Breast Cancer Subtypes Classification Using SDAE and 1D CNN Based on Omics Variation

Presented by Group 6:
Chong Kah Wei (A20EC0027)
Heong Yi Qing (A20EC0043)
Mek Zhi Qing (A20EC0077)
Zereen Teo Huey Huey (A20EC0173)

Introduction

Breast Cancer

- Breast cancer was the **leading cause of death** for women all over the countries.
- The classification of breast cancer subtypes is significant as it **helps to categorize patients who may benefit from targeted treatment** to receive the appropriate treatment.

Omics Data

- Different omics data usually will provide different information that is related to the disease.
- The **study of one type of omics data is restricted to correlation**, which mostly reflects the reaction process.
- Different variation of omics data are used including miRNA data, DNA methylation data and Copy Number Variation (CNV) data.

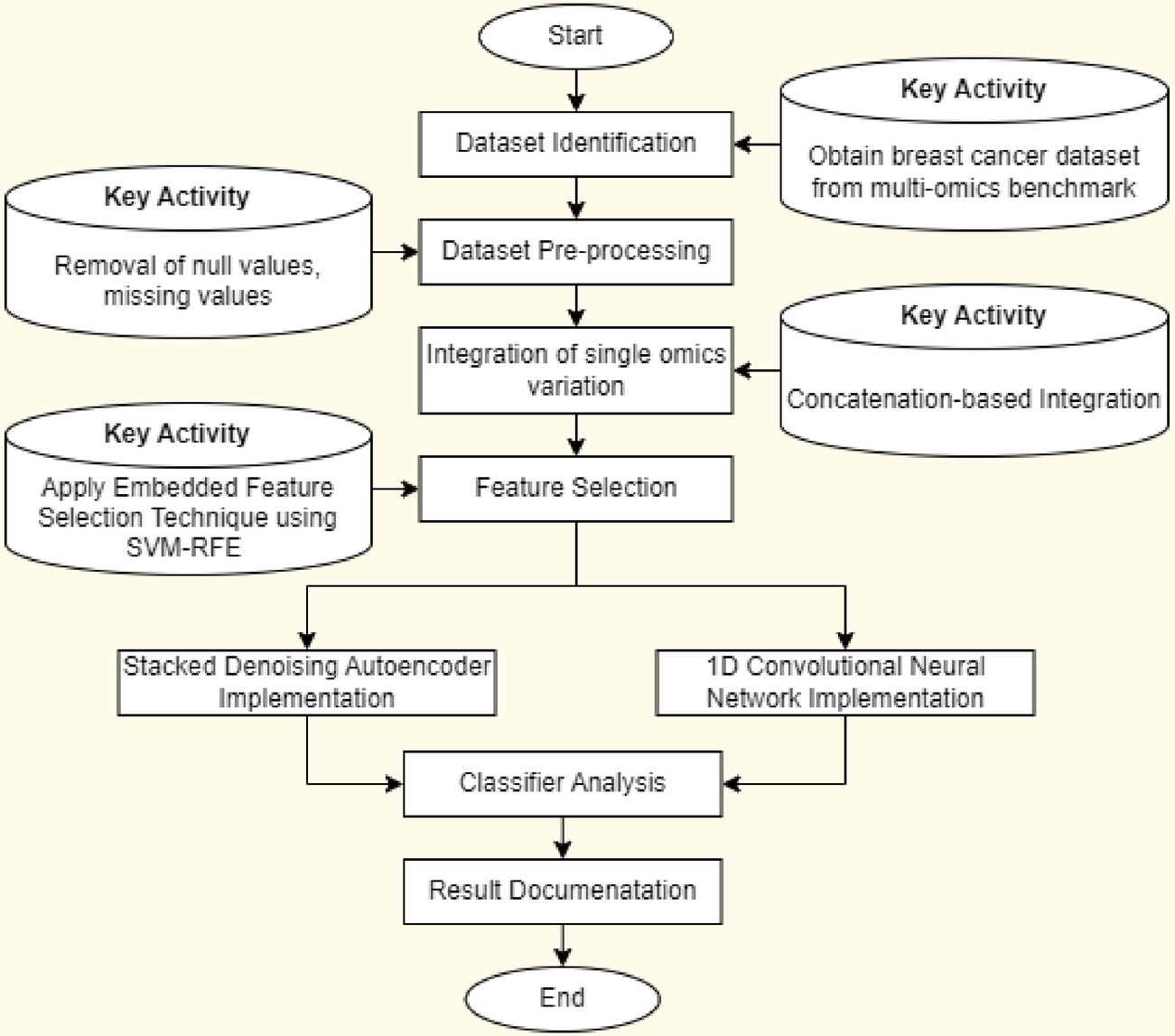
Research Aim

To classify the breast cancer subtypes based on the integration of omics data using the deep learning approaches

Literature Review

Publications	Title	Types of Data Used	Machine Learning / Deep Learning Algorithm	Findings
Lin *et al.*, 2020 [8]	Classifying Breast Cancer Subtypes Using Deep Neural Networks Based on Multi-Omics Data	• mRNA • DNA methylation • Copy Number Variation (CNV)	• Deep Neural Network (DNN)	• Multi-omics dataset outperform compared to single omics data.
Liu *et al.*, 2017 [9]	Tumor gene expression data classification via sample expansion-based deep learning	• Tumor microarray gene expression dataset of breast cancer, 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]	Breast Cancer Type Classification Using Machine Learning	• RNA Sequence data	• SVM • KNN • Naïve Bayes • Decision tree	• SVM shows more efficient compare to other 3 algorithms
