SIN-CHN SCIENTIFIC PRESS PTE.LTD.

Molecular & Cellular Biomechanics 2024, 21(4), 785.
https://doi.org/10.62617/mcb785

Article

AET-net: A framework for subtype classification based on the multi-omics data of breast cancer

Qiaosheng Zhang[1,2,*], Yalong Wei[1], Jie Hou[3], Junjie Xu[1], Zhenyu Sun[1], Heng Zhang[1], Zhaoman Zhong[1]

[1] School of Computer Engineering, Jiangsu Ocean University, Lianyungang 222000, China
[2] Jiangsu Institute of Marine Resources Development, Jiangsu Ocean University, Lianyungang 222000, China
[3] Public Teaching and Research Department, Huzhou College, Huzhou 313000, China
*** Corresponding author:** Qiaosheng Zhang, zqs@jou.edu.cn

CITATION

Zhang Q, Wei Y, Hou J, et al. AET-net: A framework for subtype classification based on the multi-omics data of breast cancer. Molecular & Cellular Biomechanics. 2024; 21(4): 785. https://doi.org/10.62617/mcb785

ARTICLE INFO

Received: 13 November 2024
Accepted: 22 November 2024
Available online: 12 December 2024

Abstract: Breast cancer (BC) is one of the most prevalent cancers worldwide and remains a significant global public health challenge. The biomechanical characteristics of tumor microenvironments provide critical insights into cellular interactions and mechanical stress responses that potentially influence cancer progression. The integration and analysis of multi-omics data for BC subtype classification present substantial challenges, including high-dimensional data complexity and difficulties in integrating heterogeneous omics data characteristics. To address these challenges, we propose an Autoencoder and Transformer integrated neural network (AET-net) classification framework. The experimental results demonstrate that our model achieves significant performance improvements in predicting BC subtypes based on integrated multi-omics datasets, with an Accuracy of 0.912 and an AUC of 0.9862. These results not only validate the high accuracy of our model in BC subtype classification, providing a valuable tool for diagnostic decision support, but also demonstrate the potential of integrated multi-omics data analysis in enhancing the precision and efficiency of BC subtype identification.

Keywords: multi-omics data; deep learning; breast cancer; autoencoder; transformer

1. Introduction

BC remains a significant global public health challenge and is currently one of the most prevalent cancers worldwide, characterized as a multifactorial disease with diverse causes [1–3]. Recent biomechanical studies have revealed that the mechanical properties of BC tissues, including cellular stiffness, extracellular matrix interactions, and tumor microenvironment mechanics, play a critical role in cancer progression, metastasis, and potential diagnostic approaches [4,5]. Furthermore, preliminary studies suggest that mechanical alterations in breast tissue can serve as potential indicators of malignant transformation, providing insights into the pathophysiological changes preceding overt tumor development [6,7]. Accurate classification of BC subtypes is crucial for understanding disease mechanisms and guiding treatment decisions, as different subtypes exhibit distinct molecular characteristics and clinical behaviors [8,9]. With the advancement of gene expression profiling technologies, BC can be categorized into different molecular subtypes (such as luminal A, luminal B, HER2 over-expression, basal-like, and normal-like), and the introduction of the prediction analysis of microarray 50 (PAM50) model has further standardized this classification approach to support individualized treatment decisions [10]. The advancement of high-throughput technologies has enabled the generation of multiple types of omics data, particularly gene expression and DNA

methylation data, which provide complementary molecular insights into BC biology and classification [11,12]. The integration of these multi-omics data types offers promising opportunities for more accurate and comprehensive BC subtype classification, where gene expression data captures the dynamic transcriptional state of cancer cells, while DNA methylation profiles reveal important epigenetic regulatory patterns [13,14]. However, the effective utilization of these data types faces two major challenges: Both gene expression and DNA methylation data are inherently high-dimensional, containing thousands of features that characterize different aspects of cancer biology, and these different omics data types possess distinct characteristics, making their integration particularly challenging for achieving accurate subtype classification. The successful integration of these multi-omics data could potentially enhance our ability to accurately classify BC subtypes, providing a valuable tool for diagnostic decision support.

In recent subtype classification research, various frameworks for molecular subtype classification of BC have emerged. For instance, Choi and Chae [15] proposed moBRCA-net, a deep learning framework for BC subtype classification that integrates multiple omics data types. moBRCA-net demonstrated superior performance compared to established machine learning methods and other state-of-the-art cancer subtype classifiers, highlighting the benefits of multi-omics data integration and attention mechanisms in improving classification accuracy. Zubair et al. [16] discussed the increasing global burden of BC and the limitations of current diagnostic methods. They highlighted the need for advanced molecular diagnostic tools and explores the potential of biomarkers, multigene assays, and portable biosensors for early detection and personalized treatment of BC, while also mentioning ongoing clinical studies aimed at improving patient outcomes. Moreover, Gao et al. [17] designed DeepCC, a computational algorithm for calculating enrichment scores of each cancer sample's gene expression profile based on selected gene sets. They employed these scores to implement a fully-connected neural network model for classifying BC subtypes. Meti et al. [18] compared the predictive performance of machine learning prediction models with standard statistical models on clinical and pathological data to assist in the early identification of BC patients who respond poorly to neoadjuvant chemotherapy. Graudenzi et al. [19] proposed a novel cancer subtype classifier based on gene expression data and applied it to two different BC datasets. This classifier, based on a support vector machine, relies on critical pathway information related to BC development to reduce the vast variable space. However, these methods may have some potential issues, including dataset imbalance and classification accuracy after data integration. For example, certain BC subtypes may be more common than others, leading to an uneven distribution of samples in the dataset. This imbalance could impact the training of classification models as the models might overfit to more common categories and neglect less common ones, potentially resulting in poorer classification performance for less common BC subtypes [20,21]. Machine learning classification methods have certain shortcomings in feature engineering, dependency on input data, handling long-term dependency relationships, parameter tuning, and large-scale data processing. In contrast, the Transformer model, with features such as self-attention mechanism and

position encoding, can handle sequence data and large-scale data more effectively, without the need for manual feature engineering [22].

