PSM 2 PRESENTATION

SESSION 2023/2024-2

Identification of Potential Ovarian Cancer Biomarkers from Imbalanced Gene Expression with Protein-Protein Interactions

PREPARED BY: MEK ZHI QING (A20EC0077)

SUPERVISOR: DR. CHAN WENG HOWE

Presentation Link: https://youtu.be/ul5YbWgv5MM

Demo Link: https://youtu.be/CEcb7PErzbY

Overview

- Introduction, Problem background, Aim, Objectives, Scopes
- Summary of Literature Review
- Summary of Research Methodology
- Summary of Research Design and Implementation
- Summary of Results and Discussion
- Conclusion

Introduction

Biomarker

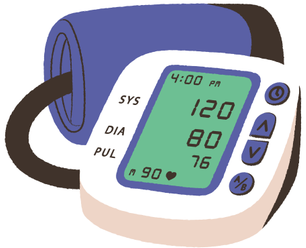
An indicator to measure various processes and responses including normal biological processes, pathogenic processes, and biological responses to the intervention (Ou *et al.*, 2021).

Example:

- Biological molecules
- pulse
- blood pressure (Strimbu & Tavel, 2010)

Importance:

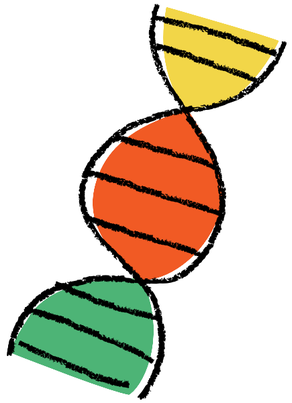
- early detection of disease
- categorization of disease
- identify the existence of high-risk cohort
- assessing response of treatment (Samprathi & Jayashree, 2020)



Introduction

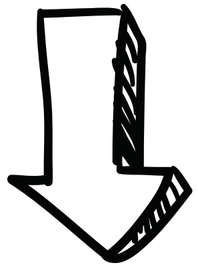
Gene Biomarker

Detected from large amount of gene expression profiling

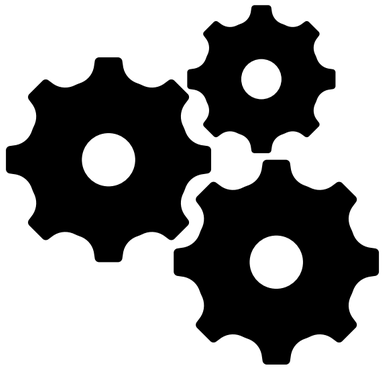


Early Phase in Discovery Biomarker

- Done in laboratory instead of using computational methods
- **Disadvantage:** Time-consuming, expensive



Use computational method (machine learning / deep learning)



Problem Background

analysis of gene expression data through computational method → Biomarker

However

1. prone to fluctuations which may **affect the accuracy of classification** (Zhong *et al.*, 2021)

2. using **gene expression data only** is considered **less informative** as most of the gene seems to be correlated with other genes to perform function effectively (Wu *et al.*, 2012)

3. **class imbalance issue** in gene expression data might decrease the ability of predictive machines to predict the minority class accurately (Koziarski *et al.*, 2020)

Problem Background

Problem Statement

The focus on high dimensional **gene expression data only** will **leave over** the **interaction** between genes and further lead to the **less informative** discovery of potential biomarkers while the **imbalance class distribution** might cause the **classifier** to become **bias.**

Research Aim

to implement different data resampling strategies for identifying potential biomarkers of ovarian cancer from imbalanced gene expression with protein-protein interactions.

Research Objectives

1 To study the **existing computational methods** in identification of potential biomarkers.

2 To generate input that considers **interactions between genes** from gene expression datasets with protein-protein interactions data.

3 To apply different **data resampling strategies** on the generated input for better identification of potential biomarkers of ovarian cancer.

4 To evaluate the **performance** of different data resampling strategies in terms of accuracy, precision, sensitivity, specificity, and F1 score.

Research Scopes

The research will focus on identifying the potential biomarkers of ovarian cancer.

The dataset mainly used is gene expression data that can be downloaded from Gene Expression Omnibus (GEO).

The performance measurement will be focused on accuracy, precision, sensitivity, specificity, and F1-score.

The source of biological context verification will be based on the published journals and articles.

Summary of Literature Review

Protein-protein Interaction (PPI)

- physical interactions between two or more proteins that have complex biological activities (Farooq *et al.*, 2021).

