Progress in Microbes and Molecular Biology

HH PUBLISHER

Original Research Article

Classification of Breast Cancer Subtypes using Transcriptomic Data with SDAE and DBN

Chang Min Xuan [1], Hanis Rafiqah Hisham Razuli [1], Lee Jia Yee [1], Nik Syahdina Zulaikha Badrul Hisham [1]

Article History

Received: 16 July 2024;

Received in Revised Form: 16 July 2024;

Accepted: 16 July 2024;

Available Online: 16 July 2024;

[1]Faculty of Computing, Universiti Teknologi Malaysia, 81310 Johor Bahru, Johor, Malaysia; changxuan@graduate.utm.my(CMX); hanisrafiqah@graduate.utm.my(HR); yee00@graduate.utm.my(LJY); niksyahdinazulaikha@graduate.utm.my(NSZ)

Abstract: BRCA is a heterogeneous disease entity for which accurate subtype classification is indispensable for effective treatment. The present research focuses on improving BRCA subtype classification through the integration of single-omic variation data, in particular, Copy Number Variation, microRNA, and mRNA. Integration of data at the single-omic variation level would allow more sophisticated analyses to be carried out based on features of several aspects of biology. However, large-scale cancer single-omic variation data analysis may be challenging due to high dimensionality. We applied Support Vector Machine-Recursive Feature Elimination (SVM-RFE) for the selection of the most informative features. In our study, we have used Synthetic Minority Over-sampling Technique (SMOTE), which creates synthetic samples for the minority class to address class imbalance. Furthermore, we will explore Stacked Denoising Autoencoder (SDAE) and Deep Belief Networks (DBN) in their ability for the discovery of latent patterns within the single-omic variation landscape of breast cancer. Our results indicated that the inclusion of SDAE and DBN, together with SVM-RFE and SMOTE, had empowered this extremely complex problem of BRCA. Results showed that the overall accuracy using DBN model is much higher compared to accuracy SDAE such as DBN with SMOTE 83% while SDAE with SMOTE 75% .

Keywords: Breast Cancer; Single-omic variation; Deep Learning; Support Vector Machine-Recursive Feature Elimination; Stacked Denoising Autoencoder (SDAE); Deep Belief Networks (DBN)

1. Introduction

Breast cancer forms the single most prevalent type of cancer in women worldwide and rears its evil head once every year in millions. In short, breast cancer results from abnormal growth and proliferation of cells inside the tissue of the breast. Unless checked, these cells can migrate to other parts of the body. The risk factors are many for the disease, including gender, age, heredity, and several lifestyle aspects such as obesity and consuming alcohol. The symptoms of breast cancer may be anywhere from nothing at all to something very noticeable. Accordingly, regular examination of the breasts and screening tests such as mammograms are done for early detection, which paves the path for effective treatment. Adequate treatment would depend on combining all the different modes of surgery, radiotherapy, chemotherapy, hormone therapy, and targeted drug therapies. Breast cancer is one of the most common types of cancer in women, with nearly one woman out of eight suffering from this disease at some point in her lifetime. On the other hand, it is relatively a rare cancer in men[1].

Breast cancer is estimated to have around 2.3 million new cases diagnosed in 2020, making it one of the most frequently diagnosed cancers in women worldwide[2]. Such primary subtypes of breast cancer can be further classified based on the absence or presence of specific receptors: estrogen receptor, progesterone receptor, and human epidermal growth factor receptor 2. These include ER-positive/PR-positive, which is the most frequent, HER2-positive, triple negative, and lastly, Luminal A and B, that further sub-classify ERpositive cancers. All these subtypes are characterized by different characteristics, prognosis, and outlines of treatment. Accurately identifying the breast cancer subtype may help guide personalized therapeutic strategies for improving patient outcomes. Luminal-A is the most common type of breast cancer; it expresses at high levels of hormone receptors, such as estrogen and progesterone[3]. The prognosis in the Luminal B1 subtype is poorer than in the Luminal A subtype due to a much higher cellular proliferation rate[4]. HER2 is a human epidermal growth factor receptor with tyrosine kinase activity. The contribution of HER2 to inducing breast cancer can be through gene amplification, mutation, or overexpression of the receptor[5].

