been identified as a microRNA of interest in cancer, particularly regarding its potential role in regulating breast cancer metastasis through its association with breast cancer metastasis suppressor 1 (BRMS1)[69].
- hsa-miR-7973: This has been identified as one of the top predicted miRNAs targeting QPRT, a gene found to be highly expressed in breast cancer, particularly in HER2 + subtypes. QPRT is associated with poor prognosis and is involved in pathways such as PI3K-AKT, Wnt signaling, and cell cycle regulation. The regulation of QPRT by miR-7973 suggests a potential role for this miRNA in modulating tumor progression and therapeutic response in breast cancer[70].

Computational cost and scalability

The genetic programming framework was executed on a system with [e.g., an Intel Xeon E5-2690 v4 CPU with 64 GB of RAM]. The total runtime for a single complete run (100 generations with a population of 50) was approximately [e.g., 48 h]. The most computationally intensive step is the fitness evaluation, which requires fivefold cross-validation for each of the 50 chromosomes in every generation. While our framework is powerful for discovering novel integration strategies, its scalability is a consideration. The computational cost grows linearly with population size and number of generations. For datasets with significantly more omics types or samples, strategies such as parallelizing fitness evaluations or using a proxy fitness function for earlier generations may be necessary to maintain feasible runtimes.

Discussion

Our adaptive multi-omics integration framework achieves strong performance in breast cancer survival analysis and outperforms several state-of-the-art deep learning approaches, highlighting the potential of biologically guided feature selection in enhancing clinical predictions. While the results are promising, several areas remain open for further improvement and exploration. First, although the framework demonstrates superiority over baseline and deep learning models, we acknowledge that the margin of improvement relative to some advanced deep learning approaches is modest. This reflects the inherent difficulty of survival prediction and suggests that hybrid strategies combining our approach with deep learning could yield further gains. Second, as with many evolutionary optimization methods, genetic programming (GP) may produce complex models that are not easily interpretable from a biological perspective. Improving interpretability through parsimony constraints, feature grouping, or pathway-level integration remains an important direction. Third, our preprocessing step restricted features to those with high variance to reduce complexity and computational cost. However, this solution may exclude low-variance but biologically informative features, such as tumor suppressors with stable expression. Future research could investigate more nuanced filtering strategies that balance variance with biological relevance. Fourth, we did not yet perform stratified analyses by molecular subtypes or tumor grades, which could uncover subtype-specific biomarker signatures and integration strategies, especially considering the heterogeneous nature of breast cancer. Fifth, while the C-index remains a widely used and accepted metric in survival analysis, incorporating complementary metrics—such as the Brier score—could offer additional insights

into model calibration and reliability. Sixth, while our study focused on integrating genomics, transcriptomics, and epigenomics data, certain other data types available in TCGA, such as whole slide images, exon expression, SNP data, and phenotype information like patient age were not included in our analysis. Incorporating these data types in future research could further enhance the predictive accuracy of survival models and provide a more comprehensive insight into breast cancer biology. Moreover, our use of a single-objective genetic programming approach does not explicitly balance the trade-off between model complexity and predictive performance. Future work could benefit from exploring multi-objective optimization strategies, such as Multi-Objective Genetic Algorithms (MOGAs), to yield more parsimonious yet effective models. Although our method does not explicitly address inter-omics redundancy, we note that such redundancy can sometimes reflect convergent biological signals, which may be meaningful. Nonetheless, incorporating mechanisms to distinguish between informative and uninformative redundancy, as well as leveraging prior biological knowledge (e.g., pathways or gene sets) during feature selection, may enhance interpretability. Finally, the computational demands of the current approach pose challenges for scaling to larger datasets. Optimization strategies, such as parallelized evolution or surrogate fitness models, offer promising directions for improving efficiency. Collectively, these considerations represent meaningful opportunities for advancing multi-omics integration frameworks toward greater clinical utility and biological insight.

