frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 02 August 2024
DOI 10.3389/fgene.2024.1378809

OPEN ACCESS

EDITED BY
Zhi-Ping Liu,
Shandong University, China

REVIEWED BY
Advait Balaji,
Occidental Petroleum Corporation, United States
Xuefeng Cui,
Shandong University, China

*CORRESPONDENCE
Huimin Luo,
luohuimin@henu.edu.cn

RECEIVED 30 January 2024
ACCEPTED 22 July 2024
PUBLISHED 02 August 2024

CITATION
Zhang G, Ma C, Yan C, Luo H, Wang J, Liang W and Luo J (2024), MSFN: a multi-omics stacked fusion network for breast cancer survival prediction.
Front. Genet. 15:1378809.
doi: 10.3389/fgene.2024.1378809

MSFN: a multi-omics stacked fusion network for breast cancer survival prediction

Ge Zhang[1,2,3], Chenwei Ma[2], Chaokun Yan[1,2,3], Huimin Luo[1,2,3]*, Jianlin Wang[1,2,3], Wenjuan Liang[1,2,3] and Junwei Luo[4]

[1]Academy for Advanced Interdisciplinary Studies, Henan University, Kaifeng, Henan, China, [2]School of Computer and Information Engineering, Henan University, Kaifeng, Henan, China, [3]Henan Key Laboratory of Big Data Analysis and Processing, Henan University, Kaifeng, Henan, China, [4]College of Computer Science and Technology, Henan Polytechnic University, Jiaozuo, Henan, China

Introduction: Developing effective breast cancer survival prediction models is critical to breast cancer prognosis. With the widespread use of next-generation sequencing technologies, numerous studies have focused on survival prediction. However, previous methods predominantly relied on single-omics data, and survival prediction using multi-omics data remains a significant challenge.

Methods: In this study, considering the similarity of patients and the relevance of multi-omics data, we propose a novel multi-omics stacked fusion network (MSFN) based on a stacking strategy to predict the survival of breast cancer patients. MSFN first constructs a patient similarity network (PSN) and employs a residual graph neural network (ResGCN) to obtain correlative prognostic information from PSN. Simultaneously, it employs convolutional neural networks (CNNs) to obtain specificity prognostic information from multi-omics data. Finally, MSFN stacks the prognostic information from these networks and feeds into AdaboostRF for survival prediction.

Results: Experiments results demonstrated that our method outperformed several state-of-the-art methods, and biologically validated by Kaplan-Meier and t-SNE.

KEYWORDS

deep learning, breast cancer survival prediction, multi-omics data, residual graph neural network, convolutional neural network, stacking integration

1 Introduction

According to the Global Cancer Statistics 2020, 2.26 million new cases of breast cancer were diagnosed in 2020, and the deaths from breast cancer were in the fifth rank of all cancers (Sung et al., 2021). Breast cancer has become the most prevalent cancer in the world (Arnold et al., 2022). Survival prediction is an essential part of cancer prognosis. It aims to predict the survival risk of cancer patients and provide recommendations for pathologists and doctors in treatment (Hagerty et al., 2005). Accurate and reliable survival prediction can provide doctors with scientific guidance and improve the survival rate of patients. More importantly, the survival prediction tools could formulate reasonable treatment strategies for patients, avoid unnecessary pain caused by over-treatment, and improve the quality of life of patients. Meanwhile, it reduces the burden of doctors and avoids the wastage of medical resources (Deepa and Gunavathi, 2022). Therefore, developing accurate and reliable survival prediction methods is vital for the treatment and prognosis of breast cancer.

With the widespread application of next-generation sequencing technologies and the accumulation of medical data on cancers, plenty of survival prediction methods have been developed, including (i) statistical survival analysis methods and (ii) machine learning-based methods. Statistical survival analysis methods such as CoxPH and LogRank test use survival data and a few covariates to predict patient survival (Michaelson et al., 2002). However, these methods are difficult to model and not applicable to analyzing large amounts of data. Machine learning-based methods effectively address these limits of statistical survival analysis methods. Algorithms such as Support Vector Machines (SVM), Random Forest (RF) and Logistic Regression (LR) obtain prognostic features from large amounts of cancer data to predict survival (Xu et al., 2012). However, machine learning-based methods require researchers to perform laborious and complex feature engineering work.

In recent years, deep learning methods have provided scientists with powerful tools for extracting high-quality prognostic information from massive omics data and have been proven effective in survival prediction (LeCun et al., 2015; Deepa and Gunavathi, 2022). For instance, Ching et al. (2018) developed a neural network model Cox-nnet with a Cox regression layer to predict survival using RNA-Seq data. Katzman et al. (2018) proposed DeepSurv to predict patient survival and the effect of covariates on patient survival risk by combining DNN and Cox-PH. However, the human genome is extremely complex, and various factors influence cancer pathogenesis (Lujambio and Lowe, 2012). Multi-omics data contains a wealth of information, providing an unprecedented opportunity to investigate the occurrence and progression from multiple perspectives (Arjmand et al., 2022). But the deep learning survival prediction methods described above are inapplicable to multi-omics data. Therefore, deep learning methods based on multi-omics data have risen to prominence in survival prediction (Herrmann et al., 2021; Kang et al., 2022).

One kind of survival prediction research predicts patients' survival risk (survival rate) based on their survival time and survival status. For example, Cheerla et al. proposed introducing the COX loss function in the deep learning model to fusion clinical data, gene expression data, microRNA expression data, and WSIs (Whole Slide Images) to predict the survival rate of patients with 20 cancers (Cheerla and Gevaert, 2019). Li et al. (2022) proposed HFBSurv to predict patient survival by employing a factorized bilinear model to fuse gene expression, CNV, and pathology image features step by step. Another survival prediction research predicts the long and short survival of cancer patients. For instance, Sun et al. proposed MDNNMD, a survival prediction model that integrates clinical, CNV, and gene expression data of breast cancer by fusing three DNNs with different weights (Sun et al., 2018). AMDN extracts prognostic features of clinical and gene expression data using NMF matrix decomposition combined with attention mechanisms to predict breast cancer survival (Chen et al., 2019). Subsequently, Arya et al. proposed a stacked integration model STACKED RF to overcome the limitation that MDNNMD requires manual adjustment of fusion weights (Arya and Saha, 2020). In the follow-up research, they introduced a gated attention mechanism into STACKED RF to enhance the prediction performance, named SiGaAtCNN (Arya and Saha, 2021). However, previous survival prediction studies based on multi-omics data focus on extracting prognostic features from various multi-omics data, rather than patient similarity and correlation of multi-omics data.

TABLE 1 Overview of the dataset.