A Stacked Denoising Autoencoder (SDAE) is an enhanced model of an artificial neural network that succeeded the basic working concepts of autoencoders, specifically designed to improve the quality of feature learning by introducing noise in the training. On the other hand, an auto-encoder is a neural network trained with a method to learn the copying of their input to their output, through an encoding process followed by decoding. Key to this is SDAE's innovation, which is, denoising auto-encoders are useful in learning more robust features by reconstructing data from its corrupted version. The SDAE consists of multiple layers of denoising auto-encoders stacked on top of each other. Therefore, during the training process, each layer gets trained to denoise the output created by the one before it, resulting in this hierarchical representation of the data. There are two major stages: pretraining and fine-tuning. At the stage of pretraining, every autoencoder layer should be trained independently to minimize reconstruction error of its noisy input. In fine-tuning, the whole network is trained together in a supervised way; optimization is often done according

to the labeled dataset in order to gain expertise in one particular class of tasks, for example, classification. Due to the random noise added in training, the learning of how to reconstruct the original uncorrupted data has a consequence: this forces the network to be more easily capturable with features that are invariant to noise. The approach of SDAE is fairly insensitive and reasonably robust to noisy input data in the pre-training phase[6]. This process improves generalization to new, unseen data by learning to focus on the underlying pattern rather than the noise. Domains to which the Stacked Denoising Autoencoder has recently been applied, which includes image denoising[7], human activity recognition[8], and feature representation[9].

Deep Belief Networks (DBN) are part of the architecture of deep neural networks that have drawn a lot of attention in the field of machine learning and artificial intelligence. DBN is one of the efficient ways to overcome the problems encountered in deep neural networks, such as slow convergence and overfitting during the training process, is through the method discussed[10]. DBNs are multi-layered, consisting of stochastic, latent variables known as hidden units that are used to model probability distributions over input data. Their outstanding innovation was in being able to learn the underlying features of data in an unsupervised manner through the use of layerwise training of the network, thereby affecting the reconstruction of input. This unsupervised pre-training phase of DBNs in fact makes it possible to encapsulate the intrinsic structure of data, which can be further fine-tuned for specified tasks following supervised learning. Specifically, the hierarchical structure of DBNs allows for complex, high-dimensional data and therefore finds great application in domains like image recognition[11–12], natural language processing[13], and feature extraction[14]. Through the exploitation of deep learning in generating meaningful representations from unlabeled data, DBNs have been known for showing their success on a plethora of other machine learning tasks where the task successes lead to the recent improvements in computer vision, speech recognition, predictive analytics, and many other areas.

SDAE and DBN are deep models that can be used in classifying the different subtypes of breast cancer using gene expression data because these techniques automatically find important patterns within the gene expression data. An SDAE learns by reconstructing the given gene expression data, whereby it gets to understand the structure of the data, which it could use in classifying the breast cancer subtypes. Successful applications of Stacked Denoising Autoencoders in various domains make them very promising for enhancing cancer detection, as they improve classification performance and extract linear and nonlinear relationships involved, with the identification of relevant gene subsets as potential cancer biomarkers[15]. Whereas the DBN also learns the structure of gene expression data, it does this differently. It builds a multi-layer model that will let it recognize the different subtypes of breast cancer. Both SDAE and DBN are good at working with complex gene expression data and outperform traditional methods at classifying breast cancer subtypes. This helps doctors in forming more reliable diagnoses and providing personalized treatment to breast cancer patients.

2. Materials and Methods

Method is depicted using a framework which encompasses the whole research process conducted. It is significant to provide a brief notion about research for a further and deeper understanding of the applied methodologies. Figure 1 portrays the experimental framework of this research.

