Progress in Microbes and Molecular Biology

HH PUBLISHER

Original Article

Breast Cancer Subtypes Classification using SDAE and 1D CNN based on Omics Variation

Chong Kah Wei [1], Heong Yi Qing [1], Mek Zhi Qing [1], Zereen Teo Huey Huey [1]

[1]Faculty of Computing, Universiti Teknologi Malaysia, 81310 Johor Bahru, Johor, Malaysia; chongwei@graduate.utm.my (CKW); heong@graduate.utm.my (HYQ); mekqing@graduate.utm.my (MZQ); zereen@graduate.utm.my (ZT)

Article History

Received: 16 July 2024;

Received in Revised Form:

XX Month 20XX;

Accepted: XX Month 20XX;

Available Online: XX Month 20XX

Abstract: The classification of Breast Cancer subtypes is crucial for categorizing patients and providing appropriate treatment based on the specific type of cancer. Different types of omics data can offer unique insights into Breast Cancer subtypes. This study aims to classify Breast Cancer subtypes through the integration of omics data using deep learning approaches. The datasets used include CNV, DNA methylation, miRNA, and their integrated data. The models employed in this study are SDAE and 1D CNN, and their performance in terms of accuracy is compared based on the type of omics data and the application of SMOTE. The results show that the overall performance of the 1D CNN model is better than the SDAE model when using DNA methylation (79.70% vs. 76.24%), miRNA (76.73% vs. 74.75%), and integrated data (77.23% vs. 70.30%), except for the CNV dataset (68.81% vs. 70.79%). Additionally, applying SMOTE to the datasets improved performance, particularly in the 1D CNN model, which achieved the best performance of 81.19% with DNA methylation data. However, the integrated data did not perform as well as other datasets. For the SDAE model, the accuracy was 71.29%, lower than CNV (71.78%), DNA methylation (76.73%), and miRNA (79.21%). Similarly, for the 1D CNN model, the accuracy with integrated data was 80.20%, lower than DNA methylation (81.19%). In conclusion, SMOTE helps improve the accuracy of Breast Cancer subtype classification, especially for the 1D CNN model, which is more sensitive to class imbalance. However, the integrated data did not produce the best results for either model, indicating a need for further research on model development and techniques to better utilize integrated data.

Keywords: Breast Cancer subtype; multi-omics; 1D CNN; SDAE

1. Introduction

Breast Cancer (BRCA) was the leading cause of death for women all over the countries. 99% of the patients who are diagnosed with BRCA are women and only 1% of patients who are diagnosed with BRCA are men [1]. The classification of BRCA subtypes is significant as

it helps to categorize patients who may benefit from targeted treatment to receive the appropriate treatment [2]. This may help to avoid the side effects caused by unnecessary treatment and further saving the money of patients in recurring.

The development of technology has led to the increasing number and types of omics data that can be obtained and applied to study the BRCA subtype. Different omics data usually will provide different information that is related to the disease. However, the study of one type of omics data is restricted to correlation, which mostly reflects the reaction process rather than the causative process. Multi-omics data are intended to increase the characterization of cross-molecular biological processes and provide more extensive insights into the biological systems [3]. Hence, there are more researchers who use multi-omics data to study the BRCA subtype recently. In this research, different omics of data are used including miRNA data, DNA methylation data and Copy Number Variation (CNV) data. It was believed that through the integration of these data, the analysis of results can provide a greater and deeper understanding of the prognostic phenotypes and dissect cellular responses to therapy [4]. Meanwhile, single level data only provided limited information about the cellular function.

Stacked Denoising Autoencoder (SDAE) is an enhanced version of DAE that stacks multiple denoising auto-encoders together. In comparison to shallow neural networks, SDAE is able to handle more complex relationships from input to the output layers [5]. It has been widely used in different fields such as image denoising, disease detection, and drug discovery. Other than that, 1D Convolutional Neural Network (1D CNN) is a powerful deep neural network that has been applied by researchers in handling multi-omics data. This is because 1D CNN showed good ability in handling complex multi-omics data in comparison with the traditional machine learning method [6]. Hence, both methods are applied in this research.