In this study, we introduce a novel framework for cancer subtype classification, specifically tailored for BC, which we refer to as AET-net. Our approach leverages multi-omics data, encompassing both gene expression and DNA methylation profiles. The data integration process employs an Autoencoder neural network, which effectively consolidates the disparate omics layers into a unified representation. Building upon this integrated data, we develop a Transformer-based classification framework that incorporates an attention mechanism to enhance the accuracy of BC subtype classification. To rigorously assess the efficacy of our proposed method, we perform a comparative analysis against several conventional machine learning classifiers under identical data conditions and dataset partitioning. Our experimental results demonstrate that the AET-net framework not only integrates multi-omics data with high fidelity but also achieves superior classification performance compared to other tested models. Metrics such as classification accuracy, precision, recall, and F1-score consistently highlight the advantages of our approach. The implications of our framework extend beyond theoretical advancements, offering significant potential for diagnostic decision support in BC subtype classification. By leveraging the integrated multi-omics data and superior classification performance, AET-net provides a more nuanced and accurate categorization of BC subtypes.

2. Materials and methods

2.1. Datasets collection and cleaning

In this retrospective study, we utilized the multi-omics dataset from The Cancer Genome Atlas Breast (TCGA) Invasive Carcinoma (BRCA) project, which is a comprehensive BC study under TCGA platform [23]. To conduct our analysis, we specifically obtain the multi-omics data related to gene expression and DNA methylation. The gene expression dataset consists of RNA sequencing data from 1097 BC samples, encompassing 20,531 genes. This dataset provides a comprehensive transcriptomic profile, allowing us to analyze gene expression patterns across different BC subtypes. For the collection of these datasets, we employ the TCGAbiolinks and SummarizedExperiment (SE) libraries within the R programming environment. TCGAbiolinks is a powerful R package that facilitates the integrative analysis of TCGA data, providing functions for data retrieval, preprocessing, and downstream analysis. The SE package in R allows for the efficient storage and manipulation of large, multi-omics datasets, providing a flexible structure to handle assay data and associated metadata. By leveraging these tools, we ensure that the data is consistently processed and normalized, maintaining the integrity and reproducibility of our analyses [24]. Our research leverages these datasets to investigate the molecular subtypes of BC. By integrating gene expression and DNA methylation data, we aim to uncover novel insights into the epigenetic regulation of gene expression in BC and its implications for patient prognosis and treatment strategies.

It is essential to clean the gene expression data to ensure the quality and reliability of. the subsequent analyses. The primary reason for this data cleaning is that many genes in the dataset cannot be read accurately. This can happen due to various reasons such as technical limitations of the sequencing methods, low expression levels, or errors during data collection. The presence of unreadable genes introduces a significant amount of noise, which can adversely affect the learning process of the deep learning model. Noise in the data can lead to overfitting, reduce the model's ability to generalize, and ultimately result in poor predictive performance. To address this issue, we implement a data cleaning step where we set a specific threshold for gene expression readings. If a gene is not read in the majority of samples, it is considered unreliable and thus removed from the dataset. This threshold is carefully chosen to balance the need to retain as much useful information as possible while eliminating the noise caused by unreadable genes. For instance, if more than 80% of the samples have a zero reading for a particular gene, that gene is excluded from further analysis. This thresholding helps in reducing the dimensionality of the dataset, thereby simplifying the model training process and improving the overall quality of the data. In addition to data cleaning, it is also crucial to categorize the patient samples based on their specific subtypes. For our dataset of patient samples, we first applied the Synthetic Minority Over-sampling Technique (SMOTE) to address class imbalance and ensure representative sampling across different BC subtypes. Subsequently, we performed a subtype analysis using the PAM50 algorithm, which is a widely recognized method for BC subtype classification [25]. **Table 1** shows the analysis of BC subtype information for the original 780 patient samples in the dataset. After applying the SMOTE, the number of samples in each subtype is balanced to 422 samples per subtype.

Table 1. Number of original samples in each subtype.

Data Source	Subtype	Number of Samples
TCGA-BRCA	Basal-like	137
	HER2	46
	LumA	422
	LumB	141
	Normal-like	34

2.2. Standardization of multi-omics data

We employ the z-score algorithm to standardize the data. Standardization is a crucial step, especially in the context of multi-omics data, where different types of data can have varying dimensions and magnitudes. Multi-omics data integration involves datasets such as gene expression and DNA methylation [26]. Without standardization, the differences in scales can lead to biased results, where variables with larger magnitudes dominate the analysis. The z-score normalization method transforms the data to have a mean of zero and a standard deviation of one, effectively removing the influence of scale and allowing different types of data to be compared on the same level. This process ensures that each type of omics data

contributes equally to the analysis, facilitating more accurate and meaningful integrative analysis. The formula used for z-score standardization is as follows:

$$Z = \frac{X - \mu}{\sigma} \tag{1}$$

In this formula, X represents the value from the multi-omics dataset for BC. The term μ denotes the mean of the dataset, calculated as the average value of all data points:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} X_i \tag{2}$$

where N is the total number of data points, and X_i represents each individual data. point. The standard deviation σ measures the dispersion of the data points around the mean and is calculated using the formula:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (X_i - \mu)^2} \tag{3}$$

Finally, Z represents the standardized dataset, where each data point is transformed. as follows:

$$Z_i = \frac{X_i - \mu}{\sigma} \tag{4}$$

By applying this transformation, each variable in the multi-omics dataset is rescaled. to have a mean of zero and a standard deviation of one. This rescaling is particularly important for algorithms sensitive to data scales, such as principal component analysis and many machine learning techniques, including deep learning models [27]. Standardizing the dataset helps in enhancing the model's performance by ensuring that each feature contributes equally to the learning process. It mitigates the risk of model bias towards features with higher magnitudes and improves the convergence speed during the training phase of the model.

2.3. Multi-omics data integration

The Autoencoder is an unsupervised deep learning model designed to learn low-dimensional representations of data. This model is particularly useful for integrating multi-omics data, such as gene expression and DNA methylation data, which often have high dimensionality and complex structures. By learning compact representations of these datasets, the Autoencoder facilitates more efficient and meaningful data integration [28].

We utilize an Autoencoder to integrate gene expression data and DNA methylation data, as depicted in **Figure 1**.