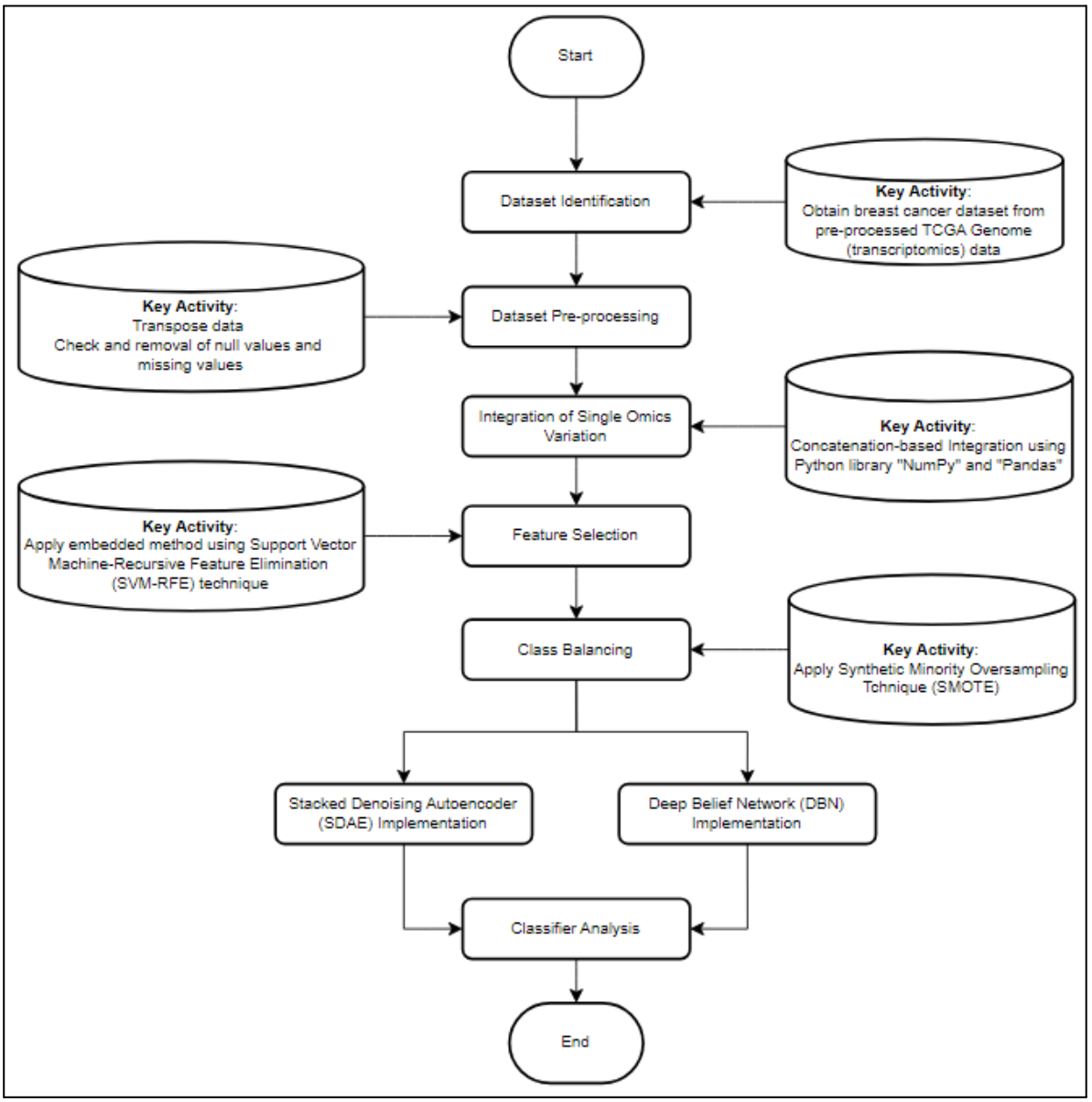

Figure 1. Experimental Framework

2.1. Omics Dataset

This study employed the breast cancer dataset which sourced from the pre-processed TCGA Genome data repository which is publically accessible. The dataset can be downloaded in the google drive sharable link from https://drive.google.com/drive/folders/1i_prPkoYj_fZtvnQuzKLculUeqmWbUb_?usp=sharing. The summary of breast cancer data is displayed in Table 1.

Table 1. Summary of Breast cancer data.

Types of dataset	Omics Field	Number of patients	Number of features
miRNA Expression	Transcriptomics	672	368
mRNA Expression	Transcriptomics	672	18206
Copy Number Variation	Transcriptomics	672	19568
Survival Data (Clinical)	-	672	1

2.2. Data Pre-Processing

To minimize useless biases and noise, data pre-processing is crucial for integrative analyses toward the research of single omics variation. The omics dataset that was obtained has already undergone pre-processing. Data transposition is carried out to ensure the samples are located in row and features in column form. After that, this step also involves checking raw data and data cleaning to eliminate duplicates and missing values to ensure that the following data analysis would not be affected by noise data. The breast cancer survival data consists of one feature which is class label.There are four distinct subtype classes which are labeled as LumA, LumB, Her2 and Basal as well as a normal class. The patients with LumA and LumB are marked as 0 and 1 whereas patients who suffer with Her2 and Basal cancer subtypes are labeled with 2 and 3 respectively. The patient samples which have label 4 represent normal patients without breast cancer. According to Table 2, the first class count revealed that 353 patients (52.53%) were under LumA subtype while 132 patients (19.64%) were under LumB class. Furthermore, the Her2 and Basal classes had 42 (6.25%) and 113 (16.82%) patients respectively. There are solely 26 patients which is 3.87% for the normal class.

Table 2. Details of classes in survival data (Clinical).

Data Type	Class				
Survival Data (Clinical)	LumA	LumB	Her2	Basal	Normal
Number of samples	353	132	42	113	31

The 80% split of the pre-processed omics data was utilized for training while the remaining 20% was applied for testing. In addition, the min-max normalization approach was used to execute data normalization during this phase. Scaling each feature into the range of 0 to 1 was the main objective of the normalization process. These data pre-processing processes have a significant impact on the latter model development and analysis, therefore it becomes essential to implement them appropriately[16].