The imbalance distribution of BRCA subtype may cause the classifier to become biased. Synthetic Minority Oversampling Technique (SMOTE) was a popular data resampling technique to generate synthetic samples for the minority class of the training data [7]. Through SMOTE, a balanced class distribution will be created for all of the BRCA subtypes and prevent the classifier from only detecting the majority subtypes.

This research aims to classify the BRCA subtypes based on the integration of omics data based on the deep learning approaches. The performance of SDAE and 1D CNN model will be evaluated and discussed according to accuracy. Additionally, the effect of applying SMOTE to the data will also be included.

2. Literature Review

The related studies that have been done by the previous researchers are reviewed and recorded to gain more understanding about the research topic. The summary of literature reviews is shown in Table 1 below.

74

Table 1. Summary of Literature Review

Publications	Title	Types of Data Used	Machine Learning / Deep Learning Algorithm	Findings
Lin et al., 2020 [8]	Classifying BRCA Subtypes Using Deep Neural Networks Based on Multi-Omics Data	• mRNA • DNA methylation • Copy Number Variation (CNV)	• Deep Neural Network (DNN)	• Multi-omics dataset outperforms compared to single omics data.
Liu et al., 2017 [9]	Tumor gene expression data classification via sample expansion-based deep learning	• Tumor gene expression data of BRCA, leukemia and colon cancer	• Sample Expansion-Based SAE (SESAE) (CNV) • Sample Expansion-Based 1DCNN (SE1DCNN)	• SE1DCNN outperforms all the competitive classifiers for all three datasets. Except for the proposed SE1DCNN and SESAE, 1DCNN outperforms the other methods.
Wu and Hicks, 2021 [10]	BRCA Type Classification Using Machine Learning	• RNA Sequence data	• SVM • KNN • Naïve Bayes • Decision tree	• SVM shows more efficient compared to other 3 algorithms
Yang et al., 2024 [11]	Comparative Evaluation of Machine Learning Models for Subtyping Triple-Negative BRCA: A Deep Learning-Based Multi-Omics Data Integration Approach	• mRNA • miRNA • Gene mutations • DNA methylation • Magnetic resonance imaging (MRI) images	• SE-ResNet101	• The model using multi-omics data achieves 80% of accuracy.
Tao et al., 2019 [12]	Classifying BRCA Subtypes Using Multiple Kernel Learning Based on Omics Data	• mRNA • DNA methylation • Copy Number Variations (CNV)	• Multiple Kernel Learning (MKL)	• Accuracy of multiple omics data higher than single omics data.
El-Nabaway et al., 2021 [13]	A Cascade Deep Forest Model for BRCA Subtype Classification Using Multi-Omics Data	• Clinical data • Gene expression • Copy Number Variations (CNV) • Copy Number Abbreviations (CNA)	• Deep Forest	• Integration of omics does not lead to improvement of result (56.7% accuracy). • The model works well with gene expression data

				alone (77.55% accuracy).
Azmi et al., 2022 [14]	Comparative Analysis of Deep Learning Algorithm for Cancer Classification using Multi-omics Feature Selection	• Gene Expression • DNA Methylation • miRNA expression • Clinical data	• Stacked Denoising Autoencoder (SDAE) • Variational Autoencoder (VAE)	• VAE outperforms SDAE with 91.86% accuracy, 22.73% specificity and 0.21% Matthews Correlation Coefficient (MCC).

3. Materials and Methods

The experimental framework of this study was created to help outline the methodology and procedures for the project. A well-considered experiment framework can help the project remain on course and accomplish its objectives. Figure 1 below illustrates the experimental framework of this study.