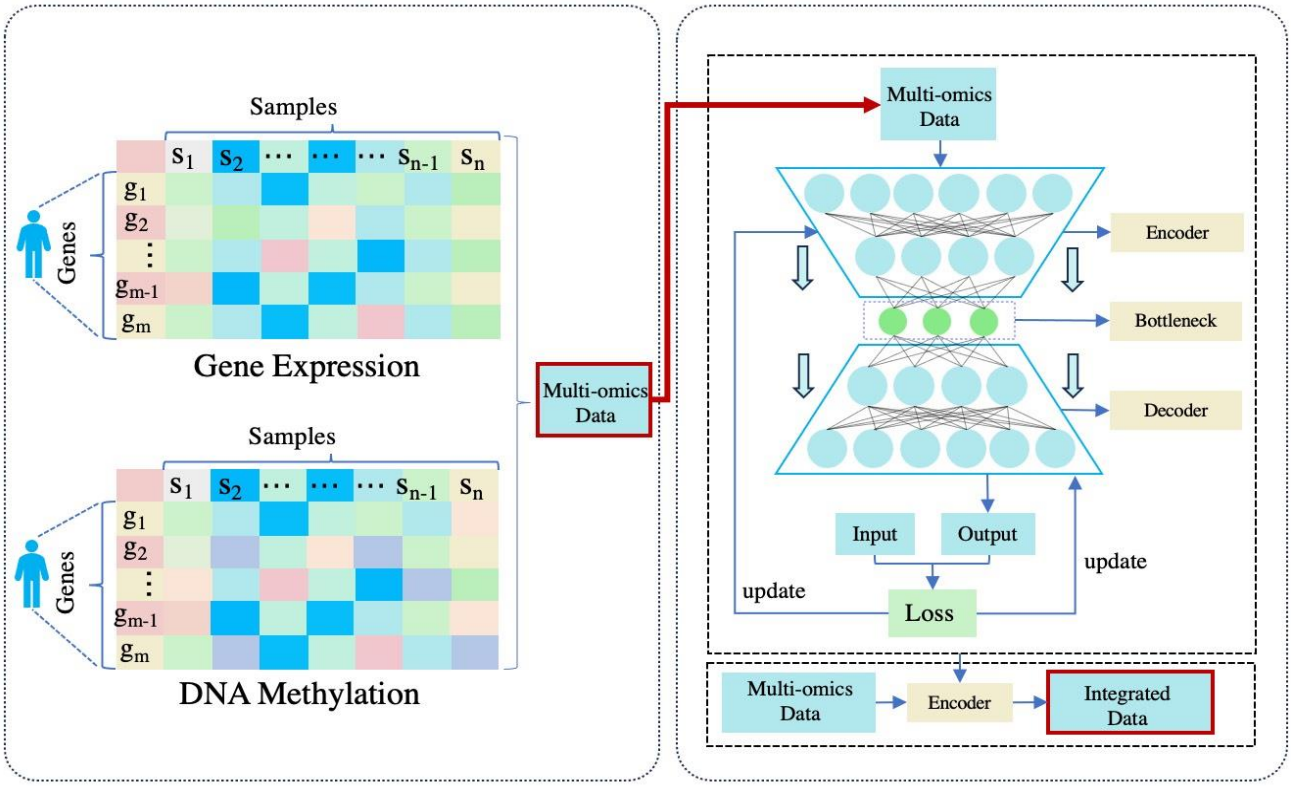

Figure 1. Multi-omics data integration framework.

The architecture of our data integration framework is designed to leverage the complementary information from these two types of omics data, thereby enhancing the downstream analysis and predictive modeling. The Autoencoder model comprises two main components: The encoder (ϕ) and the decoder (ψ). The encoder's role is to compress the original high-dimensional data into a lower-dimensional representation, effectively capturing the essential features of the data while discarding noise and redundant information. Mathematically, the encoder function can be represented as:

$$Z = \phi(X) \tag{5}$$

where X represents the input data (gene expression and DNA methylation data) and Z is the resulting low-dimensional representation. The decoder, on the other hand, aims to reconstruct the original data from this low-dimensional representation. The reconstruction process is designed to be as accurate as possible, ensuring that the learned representation retains the critical information necessary for data integration. The decoder function is expressed as:

$$\hat{X} = \psi(Z) \tag{6}$$

The training objective of the Autoencoder is to minimize the reconstruction loss, which measures the difference between the original data X and the reconstructed data $\hat{X}$. This objective can be formulated as:

$$\phi, \psi = \arg \min_{\phi,\psi} L(X, (\psi \circ \phi)X) \tag{7}$$

Here, L denotes the loss function, which is typically the mean squared error between X and $\hat{X}$:

$$L(X, \hat{X}) = \frac{1}{N}\sum_{i=1}^{N} (X_i - \hat{X}_i)^2 \quad (8)$$

By minimizing this loss, the Autoencoder learns to capture the most significant features of the input data in a lower-dimensional space, facilitating effective data integration. The integrated multi-omics data, represented in a compact and informative manner, serves as a robust input for the subsequent Transformer classification framework. The reduced dimensionality not only makes the data easier to handle and train but also helps in reducing the number of parameters in the Transformer model, thereby improving its training efficiency and generalization capability. The use of an Autoencoder for multi-omics data integration enables us to combine gene expression and DNA methylation data into a cohesive and informative representation. This integrated data enhances the performance of machine learning models by providing a comprehensive view of the biological systems under study, ultimately contributing to more accurate and insightful analyses in BC research.

2.4. Attentional neural network multi-omics classification framework

The Transformer is a neural network model based on self-attention mechanisms, widely used in natural language processing tasks such as machine translation and text classification [29]. It possesses powerful modeling capabilities, able to capture long-distance dependencies and contextual information in input data. For the multi-omics BC classification task, the Transformer effectively learns the interactions and associations between different omics data, thereby enhancing classification performance.

Multi-omics BC classification typically involves multiple data modalities, such as gene expression data, DNA methylation data, and proteomics data. The characteristics and expression methods of data in each modality differ significantly, posing a challenge for traditional machine learning classification methods which often struggle to fully utilize the relationships between multi-modal data. The Transformer model, however, adeptly handles multiple input modalities simultaneously and learns the interactions and importance between modalities through its self-attention mechanism. This capability allows for better integration of information from different modalities, improving classification performance and providing deeper insights into the underlying biological processes.

As shown in **Figure 2**, we employ an Autoencoder and Transformer to integrate multi-omics data.

