Progress in Microbes and Molecular Biology

HH PUBLISHER

Original Article

Breast Cancer Classification Through Integrated Omics Variations Using Stacked Denoising Autoencoder (SDAE) and Variational Autoencoder (VAE) Algorithms

Felicia Chin Hui Fen [1], Phang Cheng Yi [2], Goh Yitian [3], Mohd Firdaus [4]

Article History

Received: 15 July 2024;

[1]School of Computing, Universiti Teknologi Malaysia, Skudai, Johor, Malaysia, feliciahui@graduate.utm.my (F. H. F. Chin)

[2]School of Computing, Universiti Teknologi Malaysia, Skudai, Johor, Malaysia, phangyi@graduate.utm.my (C. Y. Phang)

[3]School of Computing, Universiti Teknologi Malaysia, Skudai, Johor, Malaysia, gohyitian@graduate.utm.my (Y. Goh)

[4]School of Computing, Universiti Teknologi Malaysia, Skudai, Johor, Malaysia, mohd01@graduate.utm.my (M. Firdaus)

Abstract: Breast cancer, as noted by the World Health Organisation (WHO), involves the uncontrolled growth of abnormal cells in the breast, potentially leading to fatal outcomes if untreated. This research aims to enhance the classification of breast cancer subtypes through the integration of multi-omics data using Stacked Denoising Autoencoders (SDAE) and Variational Autoencoders (VAE). We utilized a comprehensive dataset compromising copy number variation, miRNA expression, DNA methylation and mRNA expression data. Preprocessing steps included checking for missing values, transposing datasets and implementing the Synthetic Minority Oversampling Technique (SMOTE) to address data imbalance. Key methods such as Support Vector Machine-Recursive Feature Elimination (SVM-RFE) were employed for feature selection, and multi-omic integration was performed using concatenation-based methods. Results obtained indicate that the VAE model consistently outperformed the SDAE model across all omics data types, achieving the highest accuracy of 81.68% with mRNA expression data for both SMOTE and without SMOTE. The application of SMOTE generally did not improve model accuracy, likely due to the introduction of noise through synthetic features. Notably, single-omics data, particularly mRNA expression, proved to be more effective for breast cancer subtype classification than multi-omics data. In conclusion, VAE demonstrates superior performance over SDAE for breast cancer classification, and mRNA expression is identified as the most reliable marker. Future research should explore alternative omics data combinations, integration methods, and feature selection techniques to further enhance classification accuracy.

Keywords: breast cancer classification, SDAE model, VAE model, SVM-RFE, SMOTE

1. Introduction

According to the World Health Organisation (WHO)[1], breast cancer is a condition characterized by the uncontrolled growth of abnormal cells in the breast, leading to the formation of tumors. Without treatment, these tumors can spread to other parts of the body, becoming fatal. The disease typically starts in the milk ducts or milk-producing lobules of the breast. The earliest stage, known as in situ, is not life-threatening and can be detected early. If the cancer cells invade nearby breast tissue, they form tumors, causing lumps or thickening in the breast. Early detection and the removal of stigma are vital for lowering mortality rates and improving mental health[2].

Breast cancer is a highly heterogeneous disease, consisting of various biological subtypes. Each subtype has unique clinical presentations, pathological characteristics, and molecular signatures, which influence their prognosis and treatment options[3]. Therefore, accurately classifying breast cancer subtypes is crucial for precise treatment and prognosis prediction. Driven by the latest high-throughput sequencing technologies, biological data in various formats, sizes, and structures are growing rapidly[4-6]. Based on these omics data, many studies on breast cancer subtype classification have emerged, which can be categorized into two types. The first type relies on single omics data, and the second type uses multi-omics data[7]. Studies have shown that combining multiple omics datasets improves the accuracy of clinical outcome predictions, highlighting the importance of integrating multi-omics data over single-omics data[8-10]. Based on the integration method, multi-omics data integration techniques for predicting breast cancer subtypes can be classified as concatenation-based, ensemble-based, and knowledge-driven methods[11].There are 5 intrinsic subtypes which are luminal A, luminal B, HER2 overexpression, basal-like, and normal-like cancer[12]. The multi-omics data is generated by concatenating the data from copy number variation, mRNA expression, DNA methylation, and microRNA (miRNA) expression into a single dataset before training.