2.3. Integration of Single Omics Variation

In single omics variation research, data integration techniques are pivotal and challenging steps to be executed. It is a process of merging data produced by a range of

research methodologies to uncover the underlying complexity of biological phenomena and information particularly the disease, breast cancer which has several subtypes. Early integration was the integration approach adopted in this study to integrate the data from the single omics variation. This strategy was used following the data pre-processing step. Subsequently, the pre-processed data were concatenated and generated an enormous matrix result which was then used to train the model. Concatenated-based integration is frequently applied because it is straightforward, simple to execute and most importantly, it can combine features from each omics to show the consequences of the interactions between multiple layers[17]. Table 3 concludes the number of samples and features after data integration.

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

Types of dataset	Number of samples	Number of features
Single omics variation	672	30000

To concatenate the three omics data in relation to one another, the Python package Pandas has been utilized. A complete concatenated data including samples with all the relevant omics was constructed. "Sample" which denotes patient ID has been employed as the index or target data and only the samples which are available at all three data types were qualified for the further analysis.

2.4. Feature Selection

In the process of preparing data for modeling, feature selection is one of the principal phases. By using only pertinent data and eliminating data noise, feature selection is a technique for decreasing the input variables to the classification model. Since the support vector machine-recursive feature elimination (SVM-RFE) can handle data with high dimensionality and unbalanced class, it is opted for selecting features in this research.

SVM-RFE often is referred to as an embedded method which incorporates a feature ranking criterion into the SVM training process and iteratively omits the lowest ranked features until a predetermined number of features is reached[18]. The algorithm first imported the pre-processed data and trained the SVM classifier. Next, the ranking rules were computed via weight magnitude. The successive process was feature discovery with the smallest ranking criterion as well as feature ranked list updating, followed by removing the features which had the lowest ranking through RFE step. The feature's ranking score, f_l, was indicated as below:

$$Rank_1(f_l) = (w_l)^2 = \left(\sum_{i \in SV_s} a_i y_i x_{il} \right)^2 \quad (1)$$

Where SVs represent the collection of support vectors and the Lagrangian multipliers,α_i of these vectors are non-zero for SVM. In this research, SVM-RFE was applied to produce a list of the selected features with the highest ranking as the input for the following classification procedure to ensure that the model accuracy can be improved.

Table 4. Implementation of SVM-RFE method for feature selection.

Types of dataset	Before SVM-RFE	After SVM-RFE	Number of removed features
miRNA Expression	368	250	118 (32.07%)
mRNA Expression	18206	5000	13206 (72.54%)
Copy Number Variation	19568	5000	14568 (74.45%)
Total features	38142	30000	8142 (21.35%)

2.5. SMOTE

The class imbalance concerns were also tracked during the data preparation process. When the minority class has less distribution than the majority class, the poor prediction accuracy rose from models' tendency to classify new data into the majority class[19]. Hence, SMOTE, which is an oversampling method, was suggested to tackle the unbalanced class issue by increasing the sample size of the minority class in the training dataset merely to prevent the other problems such as low generalization and under-sampling[20].

The process was initiated by calculating the Euclidean distance between each selected sample and every sample in the minority class to attain its k nearest neighbors. The following step involved computing a sampling ratio based on the sample class imbalance in order to estimate the sampling magnification N. Multiple samples were randomly selected from each minority class's k neighbors via selected neighbors. After that, a new sample was synthesized from the original sample for each randomly selected neighbor[21]. The new sample was created using the equation below:

$$x_{new} = x + \lambda \times |x + x_n| \quad (2)$$

Where x_{new} denotes the newly created sample; λ denotes the random number between 0 and 1; x denotes the minority sample; x_n is the randomly chosen neighbor. These steps repeated to find the next nearest neighbor until the class distribution achieved a balanced condition.

2.6. Classifier Analysis Deep Learning Models

The accuracy of SDAE and DBN models in classifying the breast cancer subtypes was evaluated during both the training and testing phases. The model's capability to correctly categorize the omics features into 4 distinct cancer classes and 1 normal class was assessed

and the results were thoroughly analyzed. To prevent overfitting, each dataset was divided into training and testing data to ensure different data inputted into the models. The performance of these two deep learning models was measured solely via accuracy.

2.6.1. Stacked Denoising Autoencoder (SDAE)

The SDAE architecture is constructed with a batch size of 32 and 250 epochs for model training. First, the miRNA data is standardized using StandardScaler. The SDAE model is built with an input layer, two hidden encoding layers, and two hidden decoding layers, and is trained to reconstruct the input data. The input dimension for SDAE was 30000 features and the output dimension was 30000. It employed the Rectified Linear Unit (ReLU) and linear activation functions as well as the adaptive moment estimation (ADAM) optimizer. Subsequently, the SDAE with two layers, having dimensions of 64 and 32 was built using similar parameters as the SDAE mentioned previously despite the dimension for the output layer being set as the class number, 5 with softmax activation function. Finally, the trained classifier predicts class probabilities for the test data, and these probabilities are converted into class labels for evaluation. The model's performance was then observed and analyzed at the discussion stage of the research.