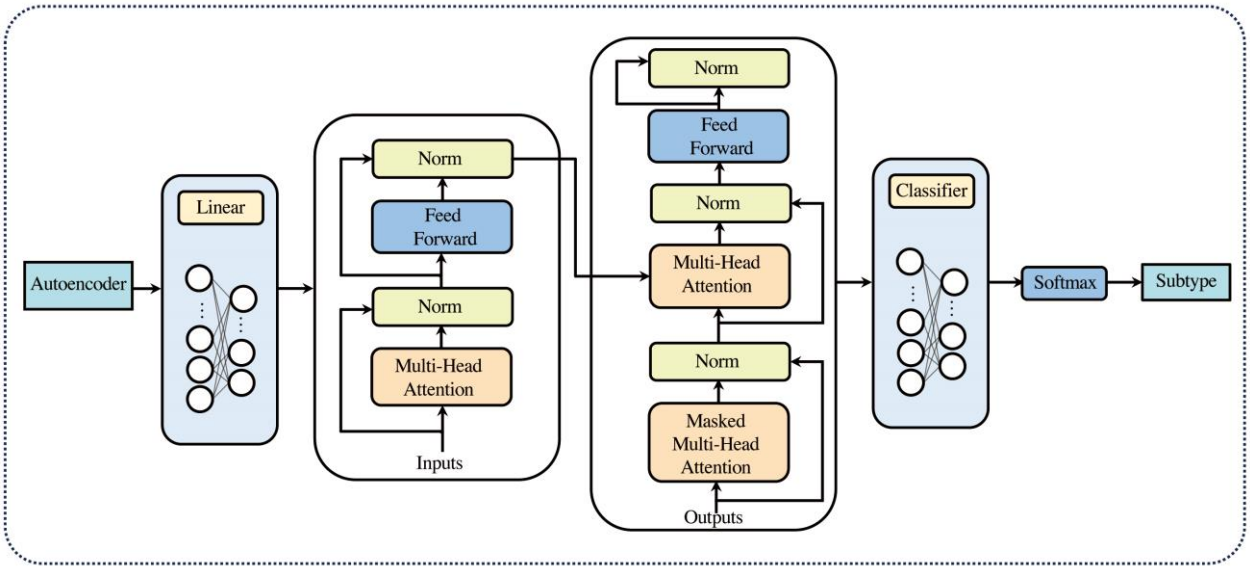

Figure 2. AET-net: Attentional neural network classification framework incorporating autoencoder.

The Autoencoder effectively compresses the high-dimensional data into a lower-dimensional latent space, preserving the most relevant features from each modality. The integrated data is then input into the Transformer neural network, which processes the data through its layers of self-attention and feed-forward neural networks. We add a Classifier layer to the output of the Transformer, utilizing the softmax function to determine the subtype categories. The softmax function converts the output scores into probabilities, facilitating the classification of BC subtypes. The softmax function used is as follows:

$$\text{Softmax}(x_i) = \frac{\exp(x_i)}{\sum_j \exp(x_j)} \tag{9}$$

To optimize the model, we use CrossEntropyLoss as the loss function. This function measures the performance of the classification model whose output is a probability value between 0 and 1. The function used is defined as:

$$\ell(x, y) = L = \{l_1, \ldots, l_N\}^\top \tag{10}$$

$$l_n = -\sum_{c=1}^{C} w_c \log \frac{\exp(x_{n,c})}{\sum_{i=1}^{C} \exp(x_{n,i})} y_{n,c} \tag{11}$$

where x represents the input data, y represents the target labels, w denotes the weight assigned to each class, and C is the number of subtype classes. This loss function penalizes the model proportionally to the error in probability estimation, thus guiding the model to improve its predictions over iterations.

The integration of the Autoencoder with the Transformer neural network and the application of the softmax function and CrossEntropyLoss ensure that our model efficiently captures the intricate relationships within the multi-omics data, leading to superior classification performance. This method demonstrates a significant advancement in the field of multi-omics data analysis, providing a robust framework

for BC subtype classification and potentially aiding in the development of personalized treatment strategies.

2.5. Comparison of machine learning classification models

We employ a suite of classical machine learning classification algorithms to evaluate the performance of our proposed AET-net data integration model. Specifically, we utilize Light Gradient Boosting Machine (LightGBM), Logistic Regression (LR), Random Forest Classifier (RF), and Extra Trees Classifier (ET). These algorithms are chosen due to their proven effectiveness in various classification tasks as highlighted by Kotsiantis et al. [30].

The LightGBM, an efficient and powerful gradient boosting framework, is particularly noted for its speed and accuracy in handling large data sets. Logistic Regression, a fundamental statistical approach, serves as a baseline for its interpretability and simplicity. Random Forest and Extra Trees, both ensemble learning methods based on decision trees, are included for their robustness to overfitting and capacity to model complex interactions within data. We conduct our experiments using the same dataset, which is split into consistent training and testing sets across all models to ensure fair comparison. The training phase involves careful hyperparameter tuning and cross-validation to optimize each model's performance. During testing, we evaluate the models based on standard metrics such as accuracy, precision, recall, and the F1-score to assess and compare their classification capabilities. Our results aim to demonstrate how our AET-net model, which integrates multiple data types using an advanced neural network architecture, compares against these traditional classifiers in terms of both performance and computational efficiency. We discuss the implications of these findings and how they validate the effectiveness of AET-net for complex data integration tasks in machine learning environments.

2.6. Evaluation metrics

We utilize the following commonly used classification metrics to rigorously evaluate the performance of the classifiers.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{12}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{13}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{14}$$

$$\text{F1-Score} = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{15}$$

These metrics are vital for assessing each model's efficacy in tackling the multi-omics BC subtype classification task.

3. Results

3.1. Analysis of integration dimensions for autoencoder

The Autoencoder excels in reducing data dimensionality while preserving the complex and nonlinear relationships inherent in diverse biological data types. We select integration dimensions of 2048, 1024, 512, 256, and 128 using the Autoencoder model. This choice is made to explore their impact on the overall performance of the system. Specifically, these dimensions are chosen to strike a balance between computational efficiency and the depth of biological insights they can capture. As depicted in **Figure 3**, the loss for each dimension is critically evaluated. The results indicate a trend where increasing the dimension size initially decreases the loss rate. However, beyond 2048 dimensions, the reduction in loss plateaus, suggesting a diminishing return on model complexity. The analysis suggests that a dimensionality setting around 2048 offers an optimal balance between loss reduction and computational efficiency.

