Created By : Group 3

Classification of Breast Cancer Subtypes Using Transcriptomic Data with SDAE and DBN Methods

Group Member:

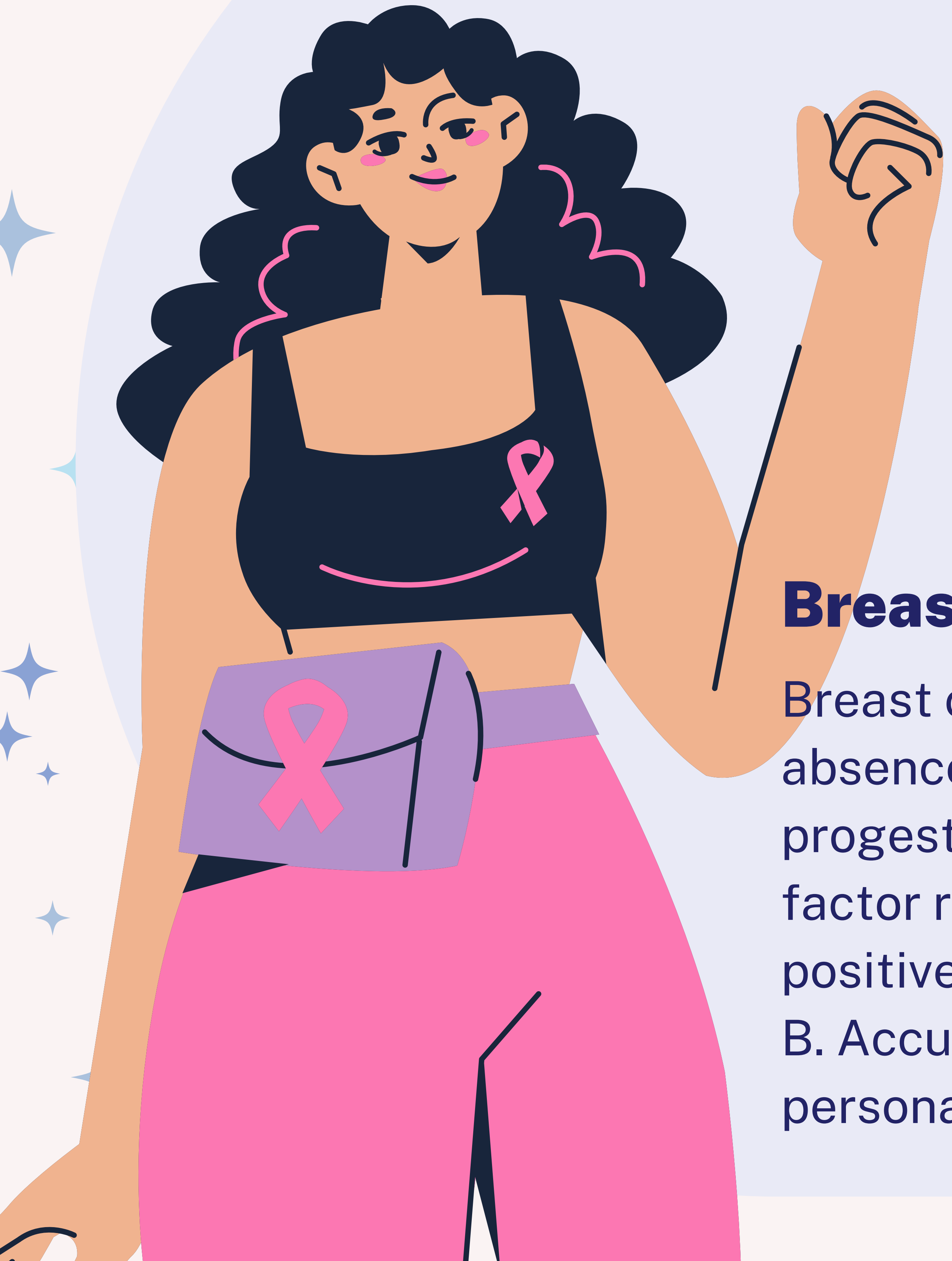
1 Chang Min Xuan

2 Hanis Rafiqah

3 Lee Jia Yee

4 Nik Syahdina

Introduction

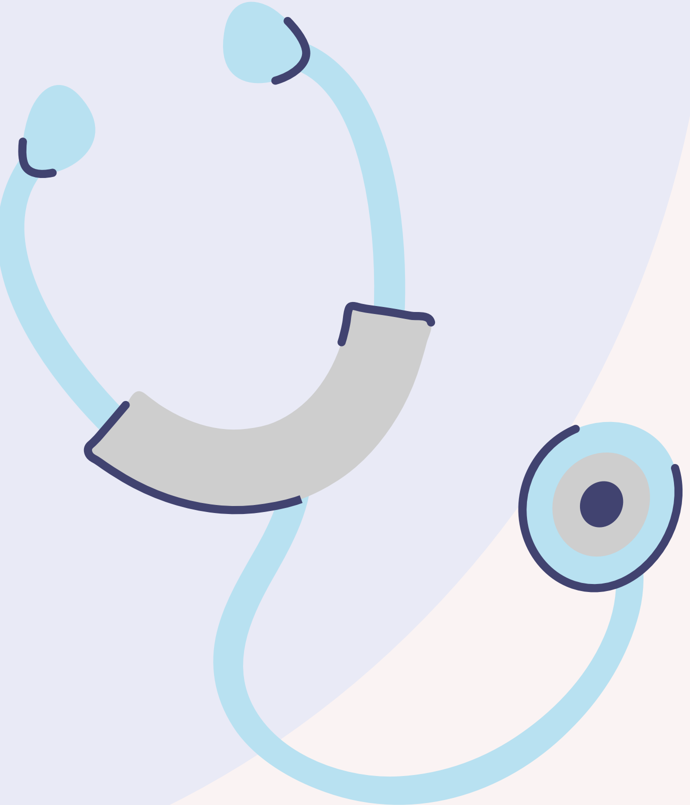


About Breast Cancer

Breast cancer is the most common cancer in women worldwide, with significant risk factors including gender, age, heredity, and lifestyle choices. Regular screening and early detection are crucial for effective treatment, which may involve surgery, radiotherapy, chemotherapy, hormone therapy, and targeted drug therapies.

Breast Cancer Subtypes