Description	Value
Threshold (years)	5
Total patients	1048
Long-time survivors	248
Short-time survivors	800
Average survival (months)	42.34

To address these issues in classification prediction studies of long and short survival, we propose a novel Multi-omics Stacked Fusion Network (MSFN) for breast cancer survival prediction. First, we construct a patient similarity network using multi-omics data. Then, we employ ResGCN to obtain similarity information of patients and correlation information of multi-omics data. Simultaneously, we construct CNNs for each omics data to obtain the specificity information. Finally, we stack the prognostic information from the hidden layers of the networks and utilize AdaboostRF for survival prediction. The superiority of MSFN is to comprehensively consider the specificity information of multi-omics data, the similarity information of patients, and the correlation information of multi-omics data, stacking these information to achieve more accurate and reliable survival prediction. The contributions of this work are summarized as follows:

- We propose a novel multi-omics stacked fusion network framework that comprehensively obtains survival-related information from multi-omics data for survival prediction.
- We integrate multi-omics data with Similarity Network Fusion (SNF) that sufficiently utilizes the similarity between patients and the correlation of multi-omics data to generate a comprehensive patient similarity network.
- We use ResGCN to extract the prognostic information of the patient similarity network, leveraging its residual connectivity to achieve a deeper network structure while effectively addressing gradient vanishing.

2 Materials and methods

2.1 Datasets and preprocessing

To investigate the performance of our method, we conducted comprehensive and rigorous experiments on the BRCA multi-omics dataset from TCGA (The Cancer Genome Atlas). We obtained this dataset from the UCSC Xena platform (http://xena.ucsc.edu/) and removed samples and features with missing values above 20%. 1048 patient samples were finally selected, each sample contained clinical, gene expression, CNV, and survival data. This is because

TABLE 2 Feature selection.

Data type	Total features	Selected features
Clinical	190	33
Gene Expression	60,488	400
CNV	19,729	200

clinical, gene expression and CNV data are highly associated with cancer occurrence and progression, and they have been used extensively in previous survival prediction studies (Shlien and Malkin, 2009; Li et al., 2017; Kalafi et al., 2019). Then, we divided patients into long-term and short-term survivors using a threshold of 5-year survival, with long-term survivors labeled as 1 and short-term survivors labeled as 0. The overview description of the dataset is shown in Table 1.

For the clinical data, we first removed not reported data and features and samples with more than 20% missing values. Then, we removed irrelevant text descriptions, markers, and years from the clinical data. Subsequently, according to the data processing procedure in the study by Sun et al. (2018); Arya and Saha (2020); Arya and Saha (2021), we screened clinical features such as age, tumor size, and TNM stage, and performed label coding and binarization for the categorical features. Finally, we obtained 33 features as clinical features. Since there were no missing values in the gene expression and CNV data, we only estimated missing values for clinical data. Specifically, we divided the 33 clinical features into 24 discrete-valued features and 9 continuous-valued features. For continuous features, we use the k-Nearest Neighbor algorithm (KNN) for interpolation then normalized them using the min-max normalization with the range set to [0,1] (Troyanskaya et al., 2001; Patro and Sahu, 2015). For the discrete features we used the mode interpolation (García-Laencina et al., 2015). For gene expression data, we also used the max-min normalization for normalization with the range set to [0,1]. For CNV data, we directly use the discretized raw data. The gene expression and CNV data for each patient in the dataset has 60,488 and 19,729 features. This high dimensionality of data leads to the "dimensionality catastrophe" that negatively affects the performance of deep learning methods (Berisha et al., 2021). Therefore, we used the renowned mRMR algorithm for feature selection (Peng et al., 2005). Then, we searched the optimal number of gene expression and CNV features in steps of 100 (Arya and Saha, 2020; Arya and Saha, 2021). Finally, we selected 400 gene expression features, 200 CNV features, and all 33 clinical features as model inputs, as shown in Table 2.

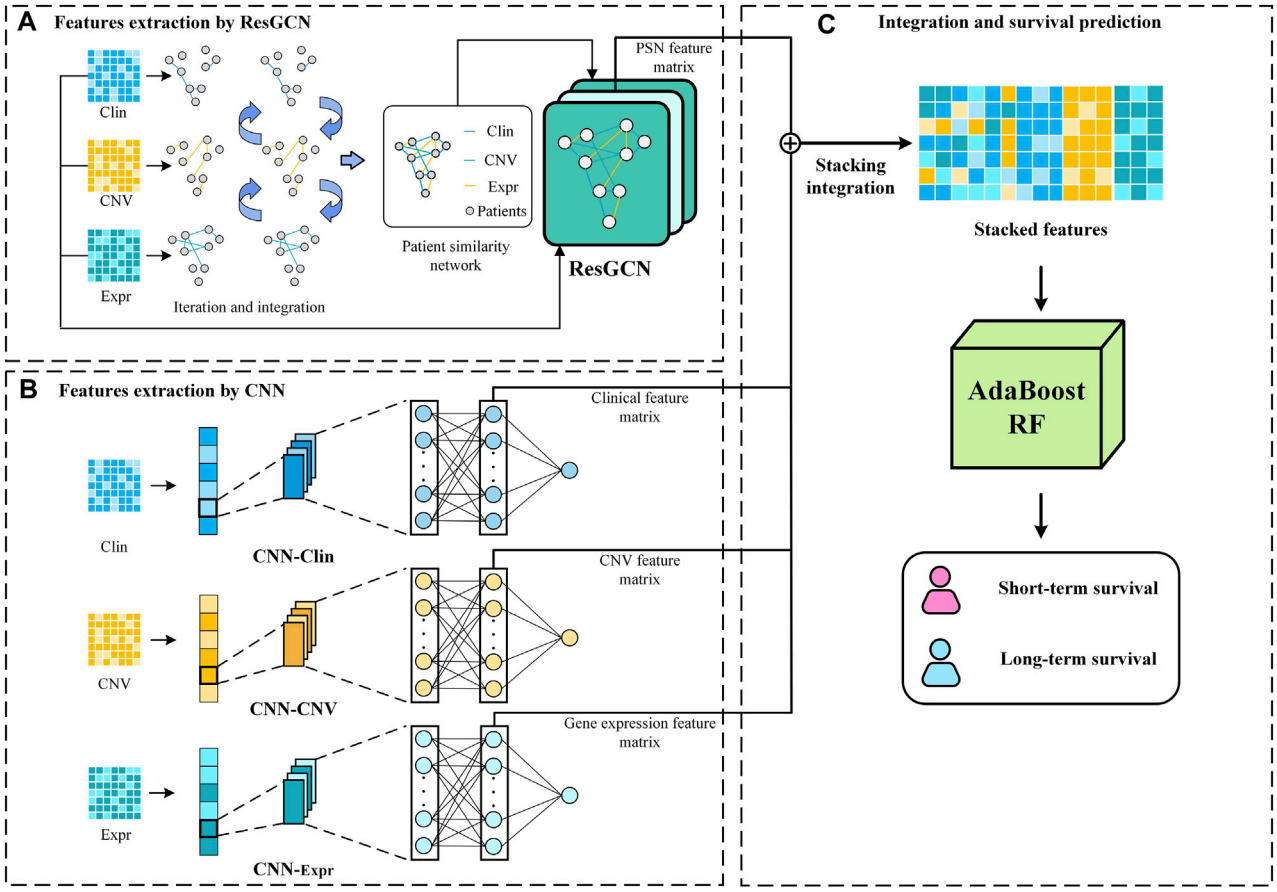

FIGURE 1
The framework of MSFN.