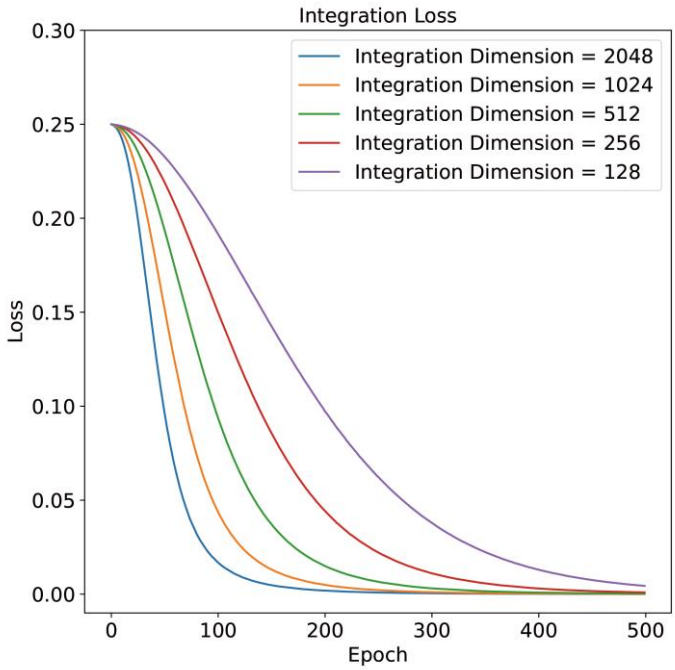

Figure 3. Autoencoder multi-omics dataset integrated loss.

3.2. Analysis of AET-net parameter settings

We adjust the hyperparameters of the model to maximize its performance. For efficient memory utilization, the input size for the batch size parameter is set to 128. This setting balances the need for sufficient data per batch with the constraints of the available memory, allowing for optimal processing efficiency. The training process comprises 5000 epoches, providing the model with ample opportunity to learn and converge. This extensive training duration ensures that the model has enough iterations to achieve maximum performance and stability in its predictions. Key parameters within the model's architecture are also finely tuned. The transformer architecture is configured with nhead value of 8, indicating the number of attention

heads used in the multi-head attention mechanism. This setting enhances the model's ability to focus on different parts of the input sequence simultaneously, improving its overall interpretative power. The number of layers, nlayers, is set to 6. This depth allows the model to capture complex patterns and relationships within the data, contributing to its high performance. To prevent overfitting and enhance generalization, a dropout rate of 0.1 is applied. This technique randomly omits a fraction of the neurons during training, which helps in making the model robust to unseen data. Additionally, the encoder dimension is set to 1024. This dimension size defines the internal representation space of the model, providing a rich and detailed encoding of the input data. The combination of these hyperparameter settings results in a powerful and efficient model, capable of high performance across various tasks.

3.3. Results of multi-omics subtype classification

In **Figure 4**, we observe that as the number of training epochs increases, the training loss of our multi-omics integration classification network model gradually decreases. This indicates that the model is progressively optimized during the learning process, with the degree of fit to the training data continually improving. When the epoch reaches 5000, the model appears to have reached a stable state, which may suggest that the model has found the optimal solution on the current training data, i.e., the model has converged. Convergence refers to the state in the training process where the update of parameters tends to be stable, and the performance of the model no longer significantly improves.

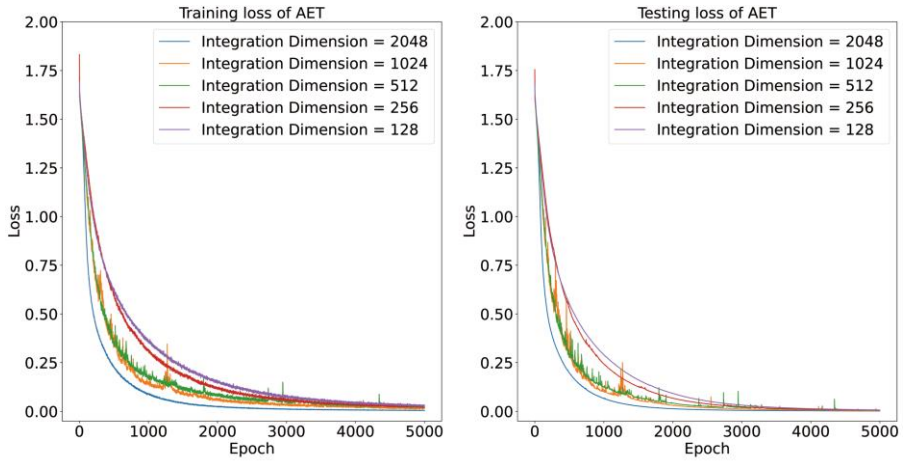

Figure 4. Training and testing losses with different integration dimensions of AET.

The **Figure 5** presents performance metrics for a classification model across different BC subtypes. Overall, the model shows excellent performance, with high AUC values (> 0.95) for all subtypes, indicating strong discriminative ability. Precision and F1 scores are generally high, but vary somewhat across subtypes. The model performs exceptionally well on Basal-like, HER2, and Normal-like subtypes, with near-perfect AUC and very high precision and F1 scores. Performance on Luminal A and B subtypes is still strong but slightly lower, particularly in terms of F1 score for Luminal A and precision for Luminal B.

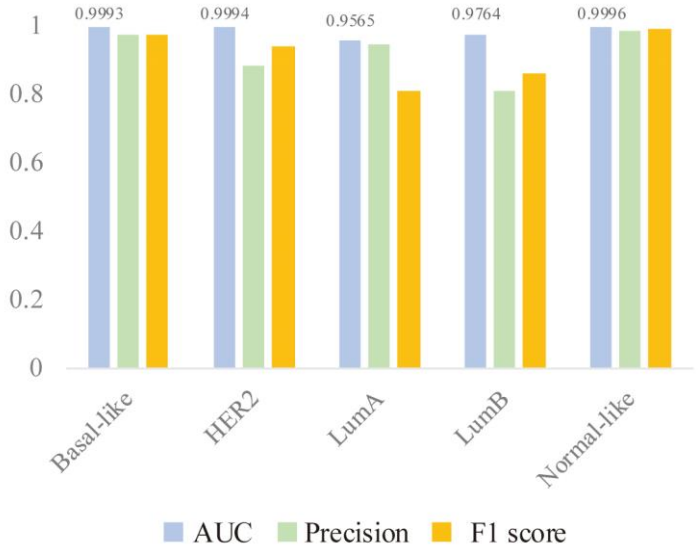

Figure 5. Performance metrics for BC subtype classification.