Conclusion

In this study, we presented an adaptive multi-omics integration framework for breast cancer survival analysis, leveraging data from The Cancer Genome Atlas (TCGA). By combining genomics, transcriptomics, and epigenomics data, we can identify complex molecular signatures that significantly impact breast cancer progression and patient survival outcomes. The integration and feature selection processes have been optimized using genetic programming, enabling us to evolve solutions for complex datasets. This approach allowed for identifying key biomarkers contributing to more accurate survival predictions. Our framework demonstrated superior performance compared to other state-of-the-art methods, including those by Chai et al., Tong et al., and Benkirane et al., Specifically, our method achieved a concordance index (C-index) of 78.31 on the training data and 67.94 on the test data, outperforming these benchmark studies. Additionally, through stepwise forward selection, we introduced several novel genes that had not been previously identified in breast cancer survival studies. These newly identified genes provide valuable insights into the genetic and epigenetic factors influencing breast cancer outcomes and open new avenues for further research. The findings of this study underscore the significance of acknowledging the complex interactions among different molecular layers in the realm of cancer biology. Our adaptive framework is not only effective for breast cancer but also offers a scalable and flexible approach that can be applied to other cancer types. By integrating diverse omics data, we can achieve a deeper understanding of oncological processes, ultimately contributing to the development of more personalized and targeted therapeutic strategies. In summary, our research emphasizes the need for innovative, multi-omics-based approaches in cancer prognosis and introduces new biomarkers that improve survival predictions. This framework paves the way for future studies to incorporate additional data types and further refine cancer survival analysis, advancing the field of cancer research.

Data availability

The datasets analyzed in this study originate from the TCGA Breast Cancer (BRCA) cohort and can be accessed and downloaded from the UCSC Xena Browser repository at the following link: https://xenabrowser.net/datapages/?cohort=GDC%20TCGA%20Breast%20Cancer%20(BRCA)&removeHub=https%3A%2F%2Fxena.treehouse.gi.ucsc.edu%3A443.

Received: 8 February 2025; Accepted: 26 September 2025
Published online: 03 November 2025

References

1. Zhu, J. W., Charkhchi, P., Adekunte, S. & Akbari, M. R. What is known about breast cancer in young women. *Cancers* **15**(6), 1917. https://doi.org/10.3390/CANCERS15061917 (2023).
2. Ochoa, S. & Hernández-Lemus, E. Molecular mechanisms of multi-omic regulation in breast cancer. *Front. Oncol.* **13**, 1148861. https://doi.org/10.3389/FONC.2023.1148861 (2023).
3. Turner, K. M., Yeo, S. K., Holm, T. M., Shaughnessy, E. & Guan, J. L. Dynamic tumor heterogeneity and cancer progression: Heterogeneity within molecular subtypes of breast cancer. *Am. J. Physiol. Cell Physiol.* **321**(2), C343. https://doi.org/10.1152/AJPCELL.00109.2021 (2021).
4. Chaitankar, V. et al. Next generation sequencing technology and genomewide data analysis: Perspectives for retinal research. *Prog. Retin. Eye Res.* **55**, 1–31. https://doi.org/10.1016/J.PRETEYERES.2016.06.001 (2016).
5. Khella, C. A., Mehta, G. A., Mehta, R. N. & Gatza, M. L. Recent advances in integrative multi-omics research in breast and ovarian cancer. *J. Pers. Med.* **11**(2), 149. https://doi.org/10.3390/JPM11020149 (2021).
6. Momeni, Z., Hassanzadeh, E., Saniee Abadeh, M. & Bellazzi, R. A survey on single and multi omics data mining methods in cancer data classification. *J. Biomed. Inform.* **107**, 103466. https://doi.org/10.1016/j.jbi.2020.103466 (2020).
7. Gao, J. et al. "Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.* **6**(269), pl1–pl1. https://doi.org/10.1126/SCISIGNAL.2004088/SUPPL_FILE/2004088_TABLES2.XLS (2013).
8. Perou, C. M. et al. Molecular portraits of human breast tumours. *Nat.* **406**(6797), 747–752. https://doi.org/10.1038/35021093 (2000).
9. Ma, S., Ren, J., Fenyö, D. Breast cancer prognostics using multi-omics data. *AMIA Summits Transl. Sci. Proc.*, 2016, 52, (2016), Accessed: Sep. 20, 2024. [Online]. Available: /pmc/articles/PMC5001766/
10. Cai, Z., Poulos, R. C., Liu, J. & Zhong, Q. Machine learning for multi-omics data integration in cancer. *iScience*. https://doi.org/10.1016/J.ISCI.2022.103798/ASSET/00667FF9-5AAA-4E60-ACFE-AE79325B0C9C/MAIN.ASSETS/GR3.JPG (2022).
11. Lin, E. & Lane, H. Y. Machine learning and systems genomics approaches for multi-omics data. *Biomark. Res.* **5**(1), 1–6. https://doi.org/10.1186/S40364-017-0082-Y/FIGURES/3 (2017).