2.2 Methods

The proposed MSFN consists of three components. In the first component, MSFN constructs patient similarity networks using SNF and employs ResGCN to obtain similarity information of patients and correlation information of multi-omics data. In the second component, MSFN constructs CNNs for each omics data to obtain the specificity prognostic information. The last component is extracting and stacking the prognostic information of ResGCN and CNNs, feeding them into AdaboostRF for survival prediction. The framework of MSFN is briefly shown in Figure 1. The implementation of our method is available at https://github.com/AckerMuse/MSFN.

2.2.1 A: Features extraction by ResGCN

2.2.1.1 Construction of patient similarity network

In order to construct the patient similarity network, we employ the similarity network fusion (SNF) to construct the patient similarity network (Wang et al., 2014). SNF can integrate multi-omics data from clinical, CNV and gene expression data to generate a comprehensive patient similarity network for fully leverages patients' similarities and the correlation of multi-omics data. Assuming there are n patients, each patient has m types of data. We denote the patient similarity network as a graph $G = (V, E)$, where V represents the set of patients,i.e.,$\{x_1, x_2, x_3, \ldots, x_n\}$. The edge E corresponds to the similarity relation between vertices $v \in V$ in the graph. The weights of these edges are represented by an $n \times n$ similarity matrix W, which is computed by Eq. 1:

$$W(i,j) = \exp\left(-\frac{\vartheta^2(x_i, x_j)}{\lambda \varepsilon_{i,j}}\right) \tag{1}$$

where λ is the hyperparameter, $\vartheta(x_i, x_j)$ is the euclidean distance between patients x_i and x_j, and $\varepsilon_{i,j}$ is used to eliminate the scaling problem (Wang et al., 2014). Then, the similarity matrix is normalized by Eq. 2:

$$P_{i,j} = \begin{cases} \dfrac{W_{i,j}}{2\sum_{k \neq i} W_{i,k}}, j \neq i \\ 1/2, j = i \end{cases} \tag{2}$$

Suppose N_i is the set of neighbor nodes of x_i. We can calculate the similarity matrix L of the single omics data by Eq. 3:

$$L_{i,j} = \begin{cases} \dfrac{W_{i,j}}{\sum_{k \in N_i} W_{j,k}}, j \in N_i \\ 0, otherwise \end{cases} \tag{3}$$

Let $P_t^{(v)}$ denote the similarity matrix after normalization of the v-th omics data $(0 < v \leq m)$ in the t-th iteration. Update $P_t^{(v)}$ according to Eq. 4:

$$P_{t+1}^{(v)} = L^{(v)}\left(\frac{\sum_{k \neq v} P_t^{(k)}}{m-1}\right)\left(L^{(v)}\right)^T \tag{4}$$

where the $L^{(v)}$ denotes the local similarity matrix of the v-th omics data. Through continuous iterative fusion, the SNF ultimately generates a patient similarity network containing correlation information from all omics data. In this work, the patient similarity network is combined with ResGCN for cancer survival prediction.

2.2.1.2 Similarity and correlation features extraction by ResGCN

Since the patient similarity network constructed by SNF is graph-structured data, we employ ResGCN to obtain the survival prediction features from it (Li et al., 2019). ResGCN modifies the data transmission mechanism in graph neural networks to mitigate the gradient vanishing problem and overcome the limitation that graph neural networks cannot construct deep networks. As shown in Figure 2, ResGCN takes the feature matrix of multi-omics data and the patient similarity network as input. After the residual graph convolution operation, outputs the feature matrix of the node. The propagation mechanism of ResGCN can be first represented as Eq. 5:

$$G(N) = f(G(N-1), A) = \sigma(AG(N-1)W(N)) \tag{5}$$

where $G(N)$ is the output of the N-th layer, and $W(N)$ is the weight matrix of N-th layer. $f()$ denotes the graph convolution operation. $\sigma()$ denotes the nonlinear activation function. However, this propagation mechanism only considers the feature vectors of all neighboring nodes, and ignores the nodes themselves. Therefore, self-connection is added to A to overcome this problem, defined as $\hat{A} = A + E$, where E denotes the identity matrix. Moreover, to avoid changes in the scale of the eigenvectors during the multiplication operation, $D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$ is defined to normalize A, where D is the diagonal node degree matrix. Therefore, the propagation mechanism is redefined as Eq. 6.

$$G(N) = f(G(N-1)) = \sigma\left(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}G(N-1)W(N)\right) \tag{6}$$

Theoretically, deeper networks possess more excellent learning capabilities than shallow neural networks to capture feature representations from more complex data (Bianchini and Scarselli, 2014). Furthermore, deeper neural networks are typically able to achieve outstanding performance with relatively less training data. These are particularly significant for multi-omics data, which are often complex and challenged with limited sample sizes (Picard et al., 2021; Zhang et al., 2021). ResGCN uses residual connections to improve the information flow in the network to alleviate the gradient vanishing problem and allow ResGCN to build deep networks (Li et al., 2019). Thus, the new propagation mechanism can be defined as Eq. 7:

$$G(N+1) = f(G(N), A) + G(N) = G(N+1)_{res} + G(N) \tag{7}$$

After $G(N)$ is transformed by f, vertex-wise addition is performed to obtain $G(N+1)$. The residual mapping f learns to take the patient similarity network as input and outputs a residual graph representation $G(N+1)_{res}$ for the next layer. After several layers of residual convolution, the fusion and MLP modules are used to fuse the data processed by multiple residual blocks and output the prediction results. Then, we extract features representing patient similarity information and multi-omics correlation information from the fusion layer in the trained ResGCN. In summary, ResGCN mitigates the gradient vanishing by residual connection mechanism, which improves data transmission in the network and allows for deeper network architecture to fit complex multi-omics data to obtain survival prediction information.

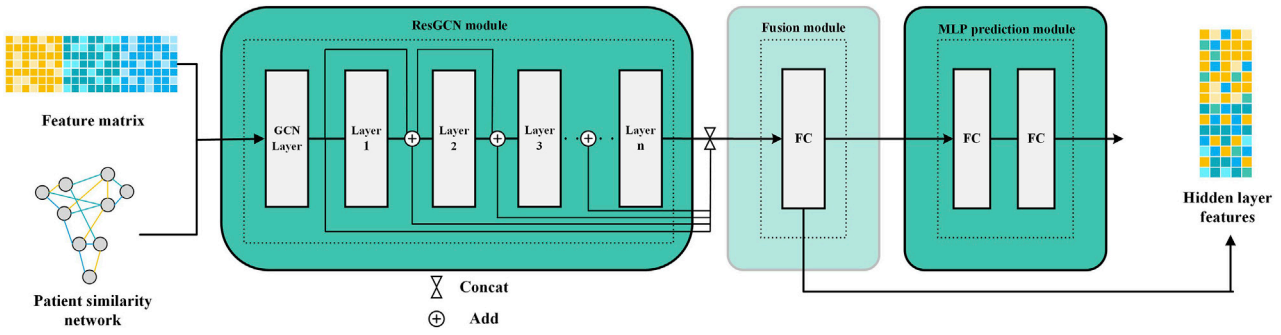

FIGURE 2
Feature extraction process by ResGCN.