There are several problem statements in this research. First, one of the methods to perform Multi-omics analysis in order to classify cancer is using Stacked Denoising Autoencoder (SDAE). Second, the high computational costs and time consumption might result from computing Stacked Denoising Autoencoder (SDAE) with millions of parameters in a single training process. Lastly, the lack of an appropriate classification technique to incorporate reliable and instructive genomic data results in significant bias. Thus, this research aims to classify breast cancer based on integration of omics variation data using SDAE and VAE. It also aims to analyze the performance of SDAE and VAE in omics data analysis. Lastly, it will evaluate the classification accuracy obtained from SDAE and VAE by implementing integrated variation of omics data. Furthermore, several research questions are outlined to help achieve the research objectives. These include: are the Variational Autoencoders (VAE) can outperform Stacked Denoising Autoencoders (SDAE) in integrated omics data for breast cancer classification? What techniques can be used for feature selection and dimensionality reduction to decrease the number of parameters in SDAE and VAE? Does integrating multiple types of omics data result in better accuracy and classification performance compared to analyzing single omics data alone? Do SDAE and VAE take into account the relationships between different types of omics data? This report consists of four chapters, which are introduction, material and methods, results, discussion and finally conclusion.

2. Literature Review

There are four articles that have been studied and reviewed. Table 1 shows the summary of literature review had been done by the previous researcher in the classification of breast cancer using omics data.

Table 1. Summary of literature review.

Author(s) & Year	Title	Methodology	Result
(Huang et al., 2024)[8]	Classifying breast cancer subtypes on multi-omics data via sparse canonical correlation analysis and deep learning	● Differential Sparse Canonical Correlation Analysis Network (DSCCN) ● Logistic regression model/multinomial model with Elastic Net (EN) regularization ● Random Forest (RF) ● DeepMo ● Sparse Multi-View Partial Least Square ● DIABLO	● 83.45% of accuracy for 5 subtypes. ● 77.55% of accuracy for 10 subtypes.
(Ren et al., 2024)[13]	Classifying breast cancer using multi-view graph neural network based on multi-omics data	● Multi-view Graph Neural Network (MVGNN)	● MVGNN model with all three types of omics data perform better than those using two or just one type of omics data.
(Choi and Chae, 2023)[14]	moBRCA-net: a breast cancer subtype classification framework based on multi-omics attention neural networks	● proposed moBRCA-net ● Support vector machine (SVM) ● Random Forest (RF) ● Logistic Regression (LR)	● moBRCA-net outperformed other machine learning with an average accuracy of 0.891.

		• Naive Bayes	
(El-Nabawy et al., 2021)[15]	A Cascade Deep Forest Model for Breast Cancer Subtype Classification Using Multi-Omics Data	• Cascade Deep Forest model	• 83.45% of accuracy for 5 subtypes. • 77.55% of accuracy for 10 subtypes.

[1] The table presents the summary of recent studies in the classification of breast cancer [8,13-15].

From Table 1, the recent studies have shown the effectiveness of deep learning in classifying breast cancer using the subtype in the omics data. However, comparisons between single omics and multi-omics data are seldom done. Furthermore, the use of autoencoders is also less implemented in the deep learning approaches by the recent studies. Thus, in this research, the classification of breast cancer will be conducted using the SDAE model and VAE model in various omics data, which includes both single omics and multi-omics.

3. Materials and Methods

The research operational framework design for this study is illustrated in Figure 1 below.