Breast cancer can be classified based on the presence or absence of specific receptors: estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2). Subtypes include ER-positive/PR-positive, HER2-positive, triple-negative, and Luminal A and B. Accurate subtype identification is essential for personalized treatment strategies.



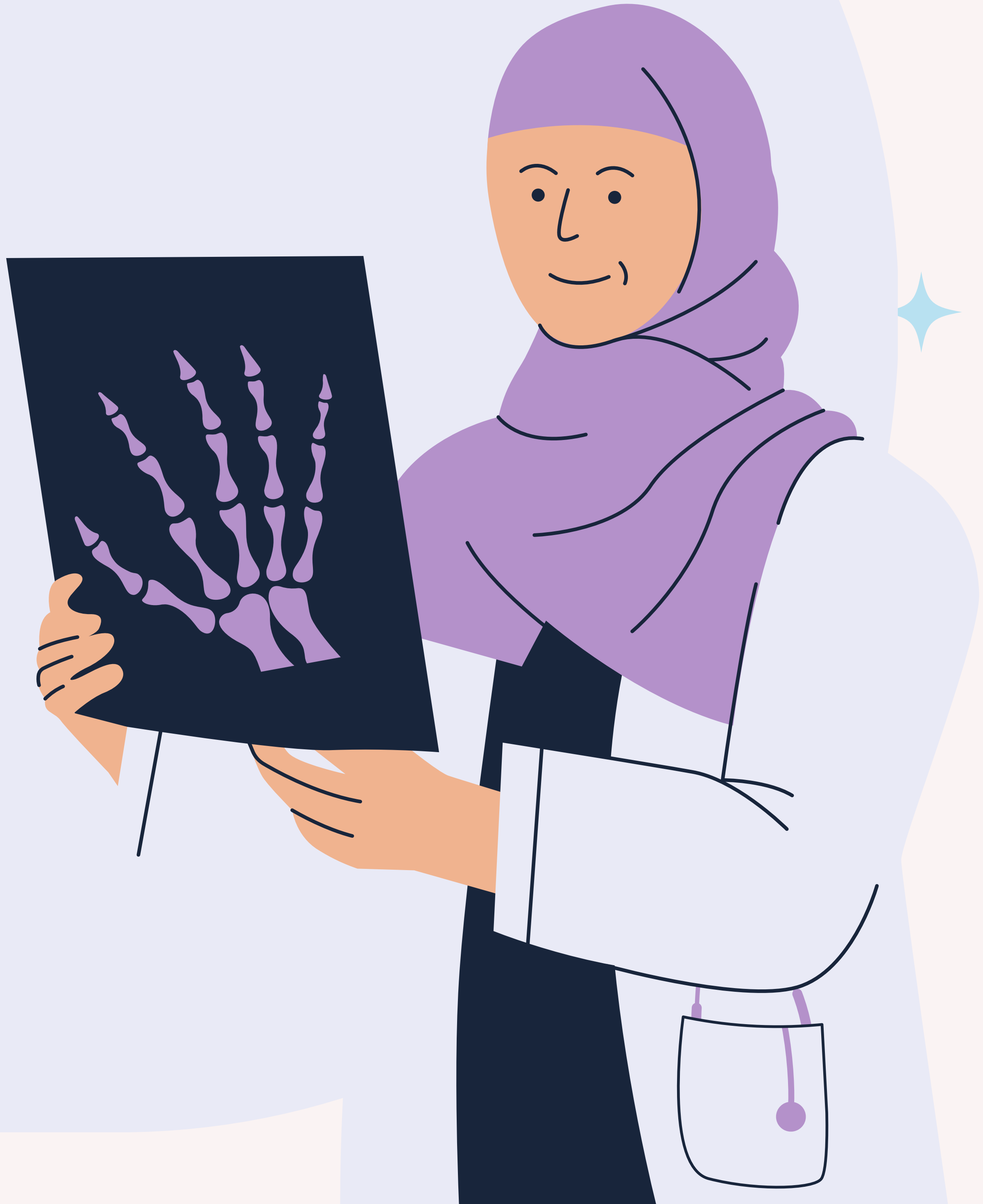
Introduction

Stacked Denoising Autoencoders (SDAE)

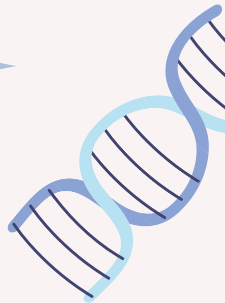
SDAE is a type of artificial neural network designed to improve feature learning by adding noise during training. It consists of multiple layers of denoising autoencoders, trained to denoise the output of the previous layer, resulting in robust feature learning and improved generalization to new data.