- Interaction between proteins will control biological processes or mechanisms that can further lead to a healthy or unhealthy condition in living organisms.

- plays an important role in identifying the molecular basis of a disease. (Atan *et al.*, 2018).

Summary of Literature Review

Protein-protein Interaction (PPI)

Publication	Purpose	Advantages
Zhang *et al*., 2022	Identify biomarkers for colon cancer	Reduce the bad influences caused by the size and heterogeneity of sample
Yu *et al*., 2020	Identify biomarkers for neurodegenerative disease	Improve the chance of successfully predicting a biomarker for the new disease
Nan *et al*., 2021	Identify biomarkers for lung cancer	Helps in understanding the function and behavior of the protein

Advantages of applying PPI in identifying biomarkers

Summary of Literature Review

Data Integration

- The process of combining different data that are obtained from a few different sources into a single dataset.

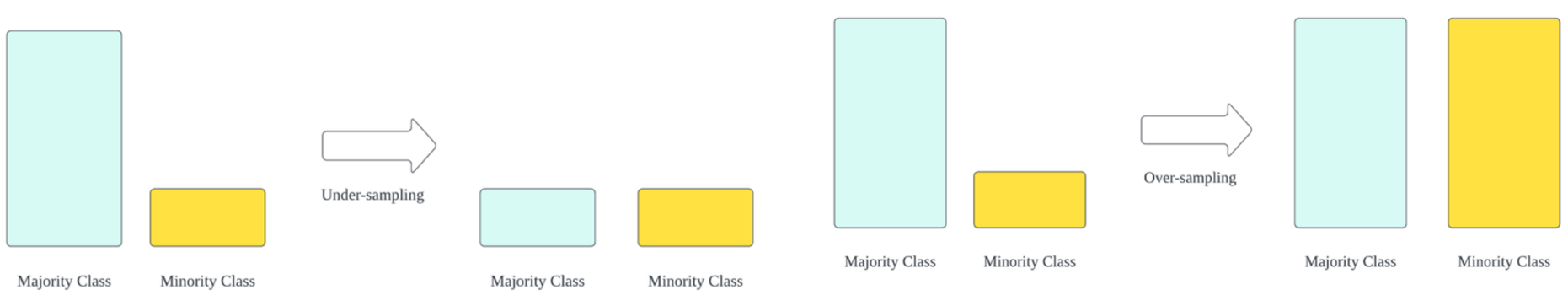
- Essential for researchers to utilize the data fully and gain more insights about biological systems (Reel *et al.*, 2021).

Publication	Dataset	Advantages	Disadvantages
Pal *et al.*, 2007; Lu *et al.*, 2014; Peng *et al.*, 2006	GE alone	Easy to understand and process	Limited information, do not considered interaction between genes
Yang *et al.*, 2018; Zeng *et al.*, 2018; Niu *et al.*, 2020	GE + PPI	Provide more information about the disease, interactions between genes are considered	Complicated, required more time to complete the whole process
Cardoso *et al.*, 2018; Li *et al.*, 2020	GE + pathway	Provide more useful insights about the biological processes	Complicated, required more time to complete the whole process

Summary of Literature Review

Data Resampling

- Applied in training data to **balance the proportion of class distribution** by decreasing the number of samples from the majority class or increasing the number of samples from the minority class (Khushi *et al.*, 2021).
- **Under-sampling** will **remove the samples from the majority classes** until it is almost the same as the number of samples from the minority classes.
- **Over-sampling** will **create new samples** according to the samples of the minority class to balance the class distribution.
- **Hybrid-sampling** combines both under-sampling and over-sampling.