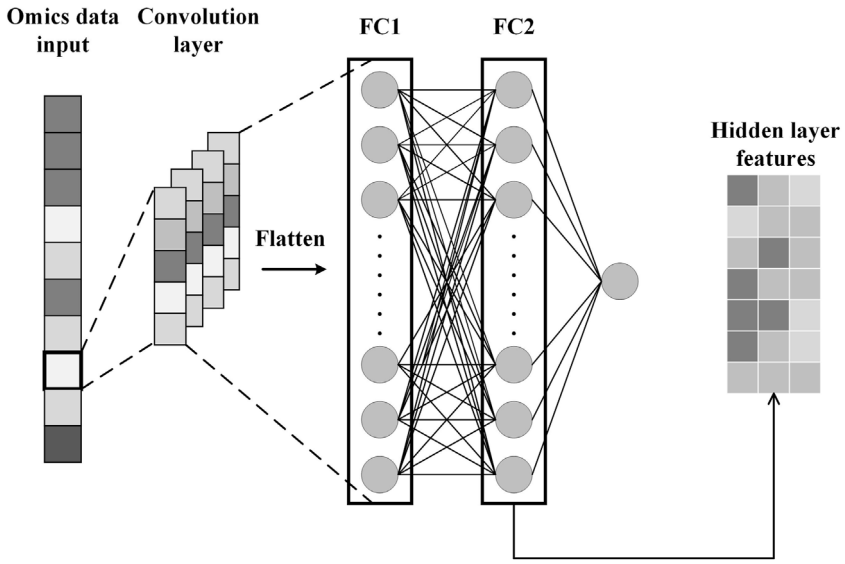

FIGURE 3
Feature extraction process by CNN.

2.2.2 B: Features extraction by CNN

To obtain specificity features for each omics data, we construct CNN for each omics data. Each CNN consists of an input layer, a convolutional layer, a fully connected layer, and an output layer, as shown in Figure 3. After the omics data is fed into the CNN, the convolution layer performs a convolution operation to generate the feature map and adds padding to the convolutional layer to control the feature map size. Subsequently, the flattening operation maps the output of the convolutional layer to a fully connected layer containing 150 units for survival prediction. In addition, the glorot initialization technique is used to generate random numbers to initialize the convolutional kernel (Glorot and Bengio, 2010). We also applied dropout and L2 regularization techniques to prevent overfitting during training (Cortes et al., 2012; Poernomo and Kang, 2018). Finally, we extract specificity features representing each omics data from the fully connected layers of the three trained CNNs.

2.2.3 C: Stack integration and survival prediction

Stacking hidden layer features of deep learning networks is an effective strategy for integrating multi-omics data for survival prediction (Arya and Saha, 2020; Arya and Saha, 2021). It allows flexible integration of feature representations from different neural network models to integrate correlation prognostic and specificity prognostic information. Moreover, this strategy allows integration in conditions that all neural network modules achieve optimal performance, rather than training all modules simultaneously. We stack the hidden layer features extracted from ResGCN and the three CNNs according to Eq. 8.

$$F_{stacked} = F_{PSN} \oplus F_{Clin} \oplus F_{Expr} \oplus F_{CNV} \tag{8}$$

where F_{PSN} represents the feature representation obtained from the ResGCN hidden layer, F_{Clin}, F_{Expr}, and F_{CNV} represent the feature representation obtained from the CNN hidden layer of each omics data, respectively. $F_{stacked}$ represents the obtained stacked feature representation, and $\oplus$ is the matrix concat operation. Then, based on

TABLE 3 Performance comparison of MSFN and comparison methods.

Methods	Accuracy	AUC	Precision	Recall	F1-score	Mcc
LR	0.837	0.788	0.548	0.707	0.615	0.523
RF	0.803	0.736	0.452	0.629	0.521	0.413
SVM	0.821	0.757	0.613	0.633	0.618	0.506
MDNNMD	0.697	0.736	0.313	0.233	0.267	0.128
PregGAN	0.814	0.756	0.617	0.580	0.597	0.477
Stacked RF	0.905	0.956	0.831	0.754	0.790	0.731
SiGaAtCNN RF	0.943	0.981	0.873	0.891	0.882	0.845
Heterogeneous stacked RF	0.891	0.826	0.807	0.695	0.758	0.688
MSFN	**0.978**	**0.991**	**0.932**	**0.964**	**0.944**	**0.930**

Bold values represent the highest performance of the model on this metric in the experiment.

Yifan et al. and Arya et al. we used $F_{stacked}$ to train the AdaBoostRF for the final breast cancer survival prediction (Arya and Saha, 2020; Yifan et al., 2021).

3 Results

3.1 Evaluation metrics and experiment settings

To comprehensively evaluate our model, we use the Area Under the Curve (AUC), accuracy, precision, Recall, F1-score, and Matthew's correlation coefficient (Mcc) as performance evaluation metrics (Goutte and Gaussier, 2005; Huang and Ling, 2005; Chicco and Jurman, 2020). The definitions of these metrics are shown in Eqs 8–14:

$$AUC = \frac{\sum\left(p_i, n_j\right)_{p_i > n_j}}{P \times N} \tag{9}$$

where P is the number of positive samples. N is the number of negative samples. p_i is the positive sample prediction score. n_j is the negative sample prediction score.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

$$Precision = \frac{TP}{TP + FP} \tag{11}$$

$$Recall = \frac{TP}{TP + FN} \tag{12}$$

$$F1 - Score = \frac{2 \times Recall \times Precision}{Recall + Precision} \tag{13}$$

$$Mcc = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN) \times (TP + FP) \times (TN + FN) \times (TN + FP)}} \tag{14}$$

where TP, FP, TN, and FN represent true positives, false positives, true negatives, and false negatives in the confusion matrix, respectively.

To overcome the variance problem caused by the limited sample size and sample imbalance, we used 10-fold cross-validation to evaluate the performance of MSFN (Rodriguez et al., 2009; Jiang and Wang, 2017). The 1048 patients were divided into 10 subsets, 9 of which were combined as the training set while the remaining 1 subset was used as the test set. The final performance was the average of the model's performance on the test set. MSFN was implemented using Pytorch. The experiments were executed on a PC with a 2.90 GHz Intel Core i7-10700 processor and NVIDIA GeForce RTX 3070 GPU.