2.6.2. Deep Belief Networks (DBN)

For the DBN model, an input of 30000 features was fed into the input layer according to the provided hidden layers and initialized the learning rates, batch size, activation function and dropout rate. The model is pre-trained by two consecutive hidden layers of 256 neurons using Restricted Boltzmann Machines (RBMs) in an unsupervised manner to learn the compressed input data with the setting of a learning rate, 0.05 for 20 epochs over the entire dataset. After pre-training, the entire network is fine-tuned using supervised learning which involves adjusting the weights of the network to minimize the error that occurs on the training data. The learning rate is set as 0.1 using backpropagation under 200 iterations and meanwhile, the batch size defined is 64 samples and the activation function used is Rectified Linear Unit (ReLU) which helps the model learn more complex patterns. Additionally, the dropout with a probability of 0.2 is applied during the training phase to prevent overfitting. Lastly, the classifier predicts the outcomes using testing data and then, the obtained results are evaluated with the label class from the same testing data to generate DBN model accuracy.

3. Results

The output of the Support Vector Machine-Recursive Feature Elimination (SVM-RFE) algorithm is presented. In addition, Stacked Denoising Autoencoder (SDAE) and Deep Belief Networks (DBN) models have been developed using the selected feature subsets identified by the SVM-RFE method. The classification performance results for these SDAE and DBN models are also reported.

3.1. SVM-RFE

SVM-RFE is employed in this research to select the most relevant features in datasets. By applying SVM-RFE as feature selection, it helps in reducing the dimensionality and focuses on more informative data. To implement SVM-RFE, the SVM model is initialized with a linear kernel. The linear kernel is used because it allows for straightforward computation of feature importance based on the model's coefficients, which is necessary for RFE to rank the features effectively. Meanwhile, C parameter refers to the penalty parameter of the error term in the SVM algorithm. It controls the tradeoff between achieving correct classifications and maintaining the smoothness of the decision boundary. The parameter setting is shown in Table 5.

Table 5. SVM-RFE classification model parameter setting description.

Parameter	Setting / Values
Model	Support Vector Regression (SVR)
Kernel	Linear
C	1.0

Moreover, the step size used is 10, meaning that 10 features are removed at each iteration. The number of features to select varies depending on the dataset size. For example, in miRNA data, which has a smaller sample size, 250 features are selected, while in CNV and mRNA data with larger sample sizes, 5000 features are selected. The selected features are determined by the coefficients of the SVM model with a linear kernel. In the context of feature selection, these coefficients indicate the importance of each feature. The Coefficient of Determination (R^2) is used to evaluate the overall performance of the model, indicating how well the selected features explain the variance in the target variable. Higher R^2 values, which are close to 1, indicate that the model explains a large portion of the variance, thus suggesting that the selected features are effective.

3.2. SMOTE

We applied the Synthetic Minority Over-sampling Technique to avoid class imbalance in our dataset during training. Class imbalance may result in effects like biasing the model toward the majority class and hence poor performance. SMOTE is that technique of oversampling which balances the distribution of classes by creating, synthetically, samples from the minority class. We initialized SMOTE with a random state of 42 for reproducibility. The fitting of the resample is done on our training data. It will involve the identification of minority class instances, then the creation of new synthetic examples by interpolating between each minority instance and its k-nearest neighbors. At last, these synthetic samples will be added to your original dataset. It is then used to train our machine learning models with this balanced training data, for improved performance and reduced bias toward the majority class. SMOTE usage would ensure that our models are robust and able to predict accurately for both the majority and minority classes.

Table 6. Summary of data before and after resampling

Type of class	Number of samples	
	Before	After
Basal	113	282
Lum A	353	282
Lum B	132	282
Her2	42	282
Normal	31	282

There was a lot of imbalance across different subtypes within the dataset as shown in Table 6, Basal had 113 samples, Lum A had 353, Lum B had 132, and only 31 for Normal. In this situation, this class imbalance may be harmful to the performance of a machine learning model since it will lead to bias toward the majority class. In order to overcome this, SMOTE was applied, generating synthetic samples for the minority classes, therefore each class would have 282 samples after resampling. In more detail, Basal class oversampled from 113 to 282 samples, whereas Lum A was under sampled from 353 to 282 samples, Lum B oversampled from 132 to 282, and lastly, Normal class underwent the greatest change, from 31 to 282 samples.