12. Wang, P., Li, Y. & Reddy, C. K. Machine learning for survival analysis: A survey. *ACM Comput. Surv.* **51**(6), 1–36. https://doi.org/10.1145/3214306/SUPPL_FILE/WANG.ZIP (2019).
13. Lai, J., Wang, H., Peng, J., Chen, P. & Pan, Z. Establishment and external validation of a prognostic model for predicting disease-free survival and risk stratification in breast cancer patients treated with neoadjuvant chemotherapy. *Cancer Manag. Res.* **10**, 2347–2356. https://doi.org/10.2147/CMAR.S171129 (2018).
14. Kumar, A., Sinha, N., Bhardwaj, A. & Goel, S. Clinical risk assessment of chronic kidney disease patients using genetic programming. *Comput. Methods Biomech. Biomed. Engin.* **25**(8), 887–895. https://doi.org/10.1080/10255842.2021.1985476 (2022).
15. D'Angelo, G. et al. Identifying patterns in multiple biomarkers to diagnose diabetic foot using an explainable genetic programming-based approach. *Futur. Gener. Comput. Syst.* **140**, 138–150. https://doi.org/10.1016/J.FUTURE.2022.10.019 (2023).
16. Espejo, P. G., Ventura, S. & Herrera, F. A survey on the application of genetic programming to classification. *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.* **40**(2), 121–144. https://doi.org/10.1109/TSMCC.2009.2033566 (2010).
17. Wang, T. et al. MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification. *Nat. Commun.* **12**(1), 1–13. https://doi.org/10.1038/S41467-021-23774-W (2021).
18. Lu, Y. et al. Multiomics dynamic learning enables personalized diagnosis and prognosis for pancancer and cancer subtypes. *Brief. Bioinform.* **24**(6), 1–16. https://doi.org/10.1093/BIB/BBAD378 (2023).
19. Lin, Y., Zhang, W., Cao, H., Li, G. & Du, W. Classifying breast cancer subtypes using deep neural networks based on multi-omics data. *Genes* **11**(8), 888. https://doi.org/10.3390/GENES11080888 (2020).
20. Choi, J. M. & Chae, H. moBRCA-net: a breast cancer subtype classification framework based on multi-omics attention neural networks. *BMC Bioinformatics* **24**(1), 1–15. https://doi.org/10.1186/s12859-023-05273-5 (2023).
21. Mohaiminul Islam, M. et al. An integrative deep learning framework for classifying molecular subtypes of breast cancer. *Comput. Struct. Biotechnol. J.* **18**, 2185–2199. https://doi.org/10.1016/J.CSBJ.2020.08.005 (2020).
22. Fan, Y. et al. Integrated multi-omics analysis model to identify biomarkers associated with prognosis of breast cancer. *Front. Oncol.* **12**, 899900. https://doi.org/10.3389/FONC.2022.899900/BIBTEX (2022).
23. Poirion, O. B., Jing, Z., Chaudhary, K., Huang, S. & Garmire, L. X. DeepProg: an ensemble of deep-learning and machine-learning models for prognosis prediction using multi-omics data. *Genome Med.* **13**(1), 1–15. https://doi.org/10.1186/S13073-021-00930-X/FIGURES/6 (2021).
24. Bichindaritz, I., Liu, G. & Bartlett, C. Integrative survival analysis of breast cancer with gene expression and DNA methylation data. *Bioinformatics* **37**(17), 2601–2608. https://doi.org/10.1093/BIOINFORMATICS/BTAB140 (2021).
25. Ouyang, D. et al. Integration of multi-omics data using adaptive graph learning and attention mechanism for patient classification and biomarker identification. *Comput. Biol. Med.* **164**, 107303. https://doi.org/10.1016/J.COMPBIOMED.2023.107303 (2023).
26. Cheng, L. et al. MoAGL-SA: a multi-omics adaptive integration method with graph learning and self attention for cancer subtype classification. *BMC Bioinformatics* **25**(1), 1–19. https://doi.org/10.1186/S12859-024-05989-Y/FIGURES/5 (2024).
27. Chaudhary, K., Poirion, O. B., Lu, L. & Garmire, L. X. Deep learning–based multi-omics integration robustly predicts survival in liver cancer. *Clin. Cancer Res.* **24**(6), 1248–1259. https://doi.org/10.1158/1078-0432.CCR-17-0853/116627/AM/DEEP-LEARNING-BASED-MULTI-OMICS-INTEGRATION (2018).
28. Lv, J., Wang, J., Shang, X., Liu, F. & Guo, S. Survival prediction in patients with colon adenocarcinoma via multiomics data integration using a deep learning algorithm. *Biosci. Rep.* **40**(12), BSR20201482. https://doi.org/10.1042/BSR20201482/227089 (2020).