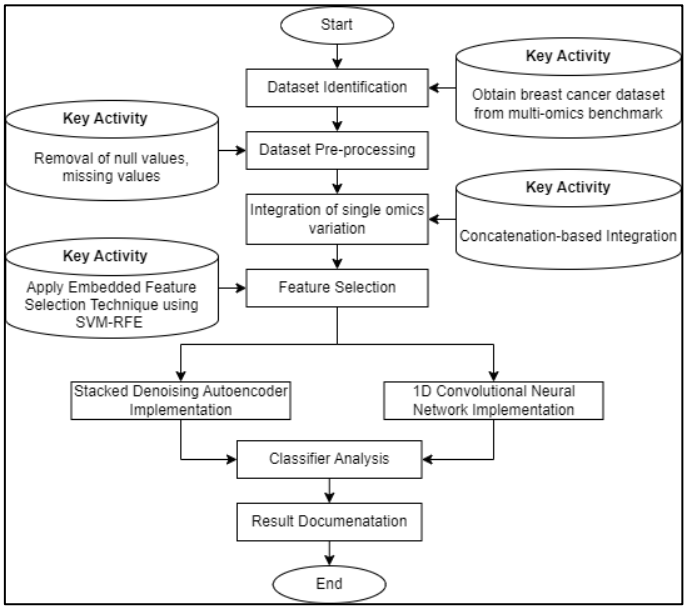

Figure 1. Experimental Framework

3.1. Omics Dataset

The datasets collected in this study are the pre-processed BRCA (BRCA) datasets extracted from TCGA Genome. All three sets of data are from Transcriptomics. Table 2 displays the summary of the BRCA datasets.

Table 2. Summary of BRCA datasets

Types of data	Number of Patients	Number of Features
Copy Number Variation (CNV)	671	19568
DNA methylation	671	19049
miRNA expression	671	368

3.2 Data Preprocessing

Data preprocessing is essential in machine learning or data analysis in reducing noise in datasets. An effective preprocessing can ensure the validity and precision of the data as well as paving the way for insightful analysis and possible clinical applications. The collected datasets are pre-processed, though the missing values and duplicate values are rechecked. To utilize the complete dataset, the omics features data and label data are combined by transposing the features data to match with label. The original nominal type of class label is replaced by numerical data. The instance counts for each class are assessed to determine the balance among classes. Table 3 shows the summary of class labels.

Table 3. Summary of Class Labels

Classes	Labels	Counts
Normal	0	31
LumA	1	353
LumB	2	132
Basal	3	113
Her2	4	42

In this study, BRCA subtype classification involves integrating three omics datasets. Normalization ensures the data are in comparable scale that facilitates the integration and enables more comprehensive analysis [15]. The omics datasets are normalized with Min-max Normalization that transforms data into a common scale. Followed by normalization, the pre-processed omics data are split into training and testing sets with the ratio of 70:30 respectively.

3.3 Omics Data Integration

To preserve data consistency and maintain the biological variability included in the omics dataset, data integration needs to be handled carefully. [16]. To implement data integration of omics datasets, the integration strategy of this study is Concatenation-based Integration (CBI). The CNV, DNA methylation and miRNA expression datasets are combined. This phase involving gene expression matrices alignment from different datasets. This is to ensure that the same genes are present in each dataset, allowing for seamless integration. The datasets are concatenated along the sample dimension by combining the expression values for each gene across all samples from different datasets into a single matrix [17]. For the BRCA subtype classification with integrated dataset, transposing and preprocessing steps are done after the concatenation of datasets. With data integration, the number of features changes. Table 4 below demonstrates the summary of integrated dataset.

Table 4. Summary of integrated BRCA datasets

Types of data	Number of Patients	Number of Features
CNV + DNA methylation + miRNA	671	38986

3.4 Feature Selection

BRCA omics datasets contain massive amounts of data that consume a lot of computational time and storage. By selecting the most significant features, feature selection reduces the dimensionality of the omics datasets. High dimensionality datasets are prone to overfitting. The model will learn noise in the training data rather than underlying problem [18]. Therefore, feature selection plays a critical role in optimizing computing efficiency, decreasing overfitting, and enhancing model performance.