Deep Belief Networks (DBN)

DBNs are multi-layered networks that learn data features in an unsupervised manner through layer-wise training. They are effective in modeling complex, high-dimensional data and are widely used in image recognition, natural language processing, and feature extraction. Both SDAE and DBN are effective in classifying breast cancer subtypes using gene expression data, aiding in accurate diagnosis and personalized treatment.



Omics Dataset



CNV Data

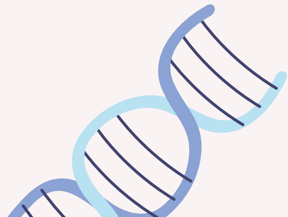
Refers to the variations in the number of copies of a particular gene or region of the genome. CNVs can involve deletions, duplications, and large-scale structural variations in the DNA.

mRNA Data

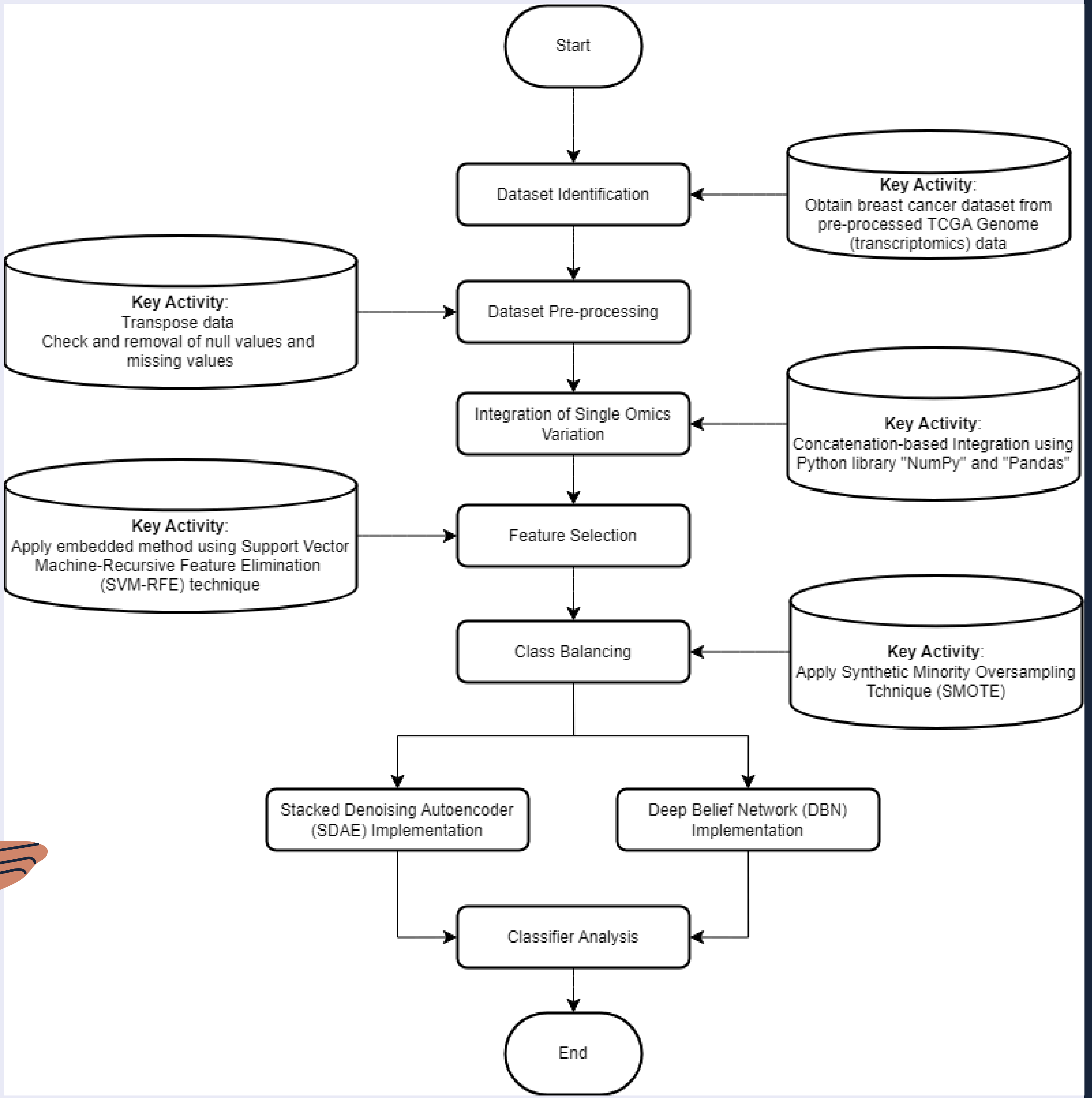
Refers to the information derived from messenger RNA (mRNA) molecules, which play a crucial role in the process of gene expression.

miRNA Data

Refers to the expression levels of various miRNAs within a sample, which can be measured using techniques like microarray analysis, next-generation sequencing (NGS), or quantitative real-time PCR (qRT-PCR).