29. Yu, J. et al. A model for predicting prognosis in patients with esophageal squamous cell carcinoma based on joint representation learning. *Oncol. Lett.* **20**(6), 1–1. https://doi.org/10.3892/OL.2020.12250/HTML (2020).
30. Zhang, X. et al. Robust prognostic subtyping of muscle-invasive bladder cancer revealed by deep learning-based multi-omics data integration. *Front. Oncol.* **11**, 689626. https://doi.org/10.3389/FONC.2021.689626/BIBTEX (2021).
31. Liu, C., Wang, X., Genchev, G. Z. & Lu, H. Multi-omics facilitated variable selection in Cox-regression model for cancer prognosis prediction. *Methods* **124**, 100–107. https://doi.org/10.1016/J.YMETH.2017.06.010 (2017).
32. Argelaguet, R. et al. MOFA+: A statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biol.* **21**(1), 1–17. https://doi.org/10.1186/S13059-020-02015-1/FIGURES/4 (2020).
33. Hargadon, K. M. Using the cancer genome atlas as a tool to improve undergraduate student understanding of cancer genetics and the hallmarks of cancer progression. *J. Cancer Educ.* **37**(5), 1357–1363. https://doi.org/10.1007/S13187-021-01962-Y/METRICS (2022).
34. Gulyaeva, L. F. & Kushlinskiy, N. E. Regulatory mechanisms of microRNA expression. *J. Transl. Med.* **14**(1), 1–10. https://doi.org/10.1186/S12967-016-0893-X (2016).
35. Mulrane, L., McGee, S. F., Gallagher, W. M. & O'Connor, D. P. miRNA dysregulation in breast cancer. *Cancer Res.* **73**(22), 6554–6562. https://doi.org/10.1158/0008-5472.CAN-13-1841/659060/P/MIRNA-DYSREGULATION-IN-BREAST-CANCERMIRNAS-AND (2013).
36. Creighton, C. J. Gene expression profiles in cancers and their therapeutic implications. *Cancer J. (United States)* **29**(1), 9–14. https://doi.org/10.1097/PPO.0000000000000638 (2023).
37. Nishiyama, A. & Nakanishi, M. Navigating the DNA methylation landscape of cancer. *Trends Genet.* **37**(11), 1012–1027. https://doi.org/10.1016/J.TIG.2021.05.002/ASSET/60367FD5-301C-488E-98CF-0DEA023A3825/MAIN.ASSETS/GR3.JPG (2021).
38. Troyanskaya, O. et al. Missing value estimation methods for DNA microarrays. *Bioinformatics* **17**(6), 520–525. https://doi.org/10.1093/BIOINFORMATICS/17.6.520 (2001).
39. Janikow, C. Z. Adapting representation in genetic programming. *Lect. Notes Comput. Sci. (Including Subser Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)* **3103**, 507–518. https://doi.org/10.1007/978-3-540-24855-2_61 (2004).
40. Chandrashekar, G. & Sahin, F. A survey on feature selection methods. *Comput. Electr. Eng.* **40**(1), 16–28. https://doi.org/10.1016/J.COMPELECENG.2013.11.024 (2014).
41. Lazar, C. et al. A survey on filter techniques for feature selection in gene expression microarray analysis. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* **9**(4), 1106–1119. https://doi.org/10.1109/TCBB.2012.33 (2012).
42. Chen, Y., Jia, Z., Mercola, D. & Xie, X. A gradient boosting algorithm for survival analysis via direct optimization of concordance index. *Comput. Math. Methods Med.* **1**, 873595. https://doi.org/10.1155/2013/873595 (2013).
43. Chansky, K., Subotic, D., Foster, N. R. & Blum, T. Survival analyses in lung cancer. *J. Thorac. Dis.* **8**(11), 3457. https://doi.org/10.21037/JTD.2016.11.28 (2016).
44. Chai, H. et al. Integrating multi-omics data through deep learning for accurate cancer prognosis prediction. *Comput. Biol. Med.* **134**, 104481. https://doi.org/10.1016/J.COMPBIOMED.2021.104481 (2021).
45. Tong, L., Mitchel, J., Chatlin, K. & Wang, M. D. Deep learning based feature-level integration of multi-omics data for breast cancer patients survival analysis. *BMC Med. Inform. Decis. Mak.* **20**(1), 1–12. https://doi.org/10.1186/S12911-020-01225-8/FIGURES/6 (2020).
46. Benkirane, H., Pradat, Y., Michiels, S. & Cournède, P. H. CustOmics: A versatile deep-learning based strategy for multi-omics integration. *PLOS Comput. Biol.* **19**(3), e1010921. https://doi.org/10.1371/JOURNAL.PCBI.1010921 (2023).
47. Goldman, M. J. et al. A user guide for the online exploration and visualization of PCAWG data. *Nat. Commun.* **11**(1), 1–9. https://doi.org/10.1038/S41467-020-16785-6 (2020).
48. Enerly, E. et al. miRNA-mRNA integrated analysis reveals roles for miRNAs in primary breast tumors. *PLoS ONE* **6**(2), e16915. https://doi.org/10.1371/JOURNAL.PONE.0016915 (2011).