Furthermore, our analysis reveals significant improvements in model performance when comparing single-omic and multi-omics approaches. As shown in **Figure 6A**, which compares gene expression data alone to the integrated multi-omics model, we observe substantial enhancements across all metrics. For instance, accuracy increased from 0.6923 to 0.9120, while the F1-score improved dramatically from 0.5339 to 0.9159. Similarly, **Figure 6B**, which compares methylation data alone to the multi-omics model, demonstrates comparable improvements. These results underscore the power of integrating multiple omics data types, as our multi-omics approach consistently outperforms single-omic models across various performance metrics. This integrated approach allows for a more comprehensive understanding of the underlying biological processes, leading to more accurate and robust predictions.

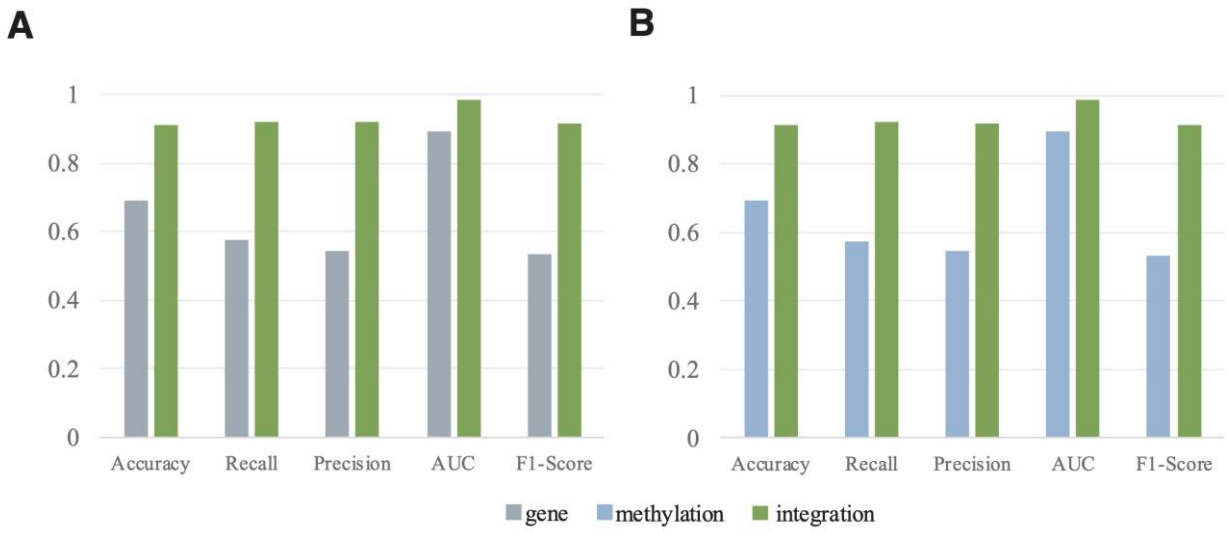

Figure 6. Comparison of classification metrics for single-omic and the multi-omics approach. **(A)** Gene expression; **(B)** methylation.

3.4. Comparison of different machine learning classifiers

The **Figure 7A** presents the classification results of gene expression dataset using machine learning classification algorithms. In terms of accuracy, the LightGBM model performs the best, achieving an accuracy of 0.7941. Regarding the AUC, the ET model exhibits the best performance with an AUC of 0.9211. In terms of recall, all models demonstrate comparable performance, with scores ranging between 0.6471 and 0.7941. Regarding precision, the LightGBM model again shows the best performance, achieving a precision of 0.8123. For the F1 score, the LightGBM model outperforms the others, reaching an F1 score of 0.7943. The **Figure 7B** presents the classification results of methylation dataset using machine learning classification algorithms. In terms of accuracy, the LR model performs the best, achieving an accuracy of 0.7353. Regarding the AUC, the LR model also exhibits the best performance with an AUC of 0.9035. In terms of recall, all models demonstrate comparable performance, with scores ranging between 0.6176 and 0.7353. Regarding precision, the LR model shows the best performance, achieving a precision of 0.7517. For the F1 score, the LR model outperforms the others, reaching an F1 score of 0.7354.

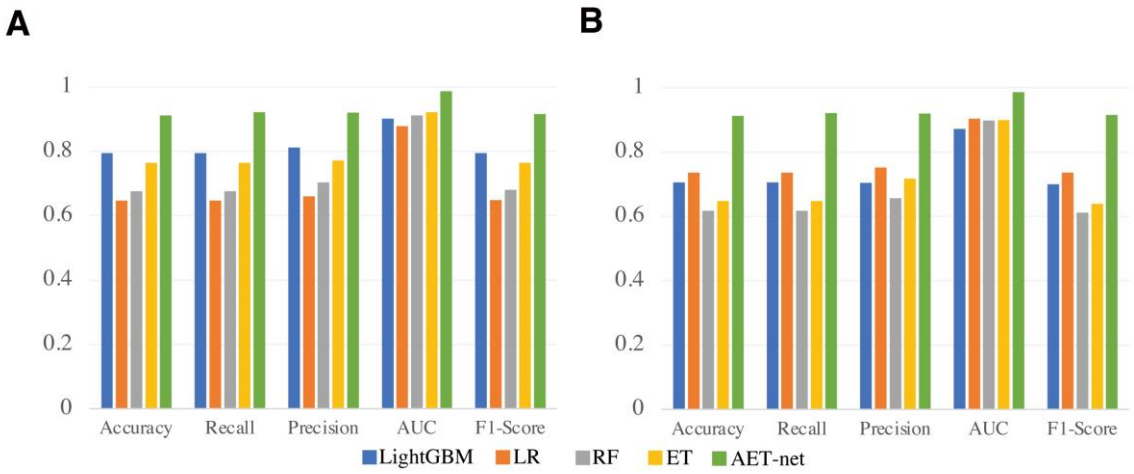

Figure 7. Comparison of classification metrics between AET-net and other classifiers on gene and methylation datasets. **(A)** Gene expression; **(B)** methylation.

From **Table 2**, the model's classification performance progressively improves with increasing integration dimensions.

Table 2. Comparison in different integration dimensions.