3.3. SDAE

Table 7 shows the performance measurement of SDAE before and after SMOTE presents the big picture of relative accuracies for the SDAE model before and after the application of SMOTE over different omics datasets. Initially, without SMOTE, the SDAE model achieved accuracies of 74% with the miRNA dataset, 72% with the mRNA dataset, and 62% with the Copy Number Variation dataset. These figures present the challenges presented by class imbalance in the original datasets, where minority classes were relatively underrepresented. Major improvement in accuracy was noticeable for all datasets following the application of SMOTE to deal with class imbalance. The accuracy increased to 79% for the miRNA dataset, up to 80% for mRNA, and for the CNV dataset, its accuracy rose to 66% post-SMOTE. Furthermore, in the single-omic variation dataset, the accuracy improved from 64% before SMOTE to 75% after SMOTE.

Table 7. Performance measurement of SDAE before and after SMOTE.

Dataset	Accuracy (%)	
	Before SMOTE	After SMOTE
miRNA	74%	79%
mRNA	72%	80%
Copy Number Variation	62%	66%

Single-omic variation	64%	75%

These improvements thus underscore the effectiveness of SMOTE in mitigating class imbalance, which currently allows SDAE models to grasp meaningful patterns and features from the data. There are marked increases in accuracy that translate to improvement in model performance regarding feature extraction and representation learning, which are very important in single-omic variation analyses. It also sheds light on the need for further research into combined omics datasets to evaluate by how much integrated data from single-omic variation sources can enhance predictive capabilities for cancer subtype classification and other biomedical applications

3.3. DBN

Table 8 shows the performance measurement of DBN before and after SMOTE. Each of these datasets have pre-processed and undergone feature selection, SVM-RFE. Later, SMOTE will be applied to balance the class in datasets. Initially, without applying SMOTE, the accuracy in miRNA data, mRNA data, CNV data and integration of these three datasets are 82%, 78%, 73% and 87% respectively. After applying SMOTE, the accuracies of miRNA and mRNA increase to 86% and 89%. However, the accuracy of CNV dropped slightly from 73% to 70% and single-omic variation remained the same at 87%.

Table 8. Performance measurement of DBN before and after SMOTE.

Dataset	**Accuracy (%)**	
	Before SMOTE	**After SMOTE**
miRNA	82%	86%
mRNA	78%	89%
Copy Number Variation	73%	70%
Single-omic variation	87%	87%

The increase of accuracy by 4% in miRNA suggests that the miRNA dataset had some level of class imbalance, and the application of SMOTE helped to mitigate this by synthetically generating new samples for the minority class. The model will likely gain a better understanding of the minority class through class distribution, and thus improve the overall performance. Moreover, there is a significant increase of accuracy in mRNA dataset by 11%. This indicates that there is a strong presence of class imbalance in mRNA dataset. SMOTE plays an important role in generating synthetic examples for the underrepresented class and producing a more balanced training set. This explains why the model learns and performs significantly better. However, CNV dataset shows that the accuracy decreases by 3% and this might be due to several factors. For instance, it may be caused by overfitting to synthetic data and data complexity. While using SMOTE, the model may learn too much from the synthetic samples which are created to balance class, however, they do not

perfectly represent real data. This causes the model to be performed well in training data but struggles to generalize new and unseen test data, resulting in low accuracy. Besides, the CNV data is complex and SMOTE may not be able to capture the complexities accurately, creating less useful synthetic samples.

4. Discussion

The gains in accuracy after applying SMOTE prove that it is indeed effective to improve the class imbalance challenge with omics datasets. Class imbalance skewens model training to biased predictions towards the majority class and reduces the overall performance on minority classes. Obviously, in this study, initial disparities in dataset sizes among miRNA, mRNA, and CNV datasets were present, where the CNV dataset contained the fewest samples. This likely lowered the initial accuracies, the CNV dataset is the lowest, at 62%. Due to class imbalance in datasets, the SDAE models were biased towards the majority classes and failed to capture representative features for the minority classes. SMOTE balanced the datasets by generating synthetic samples from the minority classes, allowing the SDAE models to learn more representative and generalized features. These improvements point toward the urgency of dataset preparation with respect to model performance. The increase in accuracy to 79% and 80% for the miRNA and mRNA datasets, respectively, further confirms that SMOTE really empowered these SDAE models to grasp relevant patterns that were otherwise obscured in these datasets. In contrast, a modest gain of 66% accuracy in the case of CNV suggests that even after using SMOTE, other factors like intrinsic complexity of the dataset or representing features can still impact models performance. Moreover, the class accuracy improved from 64% to 75% in the single-omic variation dataset, underpinning the overall effectiveness of SMOTE in a combined dataset scenario. This means that in integration into single-omic variation data, class imbalance handling is important for robust and accurate predictions. These results point, in particular, to the need for proper dataset preparation and applying techniques like SMOTE for the reduction of class imbalance in order to further improve the generalization and performance of machine learning models using a wide array of datasets.