In this study, Support Vector Machine – Recursive Feature Elimination (SVM-RFE) are selected as the method to perform feature selection process. This method is chosen as SVM-RFE has been proved to be effective in enhancing model's accuracy [19]. The SVM aims to find the hyperplane that best separates the classes with largest margin. SVM-RFE is an iterative method that recursively eliminates the least important features according to the weights of the features derived by the SVM model [20]. The feature with the lowest weight indicates the feature is least important. The iteration is repeated recursively until desired number of features is reached. The formula of SVM weight vector (w) is shown as below [21]:

$$w = \sum_{i=1}^{n} \alpha_i Y_i X_i \qquad (1)$$

Where:

- n is the number of training samples
- α_i is the Lagrange multipliers obtained from solving dual optimization problem
- Y_i is the class labels
- X_i is support vectors

In this study, the number of features desired to be selected is different for each omics data. The desired number of features for each type of omics is shown in Table 5.

Table 5. Summary of BRCA datasets

Types of data	Number of Features	Desired Number of Features
Copy Number Variation (CNV)	19568	19000
DNA methylation	19049	19000
miRNA expression	368	300
CNV + DNA methylation + miRNA	38985	38000

To implement SVM-FRE, these steps are carried out in this study to find the important features. Figure 2 below shows the pseudo code of the feature selection using SVM-RFE.

```
Input:
  - Dataset X with features (genes) X1, X2, ..., Xn
  - Labels y
  - Number of features to select k

Output:
  - Selected features (genes)

Procedure:
1. Initialize F = {X1, X2, ..., Xn}  # Start with all features
2. While |F| > k do:
    a. Train SVM model on the dataset with features in F
    b. Compute the weight vector w from the trained SVM model
    c. Calculate the ranking criterion for each feature i in F:
       R_i = (w_i)^2  # Importance of feature i
    d. Identify the feature with the smallest ranking criterion:
       f_min = argmin_{i in F} R_i
    e. Remove f_min from F:
       F = F \ {f_min}
3. Return the remaining features in F

End Procedure
```

Figure 2. Pseudo Code for Feature Selection using SVM-RFE

3.5 Synthetic Minority Oversampling Technique (SMOTE)

The omics datasets show problems with class imbalance. Underrepresentation of some classes in medical datasets, such as BRCA subtype classification, is a common problem. Synthetic Minority Oversampling Technique (SMOTE) is a technique that is commonly used in class balancing to prevent the model from being biased in favor of majority classes [22]. SMOTE is applied to prevent classifier bias. A simple concept of SMOTE is that it will generate synthetic samples for the minority class to increase the number of minority samples and further lead to a balanced class distribution. The K-Nearest Neighbors, K is determined as five. Hence, for each sample in the minority class, SMOTE will find five of its nearest neighbors. Then, one of the nearest neighbors will be selected. The difference between the selected minority sample and the selected nearest neighbor will be computed and lead to the formation of a synthetic minority sample. The sampling strategy is set to auto where it will resample only minority class. Figure 3 shows the flowchart diagram of SMOTE oversampling strategy.