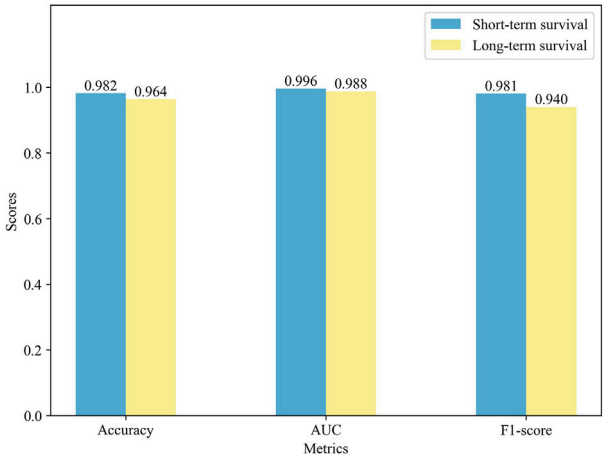

FIGURE 4
Impact of different ResGCN layers on the model performance.

3.2 Comparison with previous studies

To demonstrate the effectiveness of MSFN. We uniformly used 10-fold cross-validation to evaluate and compare it with several machine learning-based methods and deep learning-based methods. Specifically, we selected three widely used machine learning-based models as the baseline: LR (Logistic Regression) (Jefferson et al., 1997), RF (Random Forest) (Nguyen et al., 2013) and SVM (Support Vector Machine) (Xu et al., 2012). Then, we compared MSFN with

TABLE 4 Performance comparison between different variants of MSFN.

	Accuracy	AUC	Precision	Recall	F1-score	Mcc
MSFN/-ResGCN	0.902	0.964	0.869	0.735	0.770	0.726
MSFN/-CNNs	0.960	0.975	0.924	0.932	0.928	0.904
MSFN/-RF	0.965	0.951	0.921	0.909	0.937	0.919
MSFN	**0.978**	**0.991**	**0.932**	**0.964**	**0.944**	**0.930**

Bold values represent the highest performance of the model on this metric in the experiment.

TABLE 5 Performance comparison of different omics data.

Data type	Accuracy	AUC	Precision	Recall	F1-score	Mcc
Clin	0.919	0.954	0.863	0.819	0.834	0.787
CNV	0.765	0.770	0.464	0.438	0.431	0.272
Expr	0.804	0.827	0.658	0.407	0.481	0.371
Multi-omics	**0.978**	**0.991**	**0.932**	**0.964**	**0.944**	**0.930**

Bold values represent the highest performance of the model on this metric in the experiment.

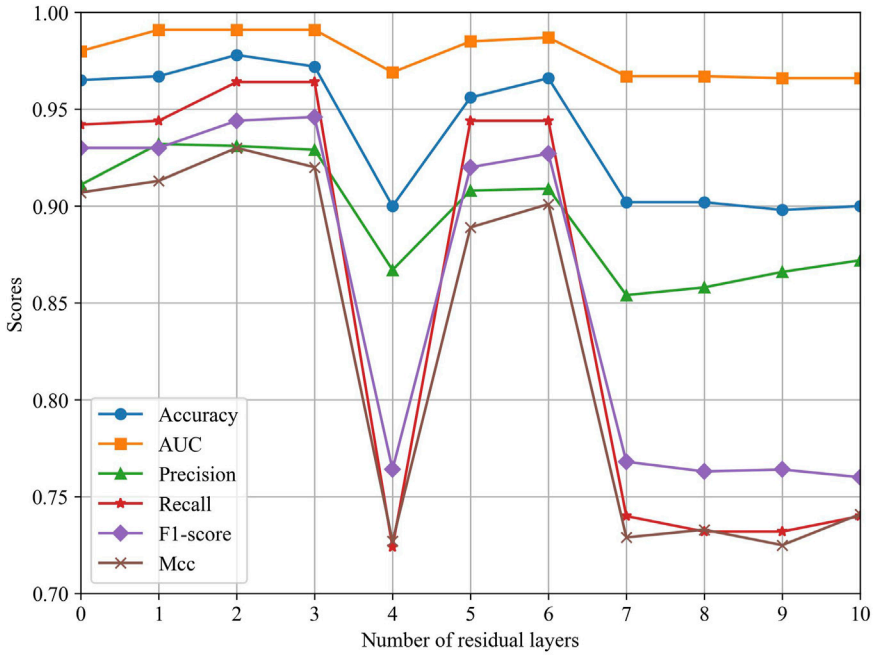

FIGURE 5
Impact of different ResGCN layers on the model performance.

five current state-of-the-art deep learning-based models. Below are brief descriptions of deep learning-based methods:

- MDNNMD (Sun et al., 2018): MDNNMD is a DNN-based cancer survival prediction method. It integrates multi-omics data through multiple DNNs and predicts breast cancer survival by setting different weights for network fusion.
- Stacked RF (Arya and Saha, 2020): Stacked RF is a CNN-based cancer survival prediction method. It trains RF to predict breast cancer survival by stacking three CNN networks' hidden layer feature representations.
- SiGaAtCNN RF (Arya and Saha, 2021): SiGaAtCNN RF is an improved method of Stacked RF. It introduces the gated attention mechanism for better feature representation and stacks hidden layer feature representations of gated

TABLE 6 The results of incremental feature number selection.

CNV	Gene expression	Accuracy	AUC	Precision	Recall	F1-score	Mcc
100	100	0.892	0.970	0.853	0.740	0.760	0.709
100	200	0.968	0.989	0.904	0.908	0.910	0.925
100	300	0.960	0.980	0.927	0.948	0.951	**0.949**
100	400	0.962	0.984	0.906	0.936	0.919	0.894
100	500	0.966	0.985	0.932	0.962	0.950	0.928
200	100	0.969	0.990	0.931	0.936	0.946	0.923
200	200	0.962	0.989	0.931	0.960	0.941	0.929
200	300	0.951	0.983	0.896	0.924	0.908	0.88
200	400	**0.978**	**0.991**	0.932	**0.964**	0.944	0.930
200	500	0.971	0.990	0.932	0.960	0.943	0.930
300	100	0.956	0.979	**0.944**	0.960	0.940	0.934
300	200	0.950	0.988	0.916	0.888	0.899	0.873
300	300	0.960	0.987	0.942	0.956	0.948	0.931
300	400	0.965	0.991	0.937	0.954	**0.951**	0.928
300	500	0.971	0.990	0.929	0.952	0.947	0.930
400	100	0.960	0.987	0.950	0.960	0.948	0.946
400	200	0.949	0.982	0.889	0.920	0.896	0.883
400	300	0.973	0.989	0.940	0.952	0.948	0.933
400	400	0.974	0.991	0.931	0.961	0.939	0.926
400	500	0.904	0.972	0.870	0.744	0.777	0.731
500	100	0.960	0.989	0.935	0.960	0.938	0.946
500	200	0.971	0.991	0.931	0.956	0.934	0.923
500	300	0.912	0.976	0.890	0.740	0.785	0.753
500	400	0.972	0.987	0.935	0.964	0.950	0.918
500	500	0.960	0.992	0.924	0.924	0.914	0.900

Bold values represent the highest performance of the model on this metric in the experiment.

attention CNNs for training RF to predict breast cancer survival.