Experimental Framework



Start

Dataset Identification

Key Activity: Obtain breast cancer dataset from pre-processed TCGA Genome (transcriptomics) data

Dataset Pre-processing

Key Activity: Transpose data Check and removal of null values and missing values

Integration of Single Omics Variation

Key Activity: Concatenation-based Integration using Python library "NumPy" and "Pandas"

Feature Selection

Key Activity: Apply embedded method using Support Vector Machine-Recursive Feature Elimination (SVM-RFE) technique

Class Balancing

Key Activity: Apply Synthetic Minority Oversampling Tchnique (SMOTE)

Stacked Denoising Autoencoder (SDAE) Implementation

Deep Belief Network (DBN) Implementation

Classifier Analysis

End

Data Preprocessing

The datasets (CNV, miRNA, mRNA) have no missing or duplicates values.

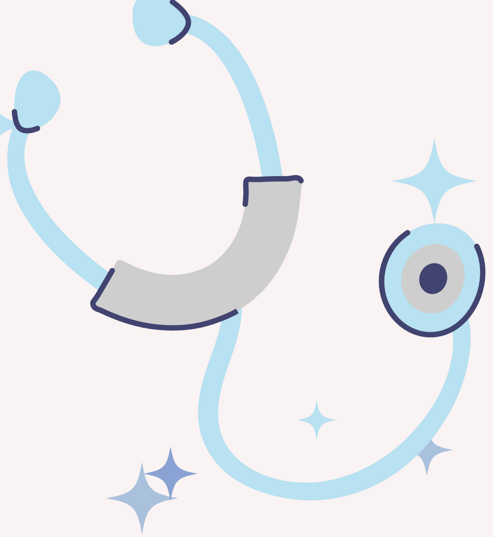
Datasets	Data Transposition	
	Before	After
CNV	(19568, 672)	(672, 19568)
miRNA	(368, 672)	(672, 368)
mRNA	(18206, 672)	(672, 18206)

Data Transposition

The process of transposing rows into columns or vice versa (samples as rows, features as columns)

Data Normalization

Min-max normalization is applied to scale every feature into the range of 0 to 1



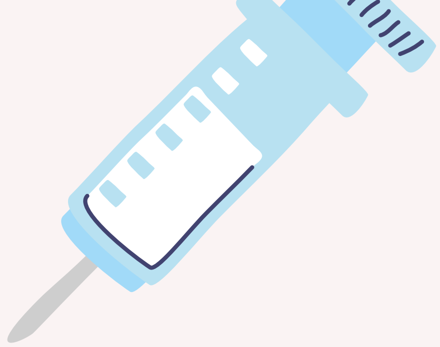
Data Integration

Concatenation-based Integration

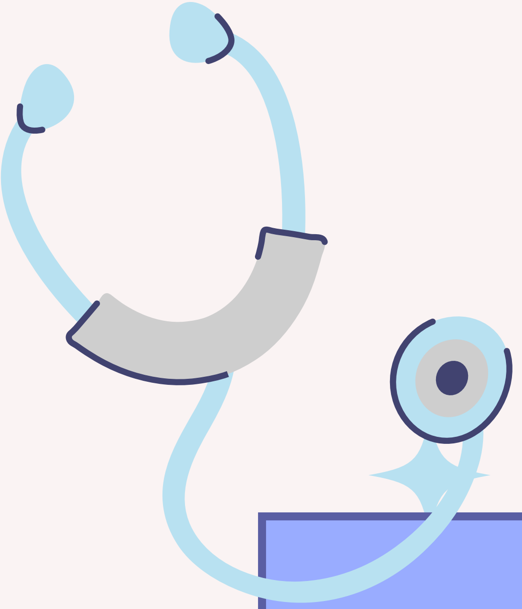
Datasets	Data Integration
CNV	(672, 19568)
miRNA	(672, 368)
mRNA	(672, 18206)
Integrated-omics	(672, 38142)

The CNV, miRNA, and mRNA datasets have the same samples, and they are concatenated by merging them by columns.

It is straightforward and simple to execute.



Feature Selection



Support Vector Machine-Recursive Feature Elimination (SVM-RFE) is used as it can **handle data with high dimensionality** and unbalanced class.

It is 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.

In this research, the number of features selected after feature selection is 30000.