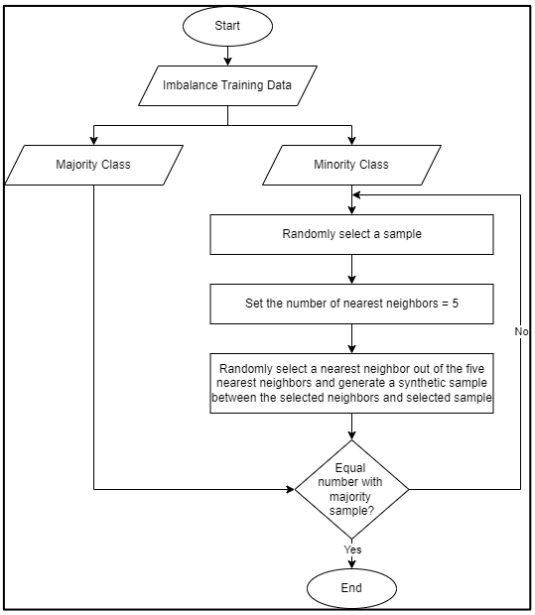

Figure 3. SMOTE Oversampling Strategy

3.6 Stacked Denoising Autoencoder (SDAE) and 1D Convolutional Neural Network (1D CNN)

The primary focus of this study is the implementation of two deep learning algorithms, which are Stacked Denoising Autoencoder (SDAE) and 1D Convolutional Neural Network (1D CNN) to classify the BRCA subtypes. To facilitate early BRCA subtypes diagnosis, deep learning models are created to analyze the importance of omics data in identifying the BRCA patients. SDAE is designed to extract reliable features from corrupted input data. It extends the basic autoencoder by stacking multiple layers of encoders and decoders as well as by incorporating noise into the input data during training. SDAE is trained in layer-wise function while each layer of the autoencoder is trained to reconstruct its input from a corrupted version, from input layer to deeper layer [23]. In this study, SDAE model with a total of 7 layers with dimensions of the number of input features, 256, 128, 100, 128, 256 and number of input features respectively. The model is compiled using the Adam optimizer with a learning rate of 0.001 and the mean squared error loss function. Then, SDAE model is built sequentially, with an early stopping with patience rate of 20 and loss rate monitoring. The DAE model stops training when the monitored metric has stopped improving. After training, the encoder part of the SDAE is used to extract features from both training and test data. Figure 4 illustrates the SDAE architecture used in this study.

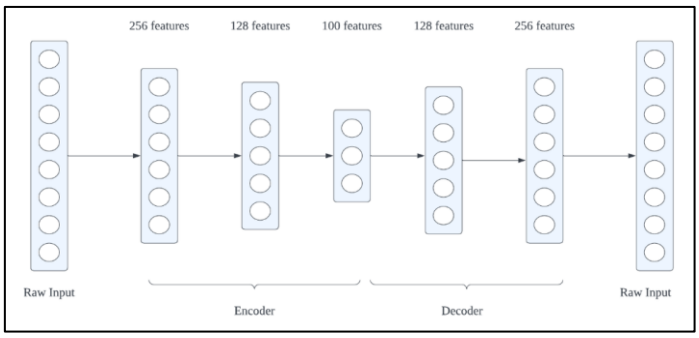

Figure 4. SDAE Architecture

1D CNN is a deep learning algorithm that is designed to handle sequential data. To discover underlying patterns and features in the dataset, it applies filters along one dimension using convolutional layers. This algorithm can perform effectively in tasks including classification, regression and anomaly detection task [24]. In this study, Keras Sequential API is used to define the CNN architecture. The model begins with a layer of 32 filters, kernel size of 3 and Rectified Linear Unit (ReLU) activation function, which processing input data with size equals to features number. Another layer with pool size of 2 is created. Furthermore, two convolutional and max-pooling layers are added. They have 64 and 128 filters respectively. To prepare it for the dense layers, the output is flattened into a one-dimensional vector. Then, a dense layer with 128 units and ReLU activation function is created, follow by another layer with a 0.5 rate to prevent overfitting. The last layer is a dense layer with five units and a softmax activation mechanism. Lastly, the model is assembled with the sparse categorical cross-entropy loss function and the Adam optimizer. The SDAE model was paired with a MLP deep learning classifier, while the 1D CNN model performed classification independently. Figure 5 shows the example of 1D CNN architecture based on 19000 features.

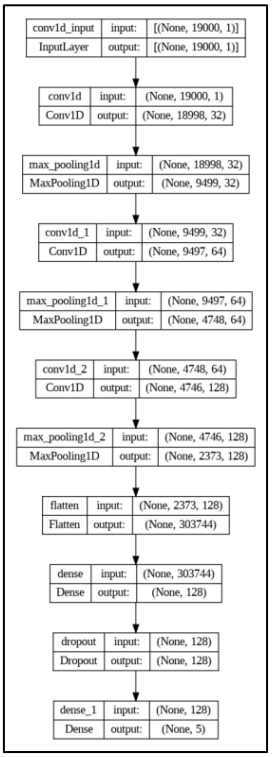

Figure 5. Architecture of 1D CNN

3.7 Analysis

To avoid overfitting, the BRCA datasets are split into training and testing set before fitting into the deep learning models. Same goes for the integrated multi-omics BRCA data was separated into two different sets. To evaluate the effectiveness of SDAE and 1D CNN, performance metric was required to analyse the models' performance. Accuracy is the performance metric of SDAE and 1D CNN models in this study. It refers to the percentage of correct value to all other instances and can be determined by comparing the projected value to the actual value of class label. Hence, accuracy of each model with various type of omics are computed and compared.