On the other hand, in the DBN Model, miRNA data and mRNA data are benefitted from SMOTE, indicating that class imbalance was an issue. The synthetic samples helped the models to better learn the minority class patterns, improving their generalization ability. The improvements also indicate that the models were initially biased towards the majority class. SMOTE helped balance the dataset, allowing the models to perform better on the minority class, thereby improving overall accuracy. While in CNV data, the decrease in accuracy suggests that the synthetic samples added noise rather than valuable information. Synthetic samples in CNV data were not representative of the minority class and the data is more complex, containing complexities that SMOTE cannot handle well. To explain the inconsistencies accuracy results occur in the model, this might be due to accuracy being more efficient in predicting the majority class than the minority class. Therefore, the accuracy in CNV data is higher before applying SMOTE. Accuracy alone is insufficient to evaluate the model's performance on imbalanced data, and the decline in accuracy may

indicate that the synthetic samples generated by SMOTE did not effectively represent the minority class in the CNV dataset. Alternative metrics and approaches are necessary to understand and address the underlying issues in class imbalance, ensuring reliable predictions for both majority and minority classes [22]. Meanwhile, for single-omic variation data, the accuracy remains the same as it might be due to the synthetic samples generated by SMOTE not adding meaningful information to the dataset, resulting in no change in performance. Generally, SMOTE is effective for datasets with class imbalance like miRNA and mRNA datasets. However, the effectiveness can be varied based on the nature of data. The characteristics and quality of the datasets play a crucial role in how they respond to SMOTE. Datasets with less noise are likely to benefit more from SMOTE.

5. Conclusions

In conclusion, the performance of Stacked Denoising Autoencoders (SDAE) and Deep Belief Networks (DBN) was evaluated and compared in terms of their accuracy on individual datasets (miRNA, mRNA, CNV) and integrated single-omic variation datasets. The results demonstrated that DBNs consistently outperformed SDAEs, showcasing their utility in classifying breast cancer subtypes. The superior performance of DBNs can be attributed to their hierarchical feature learning capabilities, which enable them to build complex representations layer by layer. This is particularly beneficial for capturing intricate patterns in high-dimensional data.

However, several limitations and drawbacks were identified that need to be addressed in future work. One notable issue was the negative impact of SMOTE on the CNV dataset when used with the DBN model. To handle class imbalance more effectively, alternative techniques such as Tomek Links, Edited Nearest Neighbors (ENN), and Adaptive Synthetic Sampling (ADASYN) should be considered. Additionally, hyperparameter tuning of the DBN model after applying SMOTE could help in achieving better performance by optimizing the model's parameters to suit the augmented data. Future research should also explore the incorporation of additional omics data to provide more informative and comprehensive results. By integrating a broader range of biological data, it may be possible to enhance the robustness and accuracy of the models in classifying breast cancer subtypes. This holistic approach will contribute to a more nuanced understanding of the disease and potentially lead to improved diagnostic and prognostic tools.

Author Contributions: Introduction, HR; literature search, HR, LJY; methodology, CMX, HR, LJY, NSZ; experiment, result analysis and validation, CMX, HR, LJY, NSZ; writing—original draft preparation, CMX HR LJY NSZ; writing—review and editing, CMX HR LJY NSZ.

Funding: No external funding was provided for this research.

Acknowledgements: The authors would like to express gratitude to Dr Azurah A Samah who gave guidance and support to this research.

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

References

1. Ajithkumar, T. (2023). Breast cancer. In Oxford University Press eBooks (pp. 219–244). https://doi.org/10.1093/med/9780198722694.003.0007

2. Mutebi, M., Unger-Saldaña, K., & Ginsburg, O. (2023). Breast cancer. In Routledge eBooks (pp. 91–97). https://doi.org/10.4324/9781003306689-14

3. Wang, S., & Lee, D. (2023). Identifying prognostic subgroups of luminal-A breast cancer using deep autoencoders and gene expressions. PLOS Computational Biology/PLoS Computational Biology, 19(5), e1011197. https://doi.org/10.1371/journal.pcbi.1011197

4. Liu, C., Chen, J., Tsai, Y., Chao, T., Wang, W., Lien, P., Hsu, C., Lai, J., Huang, C., Chiu, J., & Tseng, L. (2024). Abstract 7653: Role of ITGB2 in patients with luminal B1 breast cancer. Cancer Research, 84(6_Supplement), 7653. https://doi.org/10.1158/1538-7445.am2024-7653

5. Jiang, M., Liu, J., Li, Q., & Xu, B. (2023). The trichotomy of HER2 expression confers new insights into the understanding and managing for breast cancer stratified by HER2 status. International Journal of Cancer, 153(7), 1324–1336. https://doi.org/10.1002/ijc.34570