Datasets	Before SVM-RFE	After SVM-RFE	Number of removed features
CNV	19568	5000	14568 (74.45%)
miRNA	368	250	118 (32.07%)
mRNA	18206	5000	13206 (72.54%)
Integrated-omics	38142	30000	8142 (21.35%)

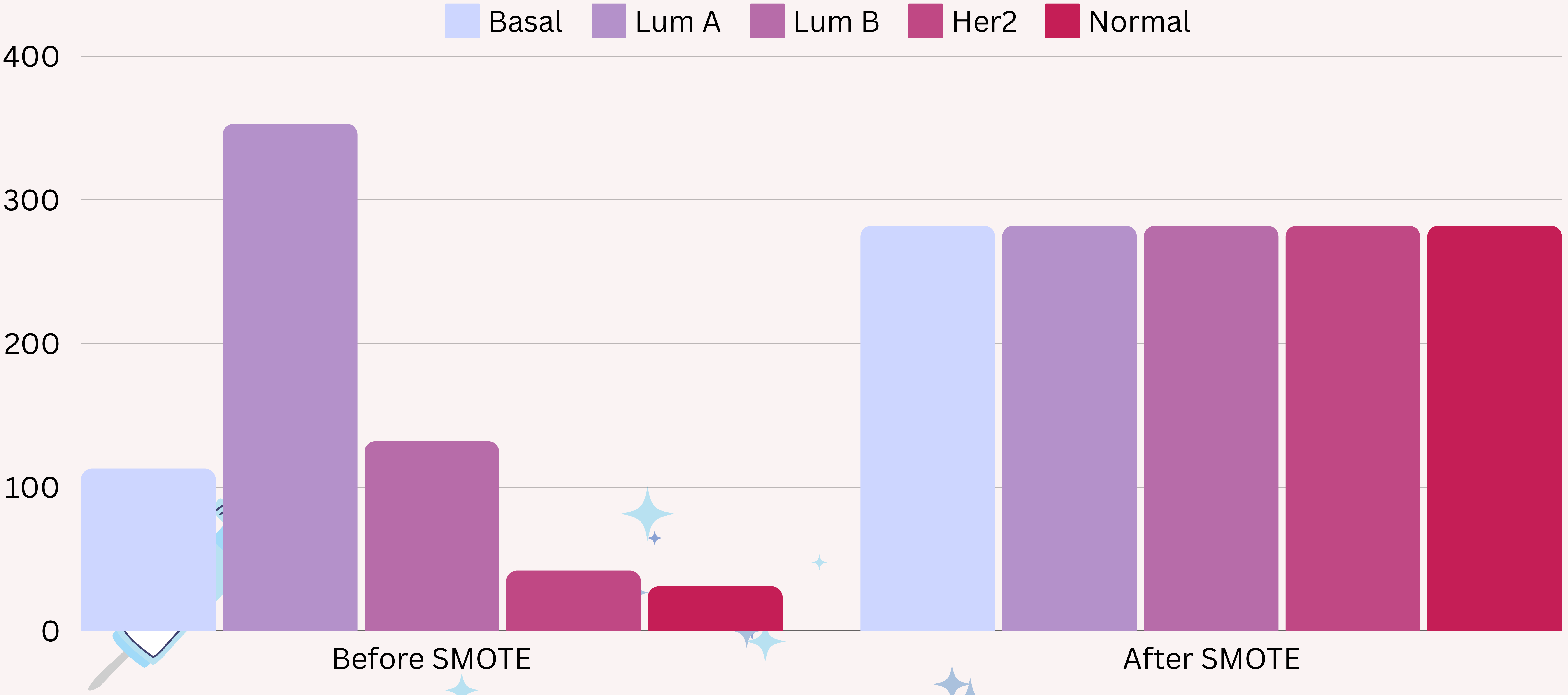
SMOTE

Summary of data before and after resampling

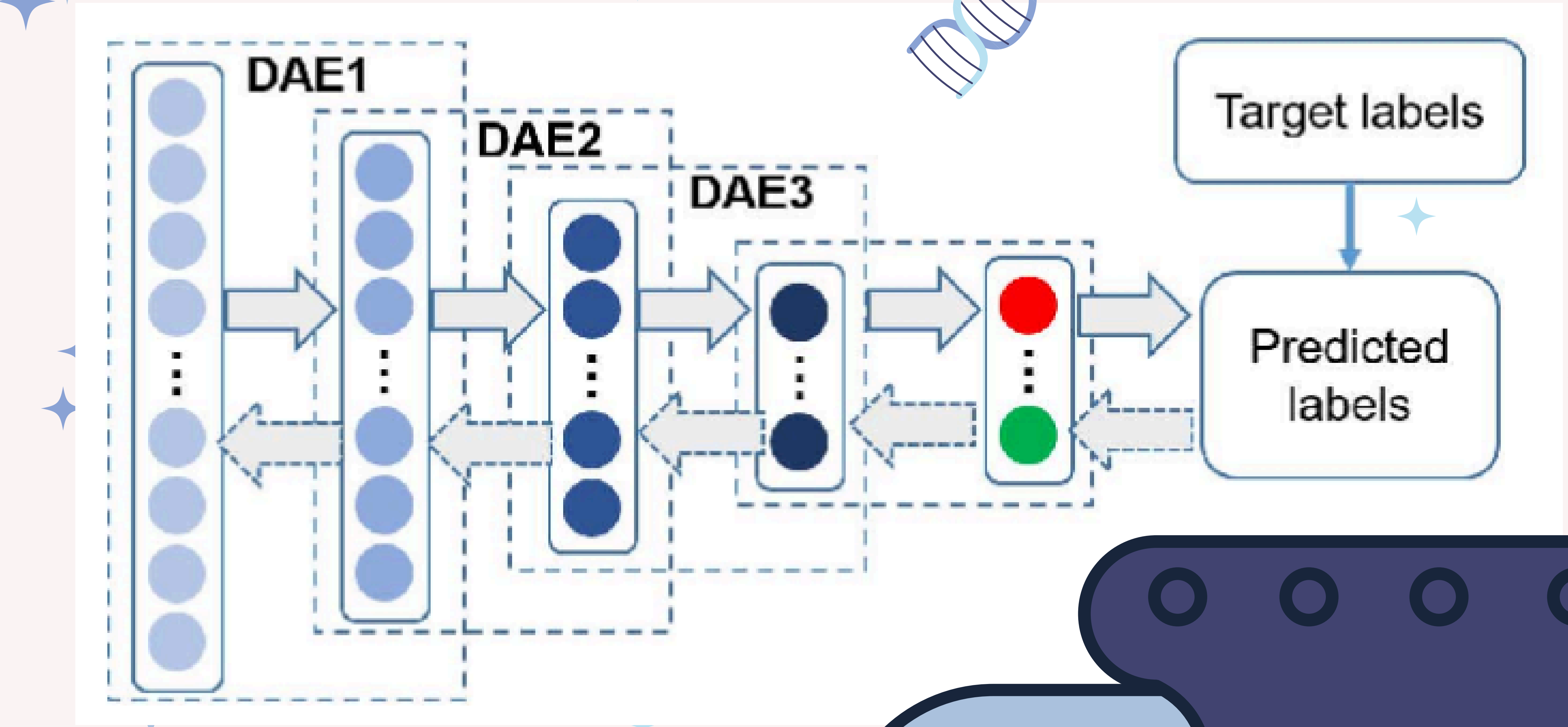
Type of class	Number of samples	
	Before	After
Basal	113	282
Lum A	353	282
Lum B	132	282
Her2	42	282
Normal	31	282