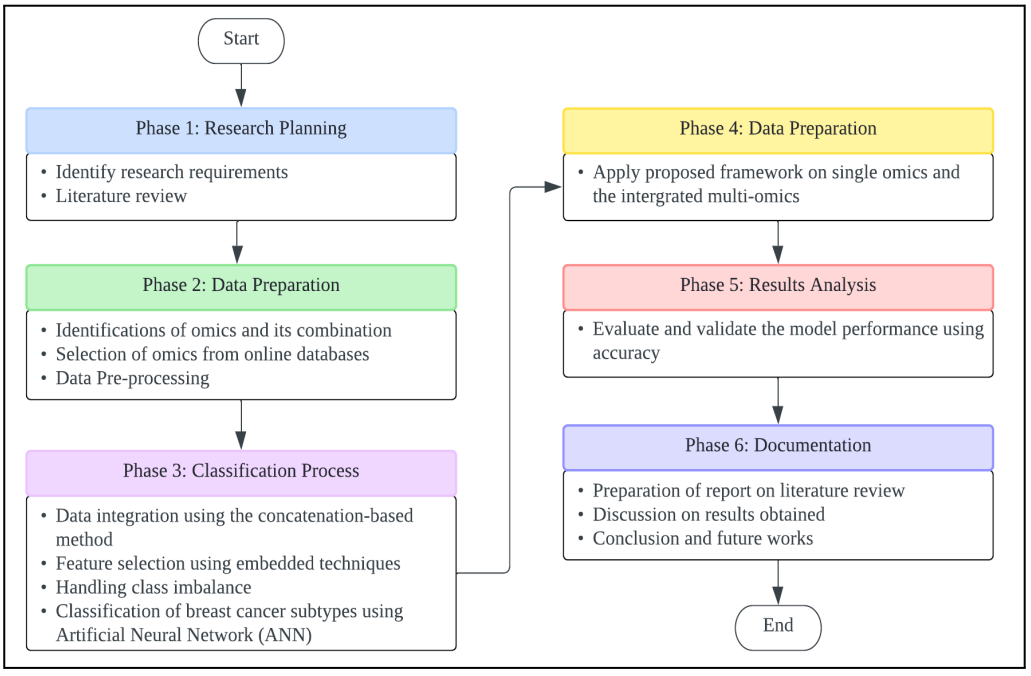

Figure 1. Research Operational Framework

The research methodology of this study is separated into four phases. First phase is research planning, second phase is data preparation, third phase is classification process, fourth phase is experimentation and results gathering, fifth phase is results analysis and the sixth phase is documentation.

To better comprehend the procedures and steps needed in implementing the algorithms and techniques involved into operation, experimental workflow has been carried out. The experimental design for this study is illustrated in Figure 2 below.

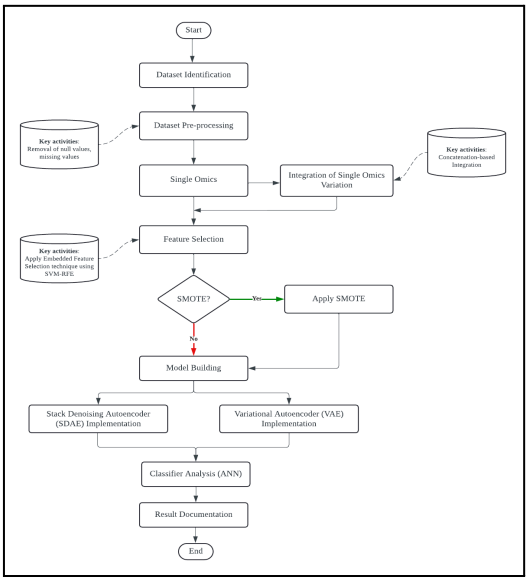

Figure 2. Experimental Design

3.1 Omics Dataset

The cancer dataset used in this study is breast cancer dataset provided by Dr. Azurah Binti A. Samah. Table 2 shows the summary of breast cancer dataset.

Table 2. Summary of Breast Cancer Dataset.

Omics Data	Omics field	Number of Patients	Number of Features
Copy Number Variation	Genomics	671	19,568
miRNA Expression	Epigenetics	671	369
DNA methylation	Methylomics	671	19,050
mRNA Expression	Transcriptomics	671	18,207

[2] The table presents the omics field, total number of patients and features for each single omic.

3.2 Data Pre-processing

The omics dataset obtained has previously undergone preprocessing. To ensure the quality of the dataset is good to be used, a rechecking step on the dataset is necessary before proceeding to the next steps. A checking on missing values is done for each omics dataset. Besides omics dataset, there is another dataset called brca_label dataset which consists of the subtypes of breast cancer identified for each patient. For this dataset, the subtypes labels in string form are being replaced with numerical values, where 0 represents

normal-like, 1 represents luminal A, 2 represents luminal B, 3 represents basal-like, and 4 represents HER2 overexpression. After completing missing values checking and label replacement using numbers, the rows and columns of the omics dataset are being transposed for the insertion of a new label column into each omics dataset. Finally, the preprocessed datasets were splitted using a train test ratio of 70:30 in which 70% is training data and 30% is testing data.