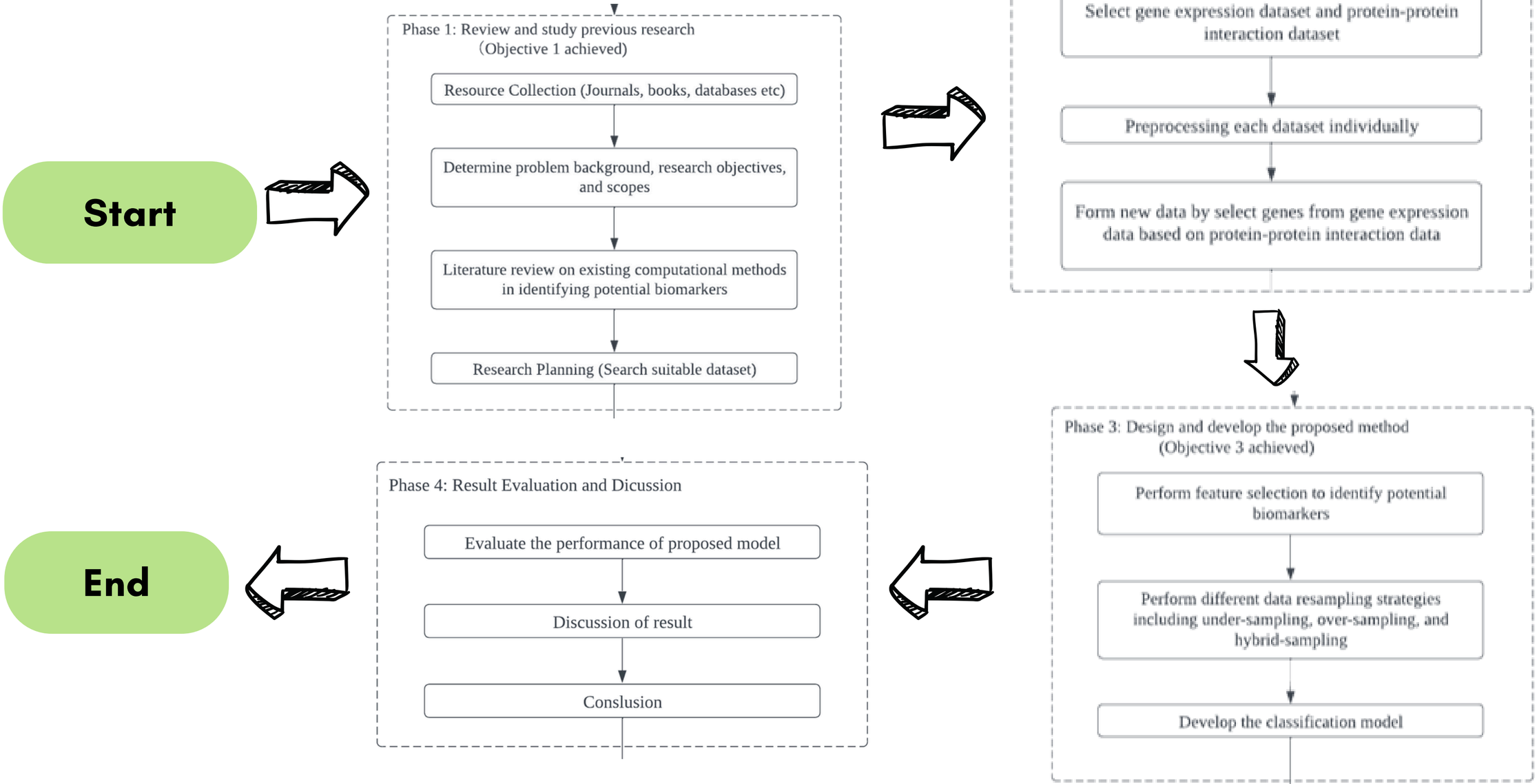
Summary of Literature Review

Publication	Computational method	Advantages	Disadvantages
Zhang *et al.*, 2021; Zhou *et al.*, 2018; Adorada *et al.*, 2018	Support Vector Machine (SVM)	• Low probability of over-fitting • Good in handling complex function	• Hard to deal with imbalanced dataset • Cannot perform well if data has noise
Toth *et al.*, 2019; Zhao *et al.*, 2019	Random Forest (RF)	• Good at handling many predictors variables • Good in dealing with imbalanced dataset	• Hard to interpret • Computationally expensive
Mofrad *et al.*, 2019; Alessandro *et al.*, 2020	Decision Tree (DT)	• Easy to interpret • Good visualization	• Prone to over-fitting • Sensitive to features
Shon *et al.*, 2019; Folego *et al.*, 2020; Abdeltawab *et al.*, 2019	Convolutional Neural Network (CNN)	• High accuracy in image classification	• Computationally expensive • Complex
Zhang *et al.*, 2021; Lu *et al.*, 2018	Deep Neural Network (DNN)	• Low error rate • Good in detecting complex relationship	• Computationally expensive • Complex

Existing computational methods and their advantages and disadvantages in identification of potential biomarkers

Summary of Research Methodology

Research Framework



Summary of Research Methodology

1

Accession Number	GSE52037	GSE10971	GSE4122	GSE6008	GSE26712
Cancerous Sample	10	13	53	99	185
Control Sample	10	24	14	4	10
Total Sample	20	37	67	103	195
Overall	422 sample				