Yang *et al.*, 2024 [11]	Comparative Evaluation of Machine Learning Models for Subtyping Triple-Negative Breast Cancer: 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 Breast Cancer 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 Breast Cancer Subtype Classification Using Multi-Omics Data	• Clinical data • Gene expression • Copy Number Variations (CNV) • Copy Number Abbreviations(CNA)	• Deep Forest	• Integration of omics do 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)	• TVAE outperforms SDAE with 91.86% accuracy, 22.73% specificity and 0.21% Matthews Correlation Coefficient (MCC).

Experimental Framework



Omics Datasets

The datasets collected in this study are the pre-processed Breast Cancer (BRCA) datasets extracted from TCGA Genome.

Types of data	Number of Patients	Number of Features
Copy Number Variation (CNV)	671	19569
DNA methylation	671	19050
miRNA expression	671	368

Data Preprocessing

1. Checking for missing and duplicate values
2. Transform original nominal type of class label into numerical data
3. Data transformation with Min-max Normalization
4. Train test split with ratio of 70:30

Data Integration

To integrate the omics datasets, Concatenation-based Integration (CBI) is used. The CNV, DNA methylation and miRNA expression datasets are combined into a single large matrix.

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

Feature Selection: SVM-RFE

- BRCA omics datasets are **large**, **consuming** computational **time** and **storage**.
- Feature selection reduces dimensionality, reducing overfitting and improving model performance
- SVM-RFE is an **iterative** method that **eliminates least important features** based on the **weights** of the features.
- The feature with the lowest weight indicates the feature is least important.
- Iteration is repeated **until desired number of features is reached**.