- PregGAN (Zhang et al., 2022): PregGAN is a CGAN-based survival prediction method. It generates high-quality pseudo-samples based on limited samples for reliable survival prediction.
- Heterogeneous stacked RF (Jadoon et al., 2023): Heterogeneous stacked RF is a heterogeneous ensembled classification prediction model that integrates CNN and DNN to predict breast cancer patient survival.

The prediction results are shown in Table 3. From the results, MSFN achieves AUC value of 0.9787 and accuracy of 0.991, which is superior to other methods. Other evaluation metrics are also obviously improved. Specifically, MSFN achieves superior prediction performance compared to SiGaAtCNN RF, Stacked RF, Heterogeneous stacked RF, and MDNNMD because the patient similarity information and multi-omics data correlation information from the patient similarity network provide more comprehensive and wealthy prognostic information for survival prediction. MSFN achieves significant performance improvement compared to traditional machine learning methods and PregGAN which directly integrate multi-omics data. This demonstrates the effectiveness and superiority of the stacked integration strategy in multi-omics data fusion compared to direct data integration.

3.3 Performance comparison of different survival cohorts

To further validate the prediction performance of MSFN, we compared its performance in different survival cohorts. We used the ten-fold cross-validation for the experiments and displayed the results in Figure 4. It is obvious that MSFN presents a better prediction performance in both long and short survival cohorts, and the gap between the prediction performance of the two cohorts

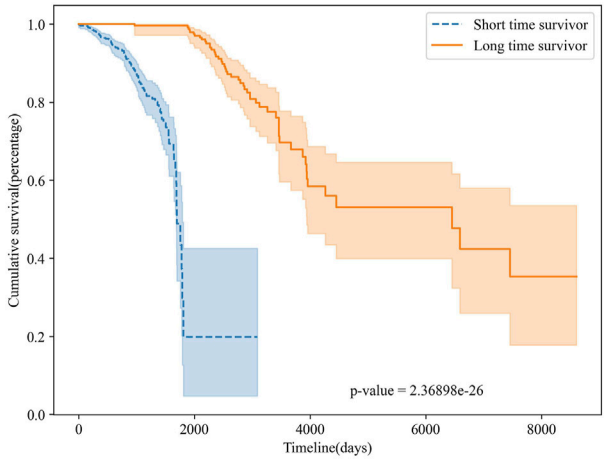

FIGURE 6
The Kaplan-Meier curves for MSFN.

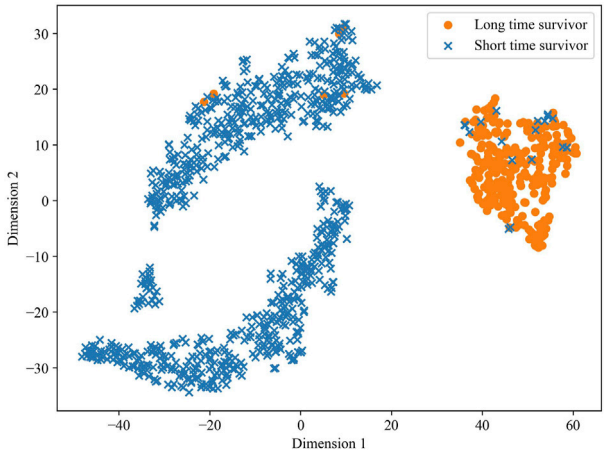

FIGURE 7
The t-SNE plot of stacked feature classification via MSFN.

is very small. This is attributed to that MSFN incorporates the prediction information from different deep learning modules, considering both correlation prognostic information and specificity prognostic information.

3.4 Ablation study

We verify how different modules of MSFN affect the performance through an ablation study and design three variants: i) MSFN/-CNNs: MSFN without CNN modules. ii) MSFN/-ResGCN: MSFN without ResGCN module. iii) MSFN/-RF: MSFN without AdaboostRF, and the prediction results are output by MLP. We compared MSFN with the variants described above. As can be seen from Table 4, both MSFN/-ResGCN and MSFN/-CNN perform lower than MSFN. Such results can be attributed to the incomplete prediction features obtained by MSFN/-ResGCN and MSFN/-CNN. This also reflects the importance of integrating prognostic information. Furthermore, MSFN performs better than MSFN/-RF. This is because the AdaboostRF is an ensemble machine learning algorithm with better feature learning ability than simple MLP and effectively deals with complex feature representations of multi-omics data.

3.5 Effect of multi-omics data

To validate the effect of multi-omics data, we constructed MSFN using each omics data, respectively. Then, we compared them with MSFN constructed using multi-omics data. As shown in Table 5, Clin, CNV, and Expr represent the MSFN constructed with clinical, CNV, or gene expression data, respectively. The accuracy and AUC only reach a maximum of 0.919 and 0.954 when using single-omics data. MSFN achieves the best performance with multi-omics data, with all evaluation metrics significantly better than single-omics data. This indicates that MSFN can obtain comprehensive prognostic information from multi-omics data and significantly improve prediction performance.

3.6 Effect of ResGCN layers

To explore the effect of different ResGCN layers on the model performance, we evaluated the performance of MSFN by changing the layers of ResGCN. As can be seen in Figure 5, the performance of MSFN gradually improves as the number of ResGCN layers increases. Several metrics achieved their maximum when the layer is set to 2. This demonstrates that the deep ResGCN constructed by residual concatenation can properly fit the multi-omics data, bringing performance improvement to the entire model. However, all metrics fluctuate and gradually decrease as the number of layers increases. This may be because ResGCN with too many layers makes the model structure too complex, leading to overfitting of the model during training.

3.7 Effect of the number of features

To explore the effect of the number of features on model performance, we used the incremental method based on previous studies to conduct experiments. Specifically, We employed the mRMR algorithm to select the top 500 features from CNV and gene expression data. Then, we searched with a step size of 100 to evaluate the performance of MSFN under different combinations of feature numbers (Sun et al., 2018; Arya and Saha, 2020). Since only 33 features were available in clinical data, we used all the clinical features. The final results are presented in Table 6. It is evident that the model's performance gradually improves as the number of features increases. MSFN achieves the best accuracy, AUC, and Recall when the number of CNV and gene expression features is set to 200 and 400. However, as the number of features increases, the model's performance remains relatively stable and then gradually decreases. This shows that too many features inevitably introduce noisy information, reducing the model's focus on valuable features and leading to performance degradation. Consequently, we selected the top 200 CNV and top 400 gene expression features along with all 33 clinical features as model inputs.