- Identify Minority Instances: Detect minority class instances in the training data.
- Create Synthetic Samples: Generate new synthetic examples by interpolating between each minority instance and its k-nearest neighbors.
- Add Synthetic Samples: Integrate these synthetic samples with the original dataset.

SMOTE


Basal
Lum A
Lum B
Her2
Normal
400
300
200
100
0
Before SMOTE
After SMOTE

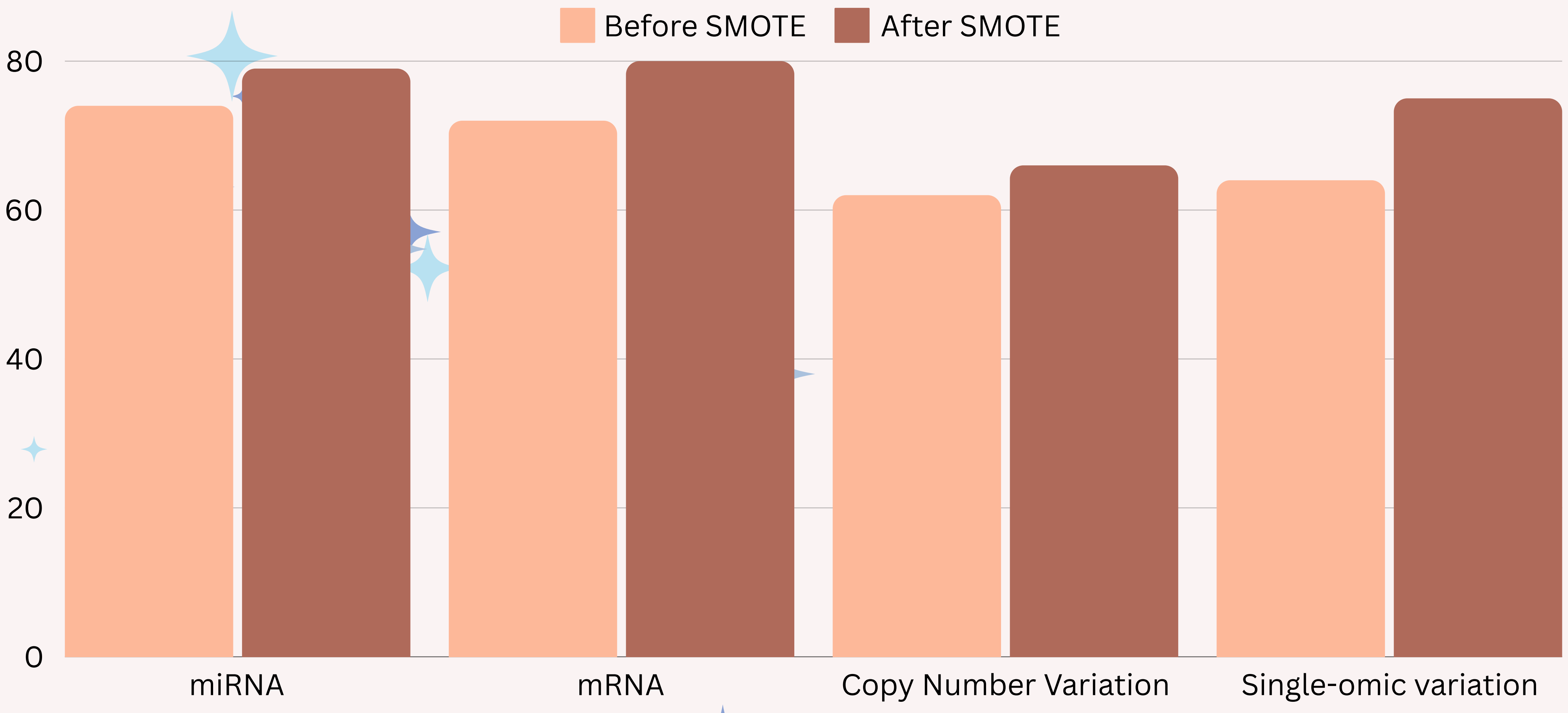
SDAE



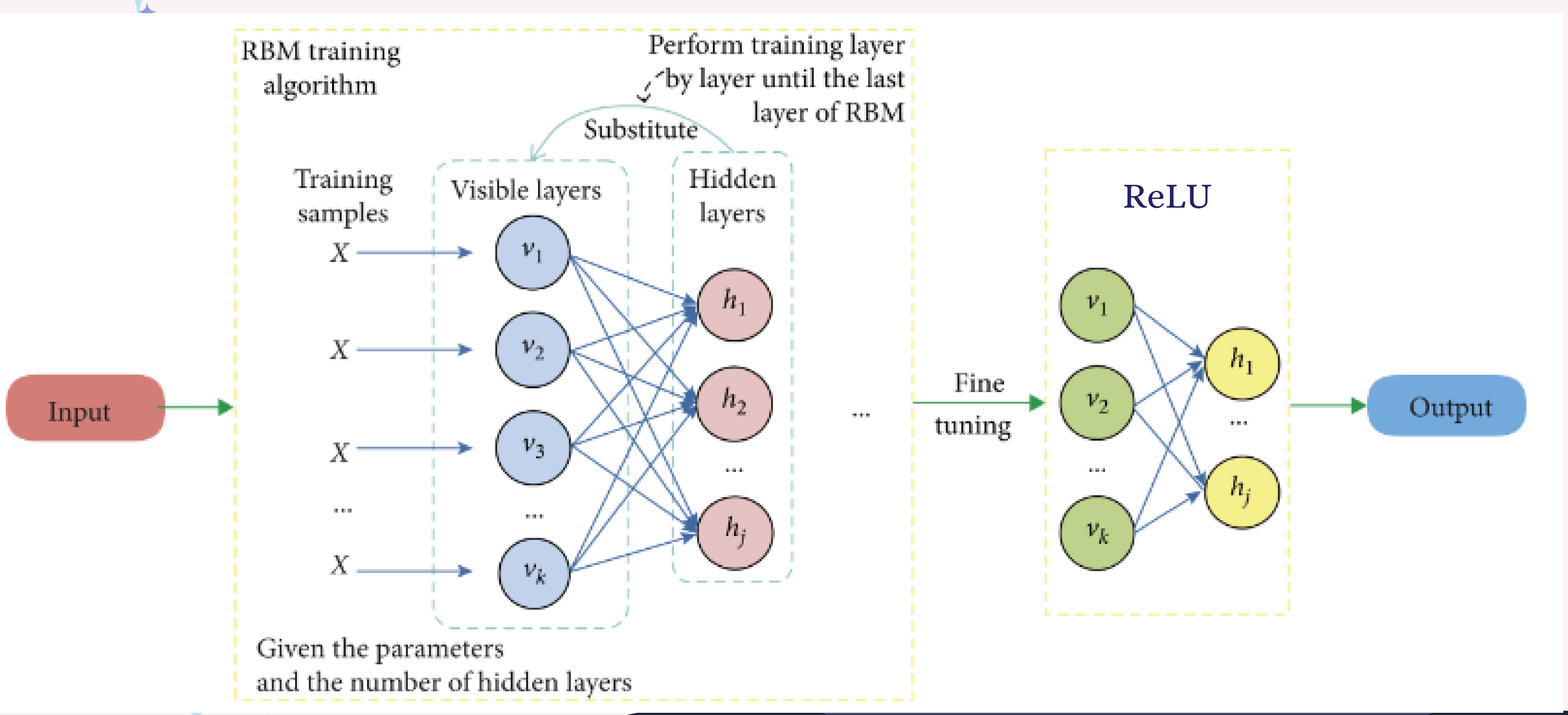
Parameters

Batch size	: 32
Epoch	: 250
input layer	: 30000
hidden layer 1	: 64
hidden layer 2	: 32
Output layer	: 30000

SDAE MODEL PERFORMANCE



The SDAE model achieved accuracies of 74% with the miRNA dataset, 72% with the mRNA dataset, and 62% with the Copy Number Variation (CNV) dataset, highlighting the challenge of class imbalance. After applying SMOTE, accuracy improved significantly: 79% for miRNA, 80% for mRNA, 66% for CNV, and in the single omics variation dataset, accuracy increased from 64% to 73%.