3.3 Multi-omics Integration

The integration process is implemented after each of the omics dataset was transposed and added a new label column. The preprocessed single omics data are integrated using a concatenate method and the final data was combined into a single, large matrix before being employed to train the model. Table 3 summarizes the number of patients and features after data integration.

Table 3. Summary of the number of samples and features after data integration.

Omics Data	Number of Patients	Number of Features
Multi-omics	671	57,192

[3] The table presents the total number of patients and features after integration of multiple single omics..

3.4 Feature Selection

Feature selection is the process of producing a subset from an original feature set based on a specific feature selection criterion, which decides the relevant features of the dataset[16]. It contributes to data compression by effectively reducing the dimensionality of the feature space by deleting unnecessary and duplicate characteristics while having no substantial impact on the trained model's decision-making quality[17].

In this study, SVM-RFE method is employed. It is a recursive feature reduction technique that utilizes SVM weights as a criterion for ranking[18]. SVM-RFE's key concept is to remove features that have the lowest squares of weight in each iteration. SVM-RFE technique searches for a subset of features by going around all the features available to be iterated and finally removing a certain amount of features, leaving the desired number of data available. For this study, a non-linear SVM-based RFE has been implemented to eliminate the least relevant features in predicting target variables.

Input: Training data $\{x_i, y_i\}_{i=1}^{N}$

Output: Ranked feature list R

Initialize: $S = \{1, 2, \ldots, D\}_i$

$R = \phi$

While S is not empty, do:

1. Restrict the features of x_j to the remaining S

2. Get weight vectors by training SVM

3. Compute the ranking criteria $c_k = w_k^2,\ k = 1, \dots, |S|$

4. Find features with lowest value of c_k, called feature p

5. Add feature p into R ($R = \{p\}\ R$)

6. Remove feature p from S ($S = S \backslash p$)

Algorithm: Feature Selection based on SVM-RFE

The procedure that has been carried out in our case study was highly similar to the recursive process of the feature removal method in general:

1. Train classifier to find the weight vector (w).
2. Calculate criteria of ranking for all features.
3. Dispose features with the lowest rating criterion value.

Features used in the iteration had to be removed with backward feature elimination. The ranking score is given according to the components of SVM's weight vector w:

$$w = \Sigma_k a_k y_k x_k \quad (1)$$

In our case, the omics dataset was aggregated into two portions termed "features" and "targets". This step is taken to ensure SVM-RFE based feature selection is made without interfering with independent variables such as class labels during the biological data ranking process. Table 4 below shows the number of the single omics and multi-omics features before and after the implementation of the feature selection technique.

Table 4. Implementation of SVM-RFE technique for Feature Selection.

Omics Data	Before SVM-RFE	After SVM-RFE
Copy Number Variation	19,568	19,000
miRNA Expression	369	300
DNA methylation	19,050	19,000
mRNA Expression	18,207	18,000
Multi-omics	57,192	57,000

[4] The table presents the comparison between the number of features of each omic before and after applying the feature selection technique - SVM-RFE.

3.5 Synthetic Minority Oversampling Technique (SMOTE)

SMOTE is one of the resampling techniques that handle imbalance data by oversampling the minority class which involves the creation of synthetic examples[19]. The imbalance data is monitored after splitting the train test data. In the training data, there are only 21 patients had normal-like cancer subtype, 25 patients had HER2 overexpression cancer subtype, 82 patients had basal-like cancer subtype, 96 patients had luminal B cancer subtype and 245 patients had luminal A cancer subtype. It can be seen that luminal A has the highest frequency which is more than 200 while the other subtypes have frequency less than 100. The class labels in the form of subtypes of breast cancer boost supervised cancer classification and facilitate omics feature prediction. Table 5 below shows the result of the breast cancer subtypes data before and after implementing SMOTE.

Table 5. The occurrences of each breast cancer subtypes.

Breast Cancer Subtypes	Before SMOTE	After SMOTE
Normal-like	21	245
Luminal A	245	245
Luminal B	96	245
Basal-like	82	245
HER2	25	245

[5] The table presents the number of samples of each breast cancer subtypes in the training data before and after applying the imbalance data technique - SMOTE.