Dimension	Accuracy	AUC	Recall	Precision	F1 score
128	0.7701	0.9516	0.7816	0.7749	0.7764
256	0.7606	0.9380	0.7737	0.7678	0.7688
512	0.8127	0.9617	0.8286	0.8168	0.8153
1024	0.8483	0.9679	0.8591	0.8508	0.8534
2048	0.9120	0.9862	0.9218	0.9197	0.9159

From **Tables 3** and **4**, machine learning classification methods do not perform well in the classification of BC subtypes. This observation underscores the need for more sophisticated methods to handle the complexity of the data and the task. Therefore, we have employed an Autoencoder data integration method and built an attention-based neural network, AET-net.

Table 5 presents the classification results of AET-net when applied to multi-omics BC subtype data, with a dataset balanced using SMOTE and a training set comprising 80% of the samples and a test set of 20%. The results demonstrate the network's robust performance in accurately classifying different BC subtypes by integrating diverse omics data.

Table 3. Classification metrics results for gene datasets compared to AET-net.

Classifier	Accuracy	AUC	Recall	Precision	F1 score
LightGBM	0.7941	0.9010	0.7941	0.8123	0.7943
LR	0.6471	0.8782	0.6471	0.6595	0.6481
RF	0.6765	0.9118	0.6765	0.7036	0.6809
ET	0.7647	0.9211	0.7647	0.7721	0.7641
AET-net	0.9120	0.9862	0.9218	0.9197	0.9159

Table 4. Classification metrics results for methylation datasets compared to AET-net.

Classifier	Accuracy	AUC	Recall	Precision	F1 score
LightGBM	0.7059	0.8720	0.7059	0.7044	0.6996
LR	0.7353	0.9035	0.7353	0.7517	0.7354
RF	0.6176	0.8982	0.6176	0.6560	0.6121
ET	0.6471	0.8989	0.6471	0.7173	0.6395
AET-net	0.9120	0.9862	0.9218	0.9197	0.9159

As shown in **Figure 7**, the classification metrics have seen substantial improvements. The accuracy has reached 0.9120, the AUC has increased to 0.9862, the recall has reached 0.9218, the precision has reached 0.9197, and the F1 score has reached 0.9159. The significant improvements can be attributed to the capability of the AET-net to effectively integrate multi-omics data and capture intricate patterns that traditional machine learning methods often overlook. The Autoencoder component of AET-net compresses the high-dimensional multi-omics data into a lower-dimensional space, preserving the essential features that are crucial for the classification task. The attention mechanism within the AET-net allows the model to focus on the most informative features, thereby increasing the model's discriminative power.

Table 5. Results of balanced dataset partitioning in AET-net.

Subtype	Train set	Test set	Correct set
Basal-like	343	79	77
HER2	346	76	76
LumA	325	97	69
LumB	331	91	84
Normal-like	343	79	79

4. Discussion

Our study introduces the AET-net framework, a novel approach for BC subtype classification that advances multi-omics data integration methodologies. A critical aspect of our research focuses on optimizing the multi-omics integration dimensions, which significantly impacts the model's performance. As demonstrated in **Table 2**, increasing the integration dimensions progressively enhances the model's classification capabilities, with the optimal results achieved at a dimension of 2048.

At this optimal dimension, the AET-net classifier demonstrates remarkable performance metrics: An accuracy of 0.912, an AUC of 0.9862, a Recall of 0.9218, a Precision of 0.9197, and an F1 score of 0.9159. These results indicate that a 2048-dimensional integration strikes an ideal balance between capturing detailed biological information and maintaining model efficiency. Comparatively, our approach shows significant improvements over recent multi-omics integration models like Choi and Chae's moBRCA-net and Gao et al.'s DeepCC algorithm. While previous research struggled with the complexities of high-dimensional omics data, our Autoencoder and Transformer-based neural network effectively addresses these challenges. The model enhances classification performance, increasing accuracy from 0.865 to 0.9120 and achieving an AUC of 0.9862. By comprehensively integrating gene expression and DNA methylation data, AET-net captures nuanced biological relationships more effectively than traditional machine learning methods. The framework provides balanced performance across different BC subtypes, offering a valuable tool for diagnostic decision support in precision oncology.

The biological significance of our approach extends beyond computational performance. By integrating gene expression and DNA methylation data, we provide insights into the molecular heterogeneity of BC subtypes. Our analysis reveals intricate interactions between genomic and epigenomic landscapes that characterize different breast cancer molecular profiles. Specifically, the model's ability to capture nuanced molecular relationships suggests distinct epigenetic modifications and transcriptional variations across BC subtypes. These molecular distinctions illuminate potential connections between genetic profiles and cellular mechanical properties, contributing to our understanding of the molecular mechanisms underlying tumor progression, cellular differentiation, and the structural characteristics of breast cancer subtypes.

The method's comprehensive approach not only improves subtype classification accuracy but also demonstrates the potential of advanced deep learning techniques in unraveling the molecular complexity of BC. However, future research should focus

on validating the approach across larger datasets, exploring additional omics data types, and developing more interpretable models to further our understanding of BC molecular heterogeneity.

5. Conclusion

Classification of BC subtypes is essential for individualized treatment, as different subtypes respond differently to treatment. Multi-omics data integration, on the other hand, provides more comprehensive biological information and improves diagnostic and prognostic accuracy. Traditional machine learning classification often relies on specific features and may ignore complex relationships between data. Deep learning has greater potential to process complex data for further inprovement. In this research, we propose a cancer subtype classification framework, called AET-net, based on multi-omics data of BC. We initially obtain multi-omics data from the BRCA project of BC, which includes gene expression data and DNA methylation data, using the R language. The obtained multi-omics data is then integrated using an Autoencoder neural network framework. We establish a Transformer attention mechanism neural network classification framework to classify BC subtypes. Simultaneously, under the same data conditions and dataset partitioning, we compare the performance of other machine learning classifiers. Our experimental results demonstrate AET-net's effectiveness in BC subtype classification, showcasing the potential of deep learning techniques in medical diagnostic support. The method consistently improved classification accuracy across multiple performance metrics, revealing promising avenues for future research. By utilizing advanced classification approaches, AET-net represents a novel method with potential to enhance diagnostic processes.

Author contributions: Conceptualization, QZ and YW; methodology, YW; investigation, JX and ZS; resources, JH; writing—original draft preparation, YW; writing—review and editing, QZ; supervision, HZ and ZZ. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the Postgraduate Research of Lianyungang (Grant No. LYG20210010), the Lianyungang Science and Technology Projects (Grant No. CG2223) and the National Natural Science Foundation of China (Grant No. 72174079).