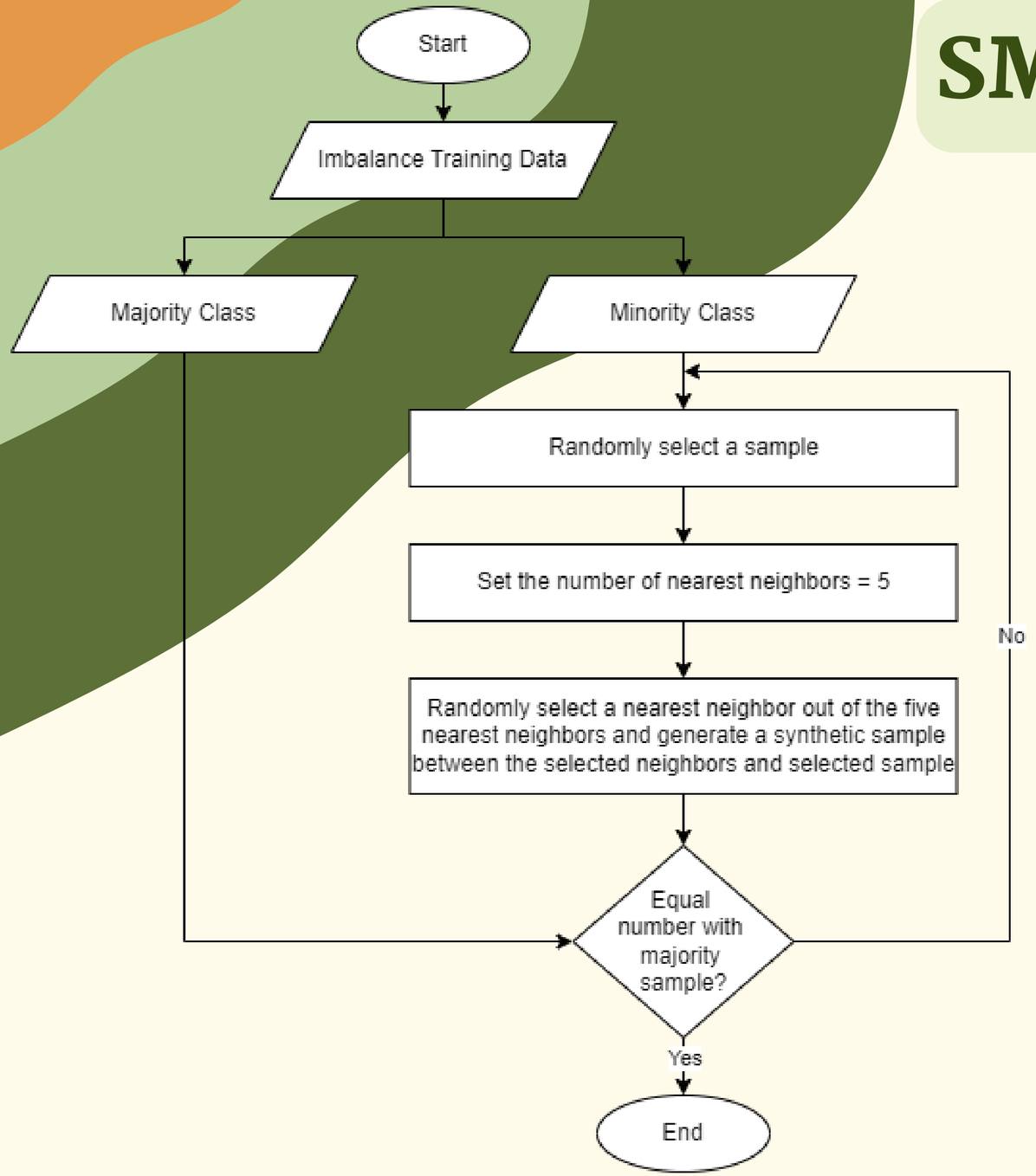
```
Input:
  - Dataset X with features (genes) X1, X2, ..., Xn
  - Labels y
  - Number of features to select k

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
```