3.6 Stacked Denoising Autoencoder (SDAE) and Variational Autoencoder (VAE) Implementation

The models used in this study were SDAE and VAE. The models were designed to analyze the contribution of integrated multi-omics data in classifying the breast cancer subtypes. SDAE model architecture consists of an input layer with the number of features equal to the input data dimensions, followed by two hidden layers with 64 and 32 neurons, respectively. The encoded representations are compressed to 16 dimensions in the bottleneck layer. The decoding part mirrors the encoding part with two hidden layers, reconstructing the input data in the output layer. The model is compiled with the Adam optimizer and mean-squared error loss. Early stopping is applied to halt training if the validation loss does not improve for 10 consecutive epochs. The SDAE is trained on the training data for 250 epochs with a batch size of 32, using validation data to monitor performance.

For the VAE model, the data is converted into numpy arrays and normalized using MinMaxScaler initially. The data is then reshaped into the required format with specified height, width and channels according to the dimension of datasets. The VAE model comprises an encoder, a sampling layer, and a decoder. The encoder consists of three

convolutional layers (16, 32, 64 neurons), followed by a flatten layer and two denser layers to compute the mean and log variance of the latent variables with a dimensionality of 64. The sampling layer generates latent vectors from these parameters. The decoder reconstructs the original input from the latent space using a dense layer, reshape layer, and three transpose convolutional layers (32, 16, 1 neurons). The VAE model integrates the encoder and decoder, tracking total loss, reconstruction loss, and KL divergence. It includes custom training and testing steps to optimize and evaluate these losses. Finally, the VAE model is compiled with the Adam optimizer and trained on the reshaped training data for 15 epochs with a batch size of 32, using validation on the reshaped test data to ensure effective feature learning and reconstruction.

3.7 Classifier Analysis (Artificial Neural Network)

The model accuracy for SDAE and VAE models, both with SMOTE and without SMOTE are evaluated by fitting the trained data into an artificial neural network (ANN) classifier. The testing data are used to make predictions on the SDAE and VAE models. The ANN classifier was built starting with the calculation of the number of unique classes in the target variable. The model consists of three Dense layers: the first layer with 64 neurons and ReLU activation, followed by a Dropout layer with a 50% rate, a second Dense layer with 32 neurons and ReLU activation then with another Dropout layer, and a final Dense layer with softmax activation for output. The model is compiled with the Adam optimizer and sparse categorical cross-entropy loss, and it tracks accuracy as a metric. Early stopping is implemented to halt training if the validation loss does not improve for 10 consecutive epochs. The classifier is trained on encoded training data for up to 250 epochs with a batch size of 32, using separate validation data to monitor performance. This setting is applied for all datasets with SMOTE and without SMOTE.

4. Results

The result for each omics data built using SDAE model and VAE model, with and without SMOTE are recorded in the Table 6 below.

Table 6. Accuracy (%) result for SDAE model and VAE model with and without SMOTE on Omics Data.

Omics Data	SDAE with SMOTE	SDAE without SMOTE	VAE with SMOTE	VAE without SMOTE
Copy Number Variation	62.38%	63.36%	65.84%	70.30%
miRNA Expression	66.34%	69.31%	75.74%	70.79%
DNA methylation	50.50%	64.85%	66.34%	70.30%
mRNA Expression	72.28%	73.76%	81.68%	81.68%
Multi-Omics	64.85%	65.84%	78.22%	76.73%

[6] The table presents the accuracy (in percentages) for various omics data types when using Stacked Denoising Autoencoder (SDAE) and Variational Autoencoder (VAE) models, with and without the Synthetic Minority Over-sampling Technique (SMOTE).

From Table 6, it shows that the range of accuracy for the SDAE model on the omics data is from 63.36% to 73.76%. When the SDAE model fits with Copy Number Variation omics data, it has the lowest accuracy value, 63.36% among the four single omics data. In contrast, when it fits with mRNA Expression omics data, it has the highest accuracy value, 73.76% compared to other three single omics data. Furthermore, the miRNA Expression has the second highest accuracy value, 69.31%. When integrating all omics data into multi-omics, the accuracy value of the SDAE model is 65.84%, the third highest accuracy value. The DNA methylation has an accuracy value, 64.85%, which is slightly higher than Copy Number Variation.