3.8 Survival analysis

To further validate the survival prediction performance of MSFN, we performed survival analyses on the classification results of MSFN. We plotted Kaplan-Meier curves to evaluate the performance of MSFN in predicting long-term and short-term survivors Rich et al. (2010), illustrated in Figure 6. The Kaplan-Meier survival curves explicitly demonstrated a statistically significant difference (p-value <10e-26) between long-term and short-term survivors predicted by MSFN. This result proves that MSFN effectively distinguishes between long-term and short-term survivors.

To validate the predictive ability of the stacked feature representations obtained in MSFN, we utilized the t-SNE algorithm to visualize the prediction results of the stacked feature representations. t-SNE attempts to minimize the difference between the conditional probabilities or similarities in the high and low dimensional spaces to map the data in the low-dimensional space (Van der Maaten and Hinton, 2008; Wattenberg et al., 2016). The visualization result is shown in Figure 7, a clear demarcation between the two groups at dimension 1 of about 25 indicates the excellent survival prediction ability of the hidden layer features extracted by MSFN.

4 Conclusion

Breast cancer is the most prevalent cancer worldwide and poses a major threat to women's health. Survival prediction can avoid the suffering caused by over-treatment and the waste of medical resources, which is significant for cancer treatment and prognosis. In this study, we propose a novel stacked fusion network (MSFN) for breast cancer survival prediction. MSFN integrates patient similarity, correlation, and specificity information of multi-omics data, providing a more comprehensive insight for survival prediction and effectively enhancing the prediction ability. First, MSFN constructs a patient similarity network and obtains patient similarity information and correlation of multi-omics data through ResGCN. Meanwhile, MSFN obtains the specificity information of multi-omics data through CNN. Finally, MSFN uses the stacking strategy to ingeniously integrate prognostic information and predict patient survival with AdaboostRF. Experiments on TCGA's breast cancer dataset showed that MSFN outperformed state-of-the-art methods in survival prediction. In future work, we will focus on exploring the survival regression issues. Furthermore, we will explore the interpretability of the survival prediction model to understand the decision-making process of the models and the interpretation of the results.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

GZ: Conceptualization, Funding acquisition, Methodology, Supervision, Writing–review and editing, Project administration, Software, Validation, Writing–original draft. CM: Conceptualization, Data curation, Formal Analysis, Methodology, Visualization, Writing–original draft, Writing–review and editing. CY: Conceptualization, Formal Analysis, Methodology, Supervision, Writing–review and editing. HL: Conceptualization, Formal Analysis, Funding acquisition, Methodology, Supervision, Writing–review and editing. JW: Formal Analysis, Funding acquisition, Methodology, Supervision, Writing–review and editing. WL: Writing–review and editing. JL: Writing–review and editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This work was supported by Science and Technology Development Plan Project of Henan Province (No. 222102210238); National Natural Science Foundation of China (No. 62006070); China Postdoctoral Science Foundation (No. 2020M672212).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Arjmand, B., Hamidpour, S. K., Tayanloo-Beik, A., Goodarzi, P., Aghayan, H. R., Adibi, H., et al. (2022). Machine learning: a new prospect in multi-omics data analysis of cancer. *Front. Genet.* 13, 824451. doi:10.3389/fgene.2022.824451

Arnold, M., Morgan, E., Rumgay, H., Mafra, A., Singh, D., Laversanne, M., et al. (2022). Current and future burden of breast cancer: Global statistics for 2020 and 2040. *Breast* 66, 15–23. doi:10.1016/j.breast.2022.08.010

Arya, N., and Saha, S. (2020). Multi-modal classification for human breast cancer prognosis prediction: proposal of deep-learning based stacked ensemble model. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* 19, 1032–1041. doi:10.1109/TCBB.2020.3018467

Arya, N., and Saha, S. (2021). Multi-modal advanced deep learning architectures for breast cancer survival prediction. *Knowledge-Based Syst.* 221, 106965. doi:10.1016/j.knosys.2021.106965

Berisha, V., Krantsevich, C., Hahn, P. R., Hahn, S., Dasarathy, G., Turaga, P., et al. (2021). Digital medicine and the curse of dimensionality. *NPJ Digit. Med.* 4, 153. doi:10.1038/s41746-021-00521-5

Bianchini, M., and Scarselli, F. (2014). On the complexity of neural network classifiers: a comparison between shallow and deep architectures. *IEEE Trans. neural Netw. Learn. Syst.* 25, 1553–1565. doi:10.1109/TNNLS.2013.2293637

Cheerla, A., and Gevaert, O. (2019). Deep learning with multimodal representation for pancancer prognosis prediction. *Bioinformatics* 35, i446–i454. doi:10.1093/bioinformatics/btz342

Chen, H., Gao, M., Zhang, Y., Liang, W., Zou, X., et al. (2019). Attention-based multi-nmf deep neural network with multimodality data for breast cancer prognosis model. *BioMed Res. Int.* 2019, 9523719. doi:10.1155/2019/9523719

Chicco, D., and Jurman, G. (2020). The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics* 21, 6–13. doi:10.1186/s12864-019-6413-7

Ching, T., Zhu, X., and Garmire, L. X. (2018). Cox-nnet: an artificial neural network method for prognosis prediction of high-throughput omics data. *PLoS Comput. Biol.* 14, e1006076. doi:10.1371/journal.pcbi.1006076

Cortes, C., Mohri, M., and Rostamizadeh, A. (2012). *L2 regularization for learning kernels*. arXiv preprint arXiv:1205.2653.

Deepa, P., and Gunavathi, C. (2022). A systematic review on machine learning and deep learning techniques in cancer survival prediction. *Prog. Biophysics Mol. Biol.* 174, 62–71. doi:10.1016/j.pbiomolbio.2022.07.004

García-Laencina, P. J., Abreu, P. H., Abreu, M. H., and Afonoso, N. (2015). Missing data imputation on the 5-year survival prediction of breast cancer patients with unknown discrete values. *Comput. Biol. Med.* 59, 125–133. doi:10.1016/j.compbiomed.2015.02.006

Glorot, X., and Bengio, Y. (2010). "Proceedings of the thirteenth international conference on artificial intelligence and statistics," in *Understanding the difficulty of training deep feedforward neural networks*, 249–256.