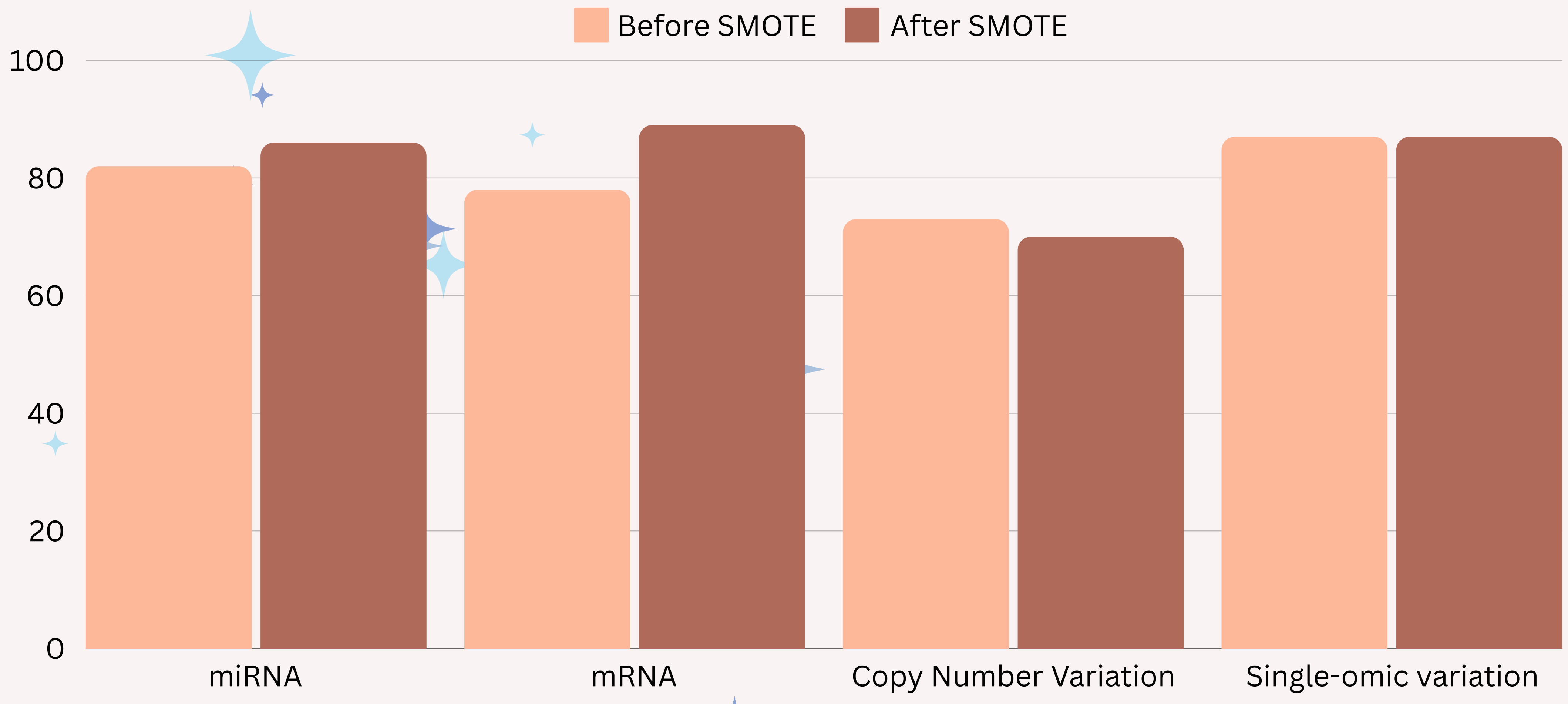
DBN

Parameters

Learning rate (RBM)	: 0.05
Epoch (RBM)	: 20
Learning rate	: 0.01

Batch size	: 64
Epoch	: 200
Input layer	: 30000
Hidden layers	: 256
Output layer	: 5

DBN MODEL PERFORMANCE



Before applying SMOTE, the DBN model achieved accuracies of 82% for miRNA, 78% for mRNA, 73% for CNV and 85% for integration of these three datasets respectively. After applying SMOTE, the accuracies of miRNA increase by 4% to 86% while mRNA data increase significantly by 11% to 89%. However, the accuracy of CNV drops slightly from 73% to 70%. Single-omic variation remains the same for 87%.

Discussion

SDAE Model

All the accuracies in each dataset and single-omic variation dataset increase.

SMOTE able to balance the datasets by generating synthetic samples from the minority classes, allowing the SDAE models to learn more representative and generalized features.

Discussion

DBN Model

The accuracies in miRNA and mRNA increase while in CNV decrease. Single-omic variation remains the same

miRNA & mRNA are benefitted from SMOTE, hence class imbalance was an issue. The synthetic samples helped the models to better learn the minority class patterns, improving their generalization ability.

Decrease accuracy in **CNV** might caused by overfitting to synthetic data and data complexity. The model may learn too much from the synthetic samples, therefore it performs well in training data but struggles to generalize new and unseen test data.

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

Conclusion

Limitations & Future Work

Negative impact of SMOTE on the CNV dataset:

- Use alternative techniques such as Tomek Links, Edited Nearest Neighbors (ENN), and Adaptive Synthetic Sampling (ADASYN)
- Hyperparameter tuning of the DBN model after applying SMOTE

Future Research

Incorporation of additional omics data to provide more informative and comprehensive results

DBNs consistently outperformed SDAEs

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.

Thank you for your attention