4. Results

To identify the best-performing model in identifying BRCA subtypes based on different types of omics data, SDAE and 1D CNN models were developed and compared. The performance of the models was also evaluated with and without using oversampling technique. The SDAE model was paired with a MLP deep learning classifier, while the 1D CNN model performed classification independently. The performance of these models was then assessed to determine which one better supports the objectives. Tables 6 and 7 show the comparative analysis of SDAE and 1D CNN models based on different types of omics data as well as the implementation of SMOTE.

Table 6. Accuracy of SDAE model based on types of omics

Model	Type of Omics	Accuracy (%)	
		Without SMOTE	With SMOTE
SDAE	CNV	70.79	71.78
	DNA methylation	76.24	76.73
	miRNA	74.75	79.21
	CNV + DNA methylation + miRNA	70.30	71.29

Table 7. Accuracy of 1D CNN model based on types of omics

Model	Type of Omics	Accuracy (%)	
		Without SMOTE	With SMOTE
1D CNN	CNV	68.81	70.79
	DNA methylation	79.70	81.19
	miRNA	76.73	80.20
	CNV + DNA methylation + miRNA	77.23	80.20

5. Discussion

The provided Tables 6 and 7 illustrate the accuracy of two models, SDAE and 1D CNN, on different types of omics data with and without the application of SMOTE. In this context, SDAE was used for feature extraction and combined with a MLP algorithm for classification purposes, while 1D CNN performed classification directly. Both models employed the datasets under the same data preprocessing steps and feature selection technique. The tables have revealed differences in the model performance and their responses to SMOTE across various types of omics data.

Looking into the results of the SDAE model as stated in Table 1, it is observed that the classification accuracy improves with the implementation of SMOTE across all the omics data types. For CNV data, the accuracy increases from 70.79% to 71.78%, while for DNA methylation data, it rises slightly from 76.24% to 76.73%. The most significant improvement is shown with miRNA data, where the accuracy increases from 74.75% to 79.21%. On the other hand, for the integrated CNV, DNA methylation and miRNA data, the SDAE model's accuracy goes from 70.30% to 71.29%. These results indicate that while SMOTE can generally improve the performance of the model, it does not really increase much except for miRNA dataset. Other than that, the result also shows that the integration of different omics data types does not seem to produce the best results.

The 1D CNN model shows a more notable improvement after applying SMOTE across all omics data types, indicating that the 1D CNN model is more sensitive to class imbalance.

Without applying SMOTE, the 1D CNN model obtains an accuracy of 68.81% for CNV data, and it increases to 70.79% with SMOTE. Meanwhile, for DNA methylation data, accuracy rises from 79.70% to 81.19%, and for miRNA data, it improves from 76.73% to 80.20%. The integrated CNV, DNA methylation, and miRNA data shows the greatest improvement after applying SMOTE, with accuracy increasing from 77.23% to 80.20%. This shows that the 1D CNN model not only performs better than the SDAE model in most cases but also benefits more from the application of SMOTE. However, similar to the SDAE model, using a combination of omics dataset does not produce the best results.

When comparing the performance of both models, the 1D CNN model outperforms the SDAE model most of the time in terms of accuracy, except for the CNV dataset. Here, we compare only the best result of each data type after applying SMOTE. For CNV data, the SDAE model (71.78%) achieves a slightly higher accuracy than the 1D CNN model (70.79%). Other than that, the 1D CNN model performs better than the SDAE model when using DNA methylation data (81.19% vs. 76.73%), miRNA data (80.20% vs. 79.21%) and combined omics data (80.20% vs. 77.23%). Among all different types of omics, the best-performing model is the 1D CNN using DNA methylation dataset with the application of SMOTE, which has achieved the highest accuracy of 81.19% in classifying the BRCA subtypes.