When SMOTE is applied in the dataset, all accuracy values of the SDAE model decrease. The accuracy value of Copy Number Variation decreased by 0.98%, miRNA Expression decreased by 3.27%, DNA Methylation decreased by 14.35%, mRNA Expression decreased by 1.48% and multi-omics data decreased by 0.99%.

For the VAE model, the range of accuracy is from 70.30% to 81.68%. The accuracy value of Copy Number Variation and DNA methylation are the lowest among the omics data, 70.30% whereas the mRNA Expression has the highest accuracy value, 81.68%. The second rank of accuracy value among omics data using VAE model is multi-omics data 76.73%, then followed by miRNA Expression, 70.79%.

When SMOTE is applied in the dataset, the accuracy value for miRNA Expression (increased by 4.95%) and multi-omics data (increased by 1.49%) are increased. In contrast, accuracy values for Copy Number Variation (decreased by 4.46%) and DNA Methylation (decreased by 3.96%) are decreased. The accuracy value for mRNA Expression remains unchanged (81.68%).

5. Discussion

From the result obtained in Table 6, from the comparison between SDAE and VAE model, it shows that VAE model is better than SDAE model, either using SMOTE or without SMOTE. For the Copy Variation Number, VAE has a higher accuracy value, 70.30% (without SMOTE) and 65.84% (with SMOTE) compared to the SDAE model, 63.36% (without SMOTE) and 65.84% (with SMOTE). Furthermore, for the DNA Methylation data, the VAE model has the same highest accuracy value of Copy Variation Number data, 70.30%. When using the multi-omics data, VAE model also outperforms the SDAE model, where the accuracy value (78.22% for SMOTE, 76.73% for without SMOTE) is higher than SDAE model for both SMOTE (64.85%) and without SMOTE (65.84%). When comparing across the omics data, both models outperform when using the mRNA Expression dataset in either SMOTE (72.28% for SDAE, 81.68% for VAE) or without SMOTE (73.76% for SDAE, 81.68% for VAE). The accuracy value for the VAE model remains unchanged in either using SMOTE or without SMOTE. Generally, the VAE model has a higher mean accuracy value of 6.54% than the SDAE model and also has a higher mean accuracy value of 14.29% than the SDAE model when using SMOTE.

Furthermore, when comparing the use of SMOTE and without SMOTE in both models, it shows that the SMOTE method generally does not improve the accuracy results, and even the accuracy result is decreased. For the SDAE model, all the accuracy values decrease with a mean accuracy value of 4.75% after applying the SMOTE method. This may due to the synthetic features generated by SMOTE may contain noise that affect the model training process which causes the accuracy value to decrease. In contrast, for the VAE

model, only the accuracy value of miRNA Expression and multi-omics data increases by mean accuracy of 3.22% after applying the SMOTE.

When comparing the single omics and multi-omics data, both models outperformed when using the single omics data, mRNA expression in classifying breast cancer, SDAE with accuracy of 73.76% and 72.28% with SMOTE; VAE with accuracy value of 81.68% for SMOTE and without SMOTE. This means that mRNA expression is suitable to be used as a marker to classify the breast cancer subtypes as both models, SDAE and VAE achieved the highest accuracy values, either using SMOTE or without SMOTE.

6. Conclusions

In conclusion, the VAE model is more suitable than the SDAE model to classify breast cancer as the accuracy obtained by the VAE model is higher than the SDAE model. Generally, SMOTE does not help to improve the accuracy value, this may be due to synthetic features generated by SMOTE consisting of noise. From the result, mRNA expression is the most suitable data for the classification of breast cancer for both SDAE and VAE models, compared to other single omics data, even the multi-omics data. In summary, the objectives of this paper have been achieved. Firstly, breast cancer has been classified successfully based on various omics data using the SDAE and VAE models, where the mRNA expression data achieved the best result of classification of breast cancer. Besides that, the performance of the SDAE model and VAE model have been analyzed, where the VAE model achieved better performance than the SDAE model. Lastly, the classification accuracy obtained from the SDAE model and VAE model using various omics data has been evaluated, where the classification accuracy obtained from the VAE model in mRNA expression data has the highest value, which is 81.68%. There are some future improvements to this research. Firstly, identify other combinations of omics data for classification. Other than that, identify the integration method for multi-omics other than concatenation based methods. Lastly, identify other feature selection techniques other than SVM-RFE.