49. Sang, M. et al. Identification of three miRNAs signature as a prognostic biomarker in breast cancer using bioinformatics analysis. *Transl. Cancer Res.* **9**(3), 1884. https://doi.org/10.21037/TCR.2020.02.21 (2020).
50. Lai, J., Wang, H., Pan, Z. & Su, F. A novel six-microRNA-based model to improve prognosis prediction of breast cancer. *Aging (Albany NY)* **11**(2), 649. https://doi.org/10.18632/AGING.101767 (2019).
51. Ding, N. et al. The role of MiR-5094 as a proliferation suppressor during cellular radiation response via downregulating STAT5b. *J. Cancer* **11**(8), 2222. https://doi.org/10.7150/JCA.39679 (2020).
52. Li, J. et al. Identification of STAT5B as a biomarker associated with prognosis and immune infiltration in breast cancer. *Med. (United States)* **102**(9), E32972. https://doi.org/10.1097/MD.0000000000032972 (2023).
53. Ma, Y. et al. PDK1 regulates the progression of esophageal squamous cell carcinoma through metabolic reprogramming. *Mol. Carcinog.* **62**(6), 866–881. https://doi.org/10.1002/MC.23531 (2023).
54. Huang, G. et al. miRNA-6516–5p regulates the proliferation and migration of renal cancer cells by targeting ODC1. *Int. J. Surg.* https://doi.org/10.3760/CMA.J.CN115396-20210601-00200 (2022).
55. Wu, X., Xiao, Y., Zhou, Y., Zhou, Z. & Yan, W. lncRNA SNHG20 promotes prostate cancer migration and invasion via targeting the miR-6516–5p/SCGB2A1 axis. *Am. J. Transl. Res.* **11**(8), 5162 (2019).
56. Jung, W. et al. Identifying the potential therapeutic effects of miR-6516 on muscle disuse atrophy. *Mol. Med. Rep.* **30**(1), 119. https://doi.org/10.3892/mmr.2024.13243 (2024).
57. Wan, Y. et al. Long noncoding RNA ADAMTS9-AS1 represses ferroptosis of endometrial stromal cells by regulating the miR-6516–5p/GPX4 axis in endometriosis. *Sci. Rep.* **12**(1), 2618. https://doi.org/10.1038/S41598-022-04963-Z (2022).
58. Song, X., Wang, X., Liu, Z. & Yu, Z. Role of GPX4-mediated ferroptosis in the sensitivity of triple negative breast cancer cells to gefitinib. *Front. Oncol.* **10**, 597434. https://doi.org/10.3389/FONC.2020.597434/BIBTEX (2020).
59. Herrera-Abreu, M. T. et al. Inhibition of GPX4 enhances CDK4/6 inhibitor and endocrine therapy activity in breast cancer. *Nat. Commun.* **15**(1), 1–19. https://doi.org/10.1038/s41467-024-53837-7 (2024).
60. Young, J. et al. Tamoxifen sensitivity-related microRNA-342 is a useful biomarker for breast cancer survival. *Oncotarget* **8**(59), 99978. https://doi.org/10.18632/ONCOTARGET.21577 (2017).