360 cancerous sample **(85%)**
62 control sample **(15%)**

Obtained from: Gene Expression Omnibus (GEO)

Summary of Research Methodology

Dataset

2

Protein-Protein Interaction Data

Consists of 39240 interaction of genes

Obtained from: Human Protein Reference Database (HPRD)

Summary of Research Methodology

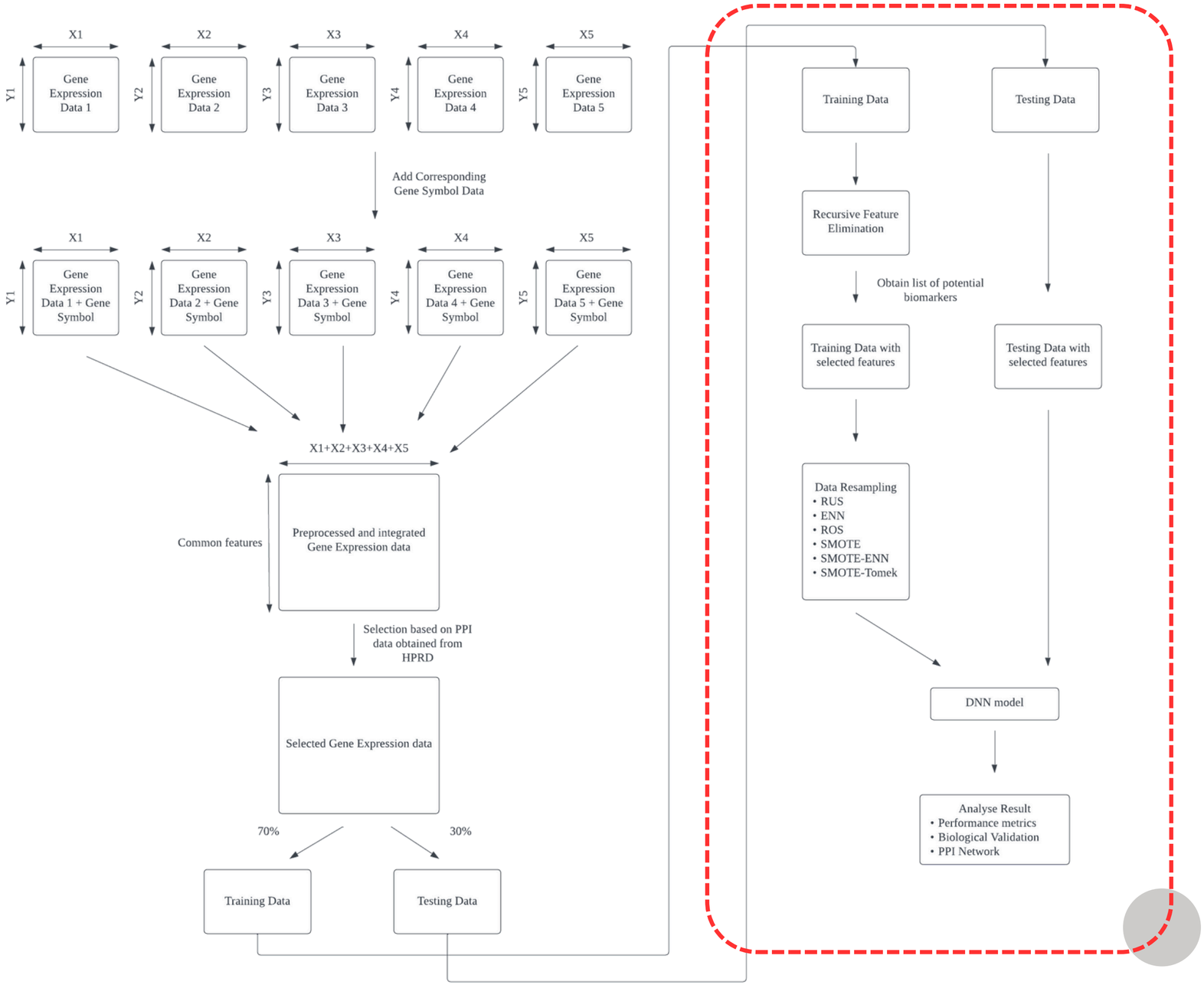
Classification Performance Measurement

Measurement	Definition	Formula
Accuracy (Acc)	It shows the value of correctly classified sample.	$\frac{TP + TN}{TP + FP + TN + FN}$
Sensitivity (Ss)	It shows the portion of positive samples that are correctly predicted.	$\frac{TP}{TP + FN}$
Precision (Pre)	It shows the quality of positive sample predicted.	$\frac{TP}{TP + FP}$
Specificity (Sp)	It shows the portion of negative samples that are correctly predicted.	$\frac{TN}{TN + FP}$
F1 score	The harmonic mean of Pre and Ss.	$\frac{2 \times Pre \times Ss}{Pre + Ss}$

Biological Context Validation

Validate the potential biomarkers from reliable sources to prove it is related to ovarian cancer.