Acknowledgments: In preparing this research, we were in contact with many friends and lecturers. They have contributed towards our understanding and thoughts. In particular, we wish to express our sincere appreciation to, Dr. Azurah Binti A. Samah, for encouragement, guidance, critics and friendship. Without her continued support and interest, this research would not have been the same as presented here. Besides that, we would like to acknowledge the unwavering support and understanding of our family. We were humbled and honored to have had the opportunity to work with such exceptional individuals who have contributed to this thesis in various ways. Their collective efforts have been instrumental in shaping this work and have undoubtedly left a lasting impact on our personal and academic growth.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. World Health Organisation. Breast cancer [Internet]. 2024 [cited 2024 Jul 13]. Available from: https://www.who.int/news-room/fact-sheets/detail/breast-cancer

2. Duffy SW, Tabar L, Yen AM, et al. Mammography screening reduces rates of advanced and fatal breast cancers: Results in 549,091 women. Cancer 2020; 126(13):2971-2979. doi: 10.1002/cncr.32859. PMID: 32390151.

3. Zeng P, Huang C, Huang Y. DiffRS-net: A novel framework for classifying breast cancer subtypes on multi-omics data. Appl Sci. 2024; 14:2728. doi: 10.3390/app14072728.

4. National Institutes of Health (US). End-of-life care. National Institutes of Health statement on the state of the science. AWHONN Lifelines 2005; 9(1):15–22.

5. Waks AG, Winer EP. Breast cancer treatment: a review. JAMA. 2019; 321(3):288–300.

6. Yersal O, Barutca S. Biological subtypes of breast cancer: prognostic and therapeutic implications. World J Clin Oncol. 2014; 5(3):412–24.

7. Khan D, Shedole S. Leveraging deep learning techniques and integrated omics data for tailored treatment of breast cancer. J Pers Med. 2022; 12:674.

8. Huang Y, Zeng P, Zhong C. Classifying breast cancer subtypes on multi-omics data via sparse canonical correlation analysis and deep learning. BMC Bioinformatics. 2024; 25(1):132. doi: 10.1186/s12859-024-05749-y.

9. Huang S, Chaudhary K, Garmire LX. More is better: recent progress in multi-omics data integration methods. Front Genet. 2017; 8:268903.

10. Argelaguet R, Velten B, Arnol D, et al. Multi-omics factor analysis—a framework for unsupervised integration of multi-omics data sets. Mol Syst Biol. 2018; 14(6)
.

11. Conesa A, Beck S. Making multi-omics data accessible to researchers. Sci Data. 2019; 6(1):251.

12. Sorlie T, Tibshirani R, Parker J, et al. Repeated observation of breast tumor subtypes in independent gene expression data sets. Proc Natl Acad Sci U S A. 2003; 100(14):8418-23. doi: 10.1073/pnas.0932692100. PMID: 12829800; PMCID: PMC166244.

13. Ren Y, Gao Y, Du W, et al. Classifying breast cancer using multi-view graph neural network based on multi-omics data. Front Genet. 2024; 15:1363896.

14. Choi JM, Chae H. moBRCA-net: a breast cancer subtype classification framework based on multi-omics attention neural networks. BMC Bioinformatics. 2023; 24:169.

15. El-Nabawy AA, Belal NA, El-Bendary N. A cascade deep forest model for breast cancer subtype classification using multi-omics data. Mathematics. 2021; 9:1574.

16. Cai J, Luo J, Wang S, Yang S. Feature selection in machine learning: a new perspective. Neurocomputing. 2018; 300:70-79.

17. Theng D, Bhoyar KK. Feature selection techniques for machine learning: a survey of more than two decades of research. Knowl Inf Syst. 2024; 66:1575-1637.

18. Lin X, Li C, Zhang Y, et al. Selecting feature subsets based on SVM-RFE and the overlapping ratio with applications in bioinformatics. Molecules. 2018; 23(1):52.

19. Chawla N, Bowyer K, Hall L, Kegelmeyer W. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res. 2002; 16:321-357.