61. Lindholm, E. M. et al. miR-342-5p as a potential regulator of HER2 breast cancer cell growth. *MicroRNA* **8**(2), 155–165. https://doi.org/10.2174/2211536608666181206124922 (2018).
62. Cittelly, D. M. et al. Downregulation of miR-342 is associated with tamoxifen resistant breast tumors. *Mol. Cancer* **9**(1), 1–12. https://doi.org/10.1186/1476-4598-9-317/FIGURES/6 (2010).
63. Crippa, E. et al. miR-342 regulates BRCA1 expression through modulation of ID4 in breast cancer. *PLoS ONE* **9**(1), e87039. https://doi.org/10.1371/JOURNAL.PONE.0087039 (2014).
64. Romero-Cordoba, S. L. et al. Abstract A47: A microRNA signature identifies subtypes of triple-negative breast cancer and reveals miR-342-3p as regulator of a lactate metabolic pathway through silencing monocarboxylate transporter 1. *Cancer Res.* **76**(6), A47–A47. https://doi.org/10.1158/1538-7445.NONRNA15-A47 (2016).
65. Krishnan, K. et al. miR-139-5p is a regulator of metastatic pathways in breast cancer. *RNA* **19**(12), 1767–1780. https://doi.org/10.1261/RNA.042143.113 (2013).
66. Cheng, C. W. et al. MiR-139 modulates cancer stem cell function of human breast cancer through targeting CXCR4. *Cancers (Basel)* **13**(11), 2582. https://doi.org/10.3390/CANCERS13112582/S1 (2021).
67. Pajic, M. et al. miR-139-5p modulates radiotherapy resistance in breast cancer by repressing multiple gene networks of DNA repair and ROS defense. *Cancer Res.* **78**(2), 501–515. https://doi.org/10.1158/0008-5472.CAN-16-3105/652748/AM/MIR-139-5P-MODULATES-RADIOTHERAPY-RESISTANCE-IN (2018).
68. Dai, H. et al. Role of miR-139 as a surrogate marker for tumor aggression in breast cancer. *Hum. Pathol.* **61**, 68–77. https://doi.org/10.1016/J.HUMPATH.2016.11.001 (2017).
69. Li, G. et al. MicroRNA-3200-5p promotes osteosarcoma cell invasion via suppression of BRMS1. *Mol. Cells* **41**(6), 523–531. https://doi.org/10.14348/MOLCELLS.2018.2200 (2018).
70. Yan, Y. et al. A comprehensive analysis of the role of QPRT in breast cancer. *Sci. Rep.* **13**(1), 1–14. https://doi.org/10.1038/S41598-023-42566-4 (2023).

Author contributions

Esmaeil Hasanzadeh: Responsible for researching, coding, and writing the manuscript. Nasrollah Moghadam Charkari: the supervisor, contributed to the design of the methods and revised and edited the manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to N.M.C.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.