Goutte, C., and Gaussier, E. (2005). "A probabilistic interpretation of precision, recall and f-score, with implication for evaluation," in *European conference on information retrieval* (Springer), 345–359. doi:10.1007/978-3-540-31865-1_25

Hagerty, R., Butow, P., Ellis, P., Dimitry, S., and Tattersall, M. (2005). Communicating prognosis in cancer care: a systematic review of the literature. *Ann. Oncol.* 16, 1005–1053. doi:10.1093/annonc/mdi211

Herrmann, M., Probst, P., Hornung, R., Jurinovic, V., and Boulesteix, A.-L. (2021). Large-scale benchmark study of survival prediction methods using multi-omics data. *Briefings Bioinforma.* 22, bbaa167. doi:10.1093/bib/bbaa167

Huang, J., and Ling, C. X. (2005). Using auc and accuracy in evaluating learning algorithms. *IEEE Trans. Knowl. Data Eng.* 17, 299–310. doi:10.1109/TKDE.2005.50

Jadoon, E. K., Khan, F. G., Shah, S., Khan, A., and Elaffendi, M. (2023). Deep learning-based multi-modal ensemble classification approach for human breast cancer prognosis. *IEEE Access* 11, 85760–85769. doi:10.1109/access.2023.3304242

Jefferson, M. F., Pendleton, N., Lucas, S. B., and Horan, M. A. (1997). Comparison of a genetic algorithm neural network with logistic regression for predicting outcome after surgery for patients with nonsmall cell lung carcinoma. *Cancer Interdiscip. Int. J. Am. Cancer Soc.* 79, 1338–1342. doi:10.1002/(sici)1097-0142(19970401)79:7⟨1338::aid-cncr10⟩3.0.co;2-0

Jiang, G., and Wang, W. (2017). Error estimation based on variance analysis of k-fold cross-validation. *Pattern Recognit.* 69, 94–106. doi:10.1016/j.patcog.2017.03.025

Kalafi, E., Nor, N., Taib, N., Ganggayah, M., Town, C., and Dhillon, S. (2019). Machine learning and deep learning approaches in breast cancer survival prediction using clinical data. *Folia Biol.* 65, 212–220. doi:10.14712/fb2019065050212

Kang, M., Ko, E., and Mersha, T. B. (2022). A roadmap for multi-omics data integration using deep learning. *Briefings Bioinforma.* 23, bbab454. doi:10.1093/bib/bbab454

Katzman, J. L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., and Kluger, Y. (2018). Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC Med. Res. Methodol.* 18, 24–12. doi:10.1186/s12874-018-0482-1

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444. doi:10.1038/nature14539

Li, G., Müller, M., Thabet, A., and Ghanem, B. (2019). "Deepgcns: can gcns go as deep as cnns?," in *2019 IEEE/CVF international conference on computer vision (ICCV)*, 9266–9275. doi:10.1109/ICCV.2019.00936

Li, R., Wu, X., Li, A., and Wang, M. (2022). Hfbsurv: hierarchical multimodal fusion with factorized bilinear models for cancer survival prediction. *Bioinformatics* 38, 2587–2594. doi:10.1093/bioinformatics/btac113

Li, Y., Kang, K., Krahn, J. M., Croutwater, N., Lee, K., Umbach, D. M., et al. (2017). A comprehensive genomic pan-cancer classification using the cancer genome atlas gene expression data. *BMC genomics* 18, 508–513. doi:10.1186/s12864-017-3906-0

Lujambio, A., and Lowe, S. W. (2012). The microcosmos of cancer. *Nature* 482, 347–355. doi:10.1038/nature10888

Michaelson, J. S., Silverstein, M., Wyatt, J., Weber, G., Moore, R., Halpern, E., et al. (2002). Predicting the survival of patients with breast carcinoma using tumor size. *Cancer Interdiscip. Int. J. Am. Cancer Soc.* 95, 713–723. doi:10.1002/cncr.10742

Nguyen, C., Wang, Y., and Nguyen, H. N. (2013). Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic. *J. Biomed. Sci. Eng.* 06, 551–560. doi:10.4236/jbise.2013.65070

Patro, S., and Sahu, K. K. (2015). *Normalization: a preprocessing stage*. arXiv preprint arXiv:1503.06462.

Peng, H., Long, F., and Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. pattern analysis Mach. Intell.* 27, 1226–1238. doi:10.1109/TPAMI.2005.159

Picard, M., Scott-Boyer, M.-P., Bodein, A., Périn, O., and Droit, A. (2021). Integration strategies of multi-omics data for machine learning analysis. *Comput. Struct. Biotechnol. J.* 19, 3735–3746. doi:10.1016/j.csbj.2021.06.030

Poernomo, A., and Kang, D.-K. (2018). Biased dropout and crossmap dropout: learning towards effective dropout regularization in convolutional neural network. *Neural Netw.* 104, 60–67. doi:10.1016/j.neunet.2018.03.016

Rich, J. T., Neely, J. G., Paniello, R. C., Voelker, C. C., Nussenbaum, B., and Wang, E. W. (2010). A practical guide to understanding kaplan-meier curves. *Otolaryngology—Head Neck Surg.* 143, 331–336. doi:10.1016/j.otohns.2010.05.007

Rodriguez, J. D., Perez, A., and Lozano, J. A. (2009). Sensitivity analysis of kappa-fold cross validation in prediction error estimation. *IEEE Trans. pattern analysis Mach. Intell.* 32, 569–575. doi:10.1109/TPAMI.2009.187

Shlien, A., and Malkin, D. (2009). Copy number variations and cancer. *Genome Med.* 1, 62–69. doi:10.1186/gm62

Sun, D., Wang, M., and Li, A. (2018). A multimodal deep neural network for human breast cancer prognosis prediction by integrating multi-dimensional data. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* 16, 841–850. doi:10.1109/TCBB.2018.2806438

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., et al. (2021). Global cancer statistics 2020: globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA a cancer J. Clin.* 71, 209–249. doi:10.3322/caac.21660

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., et al. (2001). Missing value estimation methods for dna microarrays. *Bioinformatics* 17, 520–525. doi:10.1093/bioinformatics/17.6.520

Van der Maaten, L., and Hinton, G. (2008). Visualizing data using t-sne. *J. Mach. Learn. Res.* 9.

Wang, B., Mezlini, A. M., Demir, F., Fiume, M., Tu, Z., Brudno, M., et al. (2014). Similarity network fusion for aggregating data types on a genomic scale. *Nat. methods* 11, 333–337. doi:10.1038/nmeth.2810

Wattenberg, M., Viégas, F., and Johnson, I. (2016). How to use t-sne effectively. *Distill* 1, e2. doi:10.23915/distill.00002

Xu, X., Zhang, Y., Zou, L., Wang, M., and Li, A. (2012). "A gene signature for breast cancer prognosis using support vector machine," in *2012 5th international conference on BioMedical engineering and informatics* (IEEE). doi:10.1109/BMEI.2012.6513032

Yifan, D., Jialin, L., and Boxi, F. (2021). "Forecast model of breast cancer diagnosis based on rf-adaboost," in *2021 international conference on communications, information system and computer engineering (CISCE)* (IEEE), 716–719. doi:10.1109/CISCE52179.2021.9445847

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. (2021). Understanding deep learning (still) requires rethinking generalization. *Commun. ACM* 64, 107–115. doi:10.1145/3446776

Zhang, F., Zhang, Y., Zhu, X., Chen, X., Du, H., and Zhang, X. (2022). Preggan: a prognosis prediction model for breast cancer based on conditional generative adversarial networks. *Comput. Methods Programs Biomed.* 224, 107026. doi:10.1016/j.cmpb.2022.107026