6. Xing, C., Ma, L., & Yang, X. (2016). Stacked Denoise Autoencoder based feature extraction and classification for hyperspectral images. Journal of Sensors, 2016, 1–10. https://doi.org/10.1155/2016/3632943

7. Ghosh, S. K., Biswas, B., & Ghosh, A. (2019). SDCA: a novel stack deep convolutional autoencoder – an application on retinal image denoising. IET Image Processing, 13(14), 2778–2789. https://doi.org/10.1049/iet-ipr.2018.6582

8. Gu, F., Khoshelham, K., Valaee, S., Shang, J., & Zhang, R. (2018). Locomotion activity recognition using stacked denoising autoencoders. IEEE Internet of Things Journal, 5(3), 2085–2093. https://doi.org/10.1109/jiot.2018.2823084

9. Du, B., Xiong, W., Wu, J., Zhang, L., Zhang, L., & Tao, D. (2017). Stacked convolutional denoising Auto-Encoders for feature representation. IEEE Transactions on Cybernetics, 47(4), 1017–1027. https://doi.org/10.1109/tcyb.2016.2536638

10. Hua, N. Y., Guo, N. J., & Zhao, N. H. (2015). Deep Belief Networks and deep learning. In Proceedings of 2015 International Conference on Intelligent Computing and Internet of Things (pp. 1-4). https://doi.org/10.1109/icaiot.2015.7111524

11. Zhong, P., Gong, Z., Li, S., & Schonlieb, C. (2017). Learning to Diversify Deep Belief Networks for Hyperspectral Image Classification. IEEE Transactions on Geoscience and Remote Sensing, 55(6), 3516–3530. https://doi.org/10.1109/tgrs.2017.2675902

12. Khatami, A., Khosravi, A., Nguyen, T., Lim, C. P., & Nahavandi, S. (2017). Medical image analysis using wavelet transform and deep belief networks. Expert Systems With Applications, 86, 190–198. https://doi.org/10.1016/j.eswa.2017.05.073

13. Sarikaya, R., Hinton, G. E., & Deoras, A. (2014). Application of deep belief networks for natural language understanding. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 22(4), 778–784. https://doi.org/10.1109/taslp.2014.2303296

14. Dai, X., Cheng, J., Gao, Y., Guo, S., Yang, X., Xu, X., & Cen, Y. (2020). Deep belief network for feature extraction of urban artificial targets. Mathematical Problems in Engineering, 2020, 1–13. https://doi.org/10.1155/2020/2387823

15. Danaee, P., Ghaeini, R., & Hendrix, D. A. (2016). A deep learning approach for cancer detection and relevant gene identification. In Pacific Symposium on Biocomputing 2017 (pp. 219-229). https://doi.org/10.1142/9789813207813_0022

16. Luo, J., Wu, M., Gopukumar, D., & Zhao, Y. (2016). Big data application in Biomedical Research and Health Care: A Literature review. Biomedical Informatics Insights., 8, BII.S31559. https://doi.org/10.4137/bii.s31559

17. Hasanin, T., Khoshgoftaar, T. M., Leevy, J., & Seliya, N. (2019). Investigating Random Undersampling and Feature Selection on Bioinformatics Big Data. 2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService), 346–356. https://doi.org/10.1109/bigdataservice.2019.00063

18. Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using Support Vector Machines. Kluwer Academic Publishers, 46(1/3), 389–422. https://doi.org/10.1023/a:1012487302797

19. Mahmoud, A., El-Kilany, A., Ali, F., & Mazen, S. (2021). TGT: a novel adversarial guided oversampling technique for handling imbalanced datasets. Egyptian Informatics Journal/Egyptian Informatics Journal, 22(4), 433–438. https://doi.org/10.1016/j.eij.2021.01.002

20. Xia, J., Sun, L., Xu, S., Xiang, Q., Zhao, J., Xiong, W., Xu, Y., & Chu, S. (2020). A Model Using Support Vector Machines Recursive Feature Elimination (SVM-RFE) Algorithm to Classify Whether COPD Patients Have Been Continuously Managed According to GOLD Guidelines. International Journal of Chronic Obstructive Pulmonary Disease/International Journal of COPD, Volume 15, 2779–2786. https://doi.org/10.2147/copd.s271237

21. Feng, X., Xie, W., Dong, L., Xin, Y., & Xin, R. (2024). COMBINE: A Comprehensive Multi-Omics approach for improving breast cancer prognosis classification in African American women. Research Square (Research Square). https://doi.org/10.21203/rs.3.rs-3852479/v1

22. Saad Hussein, A., Li, T., Yohannese Chubato, W. and Bashir, K. (2019). A-SMOTE: A New Preprocessing Approach for Highly Imbalanced Datasets by Improving SMOTE. International Journal of Computational Intelligence Systems. doi:https://doi.org/10.2991/ijcis.d.191114.002.