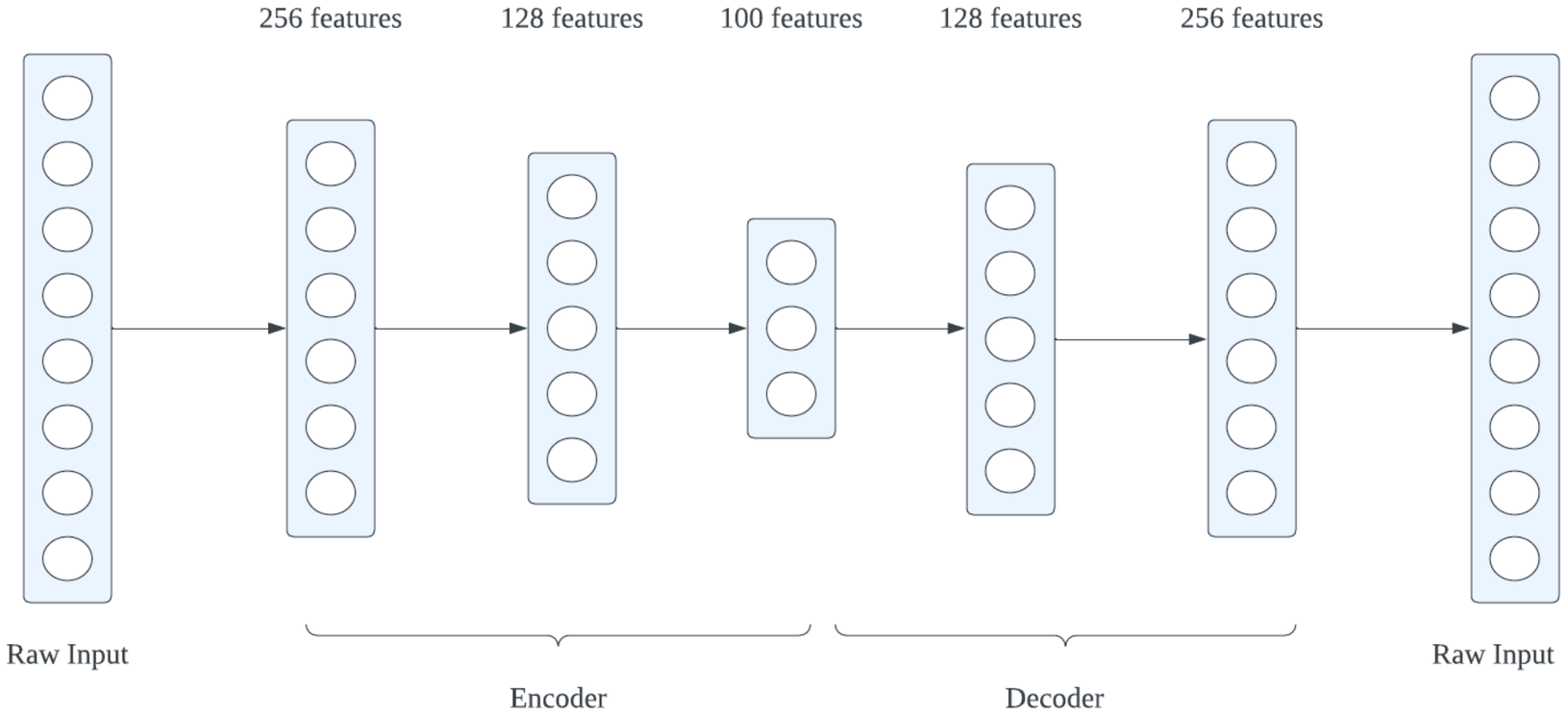
SMOTE



- Synthetic Minority Oversampling Technique (SMOTE) **prevents bias** in models favoring majority classes.
- SMOTE generates **synthetic samples** for minority class to increase minority samples and balance class distribution.
- **Nearest Neighbors** (**K**) is set at **five**, with one selected for each minority sample.
- Difference between selected minority sample and nearest neighbor is computed to form a synthetic minority sample.
- **Sampling strategy** is set to **auto**, resampling only minority class.
- Stop when the number of minority samples is equal with the number of majority samples.

SDAE(Stacked Denoising Autoencoder)