Since we observed inconsistent result when SDAE outperformed 1D CNN only when using CNV dataset, we conducted an additional experiment by reducing the feature set to 75%, resulting in 14677 features for both SDAE and 1D CNN, specifically on the CNV dataset with the application of SMOTE. The result is as shown in Table 8.

Table 8. Accuracy of SDAE and 1D CNN Models using 75% Features

Model	Type of Omics	Accuracy (%)
SDAE	CNV	66.83
1D CNN		72.27

From the result in Table 8, it is observed that the accuracy of SDAE model decreased from 71.78% to 66.83%, while the accuracy of 1D CNN model improved from 70.79% to 72.27%. Different from the previous result where 19000 features are used, 1D CNN outperforms SDAE when using 14677 features, indicating that 1D CNN works better than SDAE using a smaller number of features for CNV dataset. This shows similar trend with the previous result for the other type of omics and indicates that the models' performance is sensitive to the number of features selected from feature selection.

Overall, the 1D CNN model outperforms the SDAE model in most cases. However, it is still dependent on the architecture of the deep learning model used. Other than that, the application of SMOTE has consistently improved the classification performance across all cases. The 1D CNN model seems to be more sensitive to class imbalance as it shows greater improvement in classification accuracy after applying SMOTE. Furthermore, the integrated variations of omics do not produce the best result for both models. Specifically, the SDAE

model works best with miRNA data while the 1D CNN model obtains the best result with DNA methylation data.

6. Conclusions

In conclusion, both 1D CNN model and SDAE model can perform better after applying SMOTE means while SMOTE can help to improve the model performances by solving the class imbalance problem. However, when comparing the two models, the 1D CNN model generally achieves higher accuracy than the SDAE model, except using the CNV dataset. In this context, the number of selected features could be one of the factors that affect the models' accuracy since 1D CNN model works better on a reduced feature set of CNV dataset. The best-performed model is 1D CNN using DNA methylation data with the application of SMOTE, achieving an accuracy of 81.19% for the BRCA subtypes classification. Other than that, integrated omics data does not produce the best result. It may be due to the redundancy occurring since many types of omics data are used which can cause duplicate information, in order to increase the complexity to analyze the integrated omics data [25]. Therefore, further studies need to be carried out which focus on the model development and techniques to more effectively manage and utilize integrated omics data for the enhancement of accuracy and reliability of BRCA subtype classification.

Author Contributions: Conceptualization, methodology, software, validation, formal analysis, investigation, writing—original draft preparation, writing—review and editing, CKW, HYQ, MZQ, ZT.

Funding: No external funding was provided for this research.

Acknowledgments: In this segment, you may acknowledge any support that is not addressed by the author's contribution or funding sections.

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

References

1. Trayes, K. P., & Cokenakes, S. E. (2021). BRCA treatment. American family physician, 104(2), 171-178.
2. Orrantia-Borunda, E., Anchondo-Nuñez, P., Acuña-Aguilar, L. E., Gómez-Valles, F. O., & Ramírez-Valdespino, C. A. (2022). Subtypes of BRCA. BRCA [Internet].
3. Lin, Y., Zhang, W., Cao, H., Li, G., & Du, W. (2020). Classifying BRCA subtypes using deep neural networks based on multi-omics data. Genes, 11(8), 888.
4. Menyhárt, O., & Győrffy, B. (2021). Multi-omics approaches in cancer research with applications in tumor subtyping, prognosis, and diagnosis. Computational and structural biotechnology journal, 19, 949-960.
5. Liu, P., Zheng, P., & Chen, Z. (2019). Deep learning with stacked denoising auto-encoder for short-term electric load forecasting. Energies, 12(12), 2445.
6. Leng, D., Zheng, L., Wen, Y., Zhang, Y., Wu, L., Wang, J., ... & Bo, X. (2022). A benchmark study of deep learning-based multi-omics data fusion methods for cancer. Genome biology, 23(1), 171.
7. Elreedy, D., & Atiya, A. F. (2019). A comprehensive analysis of synthetic minority oversampling technique (SMOTE) for handling class imbalance. Information Sciences, 505, 32-64.
8. Lin, Y., Zhang, W., Cao, H., Li, G., & Du, W. (2020). Classifying BRCA subtypes using deep neural networks based on multi-omics data. Genes, 11(8), 888.