Ethical approval: Not applicable.

Conflict of interest: The authors declare no conflict of interest.

References

1. Akram M, Iqbal M, Daniyal M, Khan AU. (2017). Awareness and current knowledge of breast cancer. Biological research 50, 1–23
2. Momenimovahed Z, Salehiniya H. (2019). Epidemiological characteristics of and risk factors for breast cancer in the world. Breast Cancer: Targets and Therapy, 151–164
3. Łukasiewicz S, Czeczelewski M, Forma A, Baj J, Sitarz R, Stanisławek A. (2021). Breast cancer—epidemiology, risk factors, classification, prognostic markers, and current treatment strategies—an updated review. Cancers 13, 4287

4. Runel, G., Lopez-Ramirez, N., Chlasta, J., & Masse, I. (2021). Biomechanical properties of cancer cells. Cells 10(4), 887
5. McGhee, D. E., & Steele, J. R. (2020). Biomechanics of breast support for active women. Exercise and sport sciences reviews 48(3), 99-109
6. Tarchi, S. M., Pernia Marin, M., Hossain, M. M., & Salvatore, M. (2023). Breast stiffness, a risk factor for cancer and the role of radiology for diagnosis. Journal of Translational Medicine 21(1), 582
7. Shah, L., Latif, A., Williams, K. J., & Tirella, A. (2022). Role of stiffness and physico-chemical properties of tumour microenvironment on breast cancer cell stemness. Acta Biomaterialia 152, 273-289
8. Rajpal S, Rajpal A, Saggar A, Vaid AK, Kumar V, Agarwal M. Kumar N. (2023). XAI-MethylMarker: Explainable AI approach for biomarker discovery for breast cancer subtype classification using methylation data. Expert Systems with Applications 225, 120130
9. Rakha EA, Tse GM, Quinn CM. (2023). An update on the pathological classification of breast cancer. Histopathology 82(1), 5–16
10. Asleh K, Lluch A, Goytain A, Barrios C, Wang XQ, Torrecillas L, et al. (2023). Triple-negative pam50 non-basal breast cancer subtype predicts benefit from extended adjuvant capecitabine. Clinical Cancer Research 29, 389–400
11. Azevedo A L K, Gomig T H B, Batista M, et al. (2023). High-throughput proteomics of breast cancer subtypes: Biological characterization and multiple candidate biomarker panels to patients' stratification. Journal of Proteomics 285, 104955
12. Orsini A, Diquigiovanni C, Bonora E. (2023). Omics technologies improving breast cancer research and diagnostics. International Journal of Molecular Sciences 24(16), 12690
13. Heo YJ, Hwa C, Lee GH, Park JM, An JY. (2021). Integrative multi-omics approaches in cancer research: from biological networks to clinical subtypes. Molecules and cells 44(7), 433–443
14. Wang, Z.-z., Li, X.-h., Wen, X.-l., Wang, N., Guo, Y., Zhu, X., et al. (2023). Integration of multi-omics data reveals a novel hybrid breast cancer subtype and its biomarkers. Frontiers in Oncology 13, 1130092
15. Choi, J. M. and Chae, H. (2023). mobrca-net: a breast cancer subtype classification framework based on multi-omics attention neural networks. BMC bioinformatics 24, 169
16. Zubair, M., Wang, S., and Ali, N. (2021). Advanced approaches to breast cancer classification and diagnosis. Frontiers in Pharmacology 11, 632079
17. Gao, F., Wang, W., Tan, M., Zhu, L., Zhang, Y., Fessler, E., et al. (2019). Deepcc: a novel deep learning-based framework for cancer molecular subtype classification. Oncogenesis 8, 44
18. Meti, N., Saednia, K., Lagree, A., Tabbarah, S., Mohebpour, M., Kiss, A., et al. (2021). Machine learning frameworks to predict neoadjuvant chemotherapy response in breast cancer using clinical and pathological features. JCO Clinical Cancer Informatics 5, 66–80
19. Graudenzi, A., Cava, C., Bertoli, G., Fromm, B., Flatmark, K., Mauri, G., et al. (2017). Pathway-based classification of breast cancer subtypes. Front Biosci 22, 1697–1712
20. Mohammed, A. J., Hassan, M. M., and Kadir, D. H. (2020). Improving classification performance for a novel imbalanced medical dataset using smote method. International Journal of Advanced Trends in Computer Science and Engineering 9, 3161–3172
21. Satpathi, S., Gaurkar, S. S., Potdukhe, A., and Wanjari, M. B. (2023). Unveiling the role of hormonal imbalance in breast cancer development: A comprehensive review. Cureus 15
22. Choi, S. R. and Lee, M. (2023). Transformer architecture and attention mechanisms in genome data analysis: a comprehensive review. Biology 12, 1033
23. Thennavan, A., Beca, F., Xia, Y., Garcia-Recio, S., Allison, K., Collins, L. C., et al. (2021). Molecular analysis of tcga breast cancer histologic types. Cell genomics 1
24. Huang, R., Soneson, C., Ernst, F. G., Rue-Albrecht, K. C., Yu, G., Hicks, S. C., et al. (2020). Treesummarizedexperiment: a s4 class for data with hierarchical structure. F1000Research 9
25. Parker, J. S., Mullins, M., Cheang, M. C., Leung, S., Voduc, D., Vickery, T., et al. (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. Journal of clinical oncology 27, 1160
26. Nielsen, T. O., Leung, S. C. Y., Rimm, D. L., Dodson, A., Acs, B., Badve, S., et al. (2021). Assessment of ki67 in breast cancer: updated recommendations from the international ki67 in breast cancer working group. JNCI: Journal of the National Cancer Institute 113, 808–819

27. Voineskos, S. H., Klassen, A. F., Cano, S. J., Pusic, A. L., and Gibbons, C. J. (2020). Giving meaning to differences in breast-q scores: minimal important difference for breast reconstruction patients. Plastic and reconstructive surgery 145, 11e–20e
28. Yan, R., Zhang, F., Rao, X., Lv, Z., Li, J., Zhang, L., et al. (2021). Richer fusion network for breast cancer classification based on multimodal data. BMC Medical Informatics and Decision Making 21, 1–15
29. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. Advances in neural information processing systems 30
30. Kotsiantis, S. B., Zaharakis, I. D., and Pintelas, P. E. (2006). Machine learning: a review of classification and combining techniques. Artificial Intelligence Review 26, 159–190