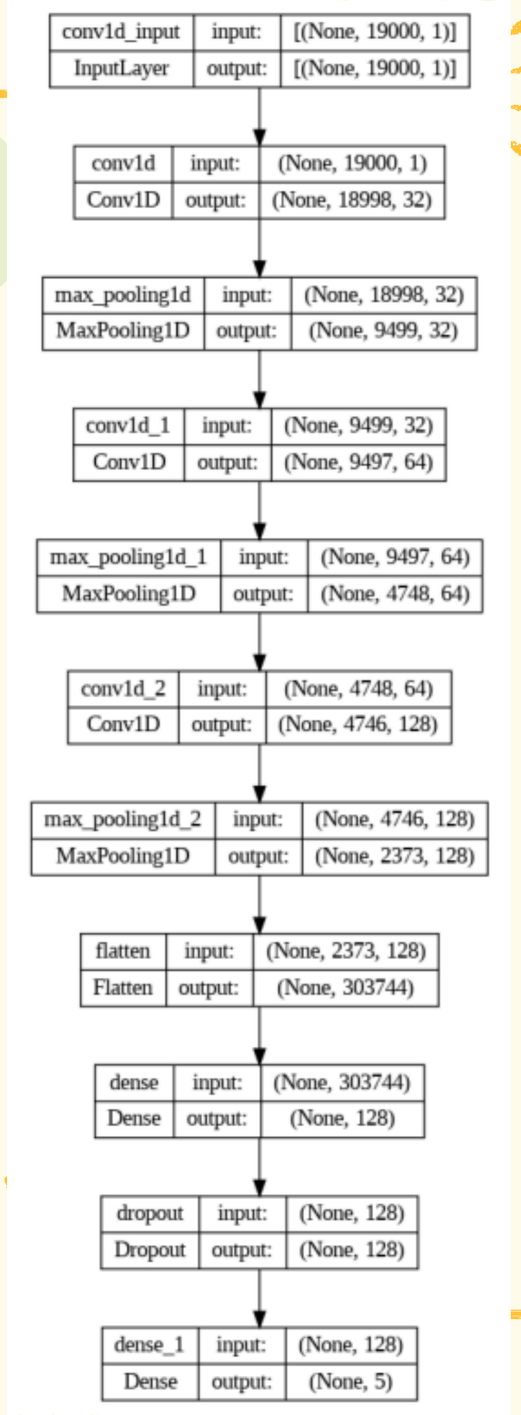
- to extract reliable features from corrupted input data.
- extends the basic autoencoder by stacking multiple layers of encoders and decoders



- Encoder reduce the dimensionality of data
- Decoder reconstructs the data
- It was pair with a MLP deep learning classifier for the classification task
- Model was built with an early stopping with patience rate of 20 and loss rate monitoring.
- The model stops training when the monitored metric has stopped improving.

1D CNN (1D Convolutional Neural Network)

- deep learning algorithm that is designed to handle sequential data
- applies filters along one dimension using convolutional layers to discover underlying patterns in the dataset
- all dataset using the same 1D CNN architecture, “input_shape” in the first convolutional layer will be varied with the number of features



Convolutional layers

- Three Conv1D layers with increasing filter sizes (32, 64, 128) to learn hierarchical features, followed by MaxPooling1D layers to reduce the spatial dimensions.

Flatten Layer

- Transforms the multi-dimensional output into a 1D vector.

Dense Layer

- One dense layer with 128 neurons and a dropout layer to prevent overfitting.

Output Layer

- A dense layer with 5 neurons (for 5 classes) and Softmax activation for multi-class classification.

Result & Discussion

Model	Type of Omic	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**

- SMOTE can generally improve the performance of the model, but does not really increase much except for **miRNA** dataset

Result & Discussion

Model	Type of Omic	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**

- 1D CNN model **more sensitive** to **class imbalance** - shows greater improvement in classification accuracy after applying oversampling technique
- Not only perform better than SDAE but also benefits more from the application of SMOTE

Comparison of the Result based on Models

	Accuracy(%)	
	SDAE	1DCNN
CNV	**71.78**	70.79
DNA methylation	76.73	**81.19**
miRNA	79.21	**80.20**
CNV + DNA methylation + miRNA	71.29	**80.20**

- 1D CNN model **outperforms** the SDAE model in most cases (except for CNV dataset)
- Application of **SMOTE has consistently improved the classification performance** across all cases
- Integrated variations of omics does not produce the best result for both models.
- SDAE model works best with miRNA data, 1D CNN model obtains the best result with DNA methylation data.

Conclusion

1D CNN model and SDAE model can **perform better after applying SMOTE**

1D CNN model generally **achieves higher accuracy** than the SDAE model, *except using the CNV dataset.*

SDAE model **works best with miRNA data** while the 1D CNN model **works best with DNA methylation data**

The best model is **1D CNN using DNA methylation data with the application of SMOTE**, with the accuracy of **81.19%**.

Thank You