9. 9Liu, J., Wang, X., Cheng, Y., & Zhang, L. (2017). Tumor gene expression data classification via sample expansion-based Deep Learning. Oncotarget, 8(65), 109646–109660. https://doi.org/10.18632/oncotarget.22762

10. Wu, J., & Hicks, C. (2021). BRCA type classification using machine learning. Journal of Personalized Medicine, 11(2), 61. https://doi.org/10.3390/jpm11020061

11. Yang, S., Wang, Z., Wang, C., Li, C., & Wang, B. (2024). Comparative evaluation of machine learning models for subtyping triple-negative BRCA: A deep learning-based multi-omics data integration approach. Journal of Cancer, 15(12), 3943–3957. https://doi.org/10.7150/jca.93215

12. Tao, M., Song, T., Du, W., Han, S., Zuo, C., Li, Y., Wang, Y., & Yang, Z. (2019). Classifying BRCA subtypes using multiple kernel learning based on OMICS DATA. Genes, 10(3), 200. https://doi.org/10.3390/genes10030200

13. El-Nabawy, A., Belal, N. A., & El-Bendary, N. (2021). A Cascade Deep Forest model for BRCA subtype classification using Multi-Omics Data. Mathematics, 9(13), 1574. https://doi.org/10.3390/math9131574

14. Azmi, N. S., A Samah, A., Sirgunan, V., Ali Shah, Z., Abdul Majid, H., Howe, C. W., Wen, N. H., & Azman, N. S. (2022). Comparative analysis of Deep Learning Algorithm for cancer classification using multi-omics feature selection. Progress In Microbes & Molecular Biology, 5(1). https://doi.org/10.36877/pmmb.a0000278

15. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321-357.

16. Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research, 43(7), e47.

17. Huber, W., Carey, V. J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B. S., ... & Morgan, M. (2015). Orchestrating high-throughput genomic analysis with Bioconductor. Nature Methods, 12(2), 115-121.

18. Taminau, J., Meganck, S., Lazar, C., Steenhoff, D., Coletta, A., Molter, C., ... & Nowé, A. (2014). Unlocking the potential of publicly available microarray data using inSilicoDb and inSilicoMerging R/Bioconductor packages. BMC Bioinformatics, 15(1), 335.

19. Saeys, Y., Inza, I., & Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. Bioinformatics, 23(19), 2507-2517.

20. Guo, J., Wang, K., & Jin, S. (2022). Mapping of Soil pH Based on SVM-RFE Feature Selection Algorithm. Agronomy, 12(11), 2742. https://doi.org/10.3390/agronomy12112742.

21. Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. Machine Learning, 46(1-3), 389-422.

22. Cristianini, N., & Shawe-Taylor, J. (2000). An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge University Press.

23. Bengio, Y., Lamblin, P., Popovici, D., & Larochelle, H. (2007). Greedy Layer-Wise Training of Deep Networks. Advances in Neural Information Processing Systems, 19, 153-160.

24. Al-Jarrah, O. Y., Yoo, P. D., Muhaidat, S., Taha, K., & Karagiannidis, G. K. (2015). Efficient Machine Learning for Big Data: A Review. Big Data Research, 2(3), 87-93.

25. Duan, R., Gao, L., Gao, Y., Hu, Y., Xu, H., Huang, M., Song, K., Wang, H., Dong, Y., Jiang, C., Zhang, C., & Jia, S. (2021). Evaluation and comparison of multi-omics data integration methods for cancer subtyping. PLOS Computational Biology/PLoS Computational Biology, 17(8), e1009224.

362