Summary of Research Design and Implementation



Data Pre-processing

Gene expression data in matrix form

ID_REF	GSM139377	GSM139378	GSM139379	GSM139380
1007_s_at	3.702688968	3.851686315	3.667172672	4.138965478
1053_at	2.704150517	2.557507202	2.382017043	2.816903839
117_at	2.853698212	2.644438589	2.73239376	2.937016107
121_at	4.124014879	4.167760266	4.139532772	4.376887057
1255_g_at	2.492760389	2.531478917	2.57054294	2.586587305
1294_at	3.240299582	3.089198367	3.202215776	3.15715444

+

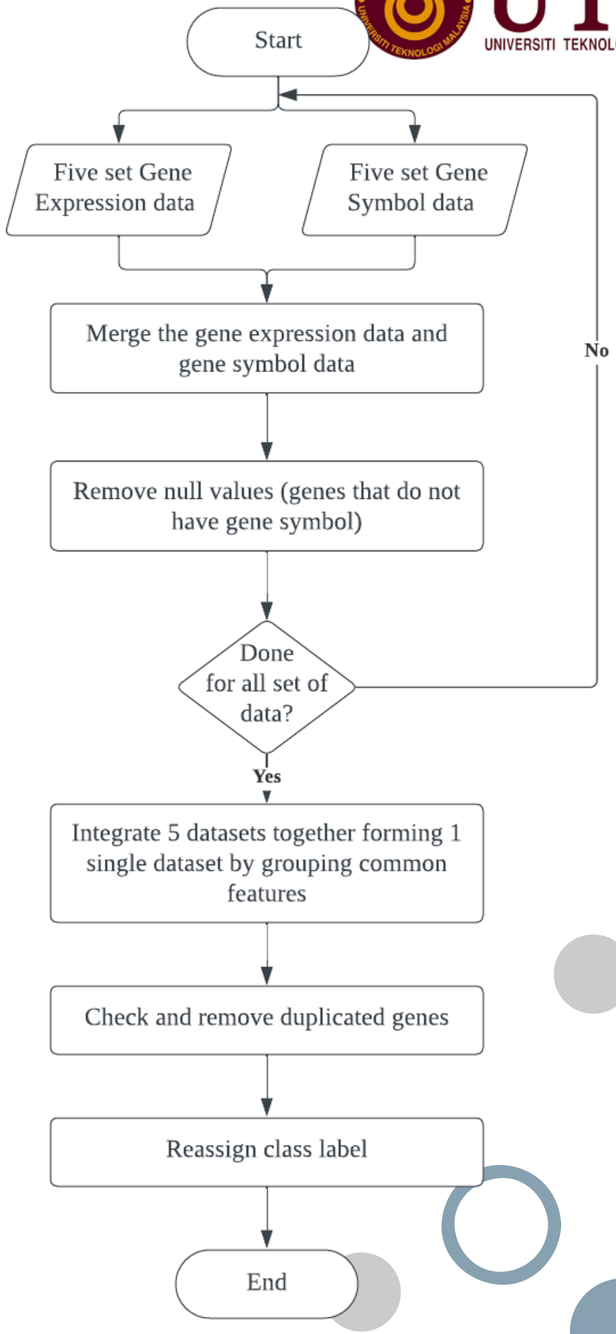
Gene Symbol data

ID_REF	Gene Symbol
1007_s_at	DDR1 /// MIR4640
1053_at	RFC2
117_at	HSPA6
121_at	PAX8
1255_g_at	GUCA1A
1294_at	MIR5193 /// UBA7
1316_at	THRA
1320_at	PTPN21

Remove missing values (genes that do not have gene symbol)

ID_REF	Gene Symbol	GSM139377	GSM139378	GSM139379	GSM139380	GSM139381	GSM139382
1007_s_at	DDR1 /// MIR4	3.702688968	3.851686315	3.667172672	4.138965478	3.835500328	3.392872745
1053_at	RFC2	2.704150517	2.557507202	2.382017043	2.816903839	2.88592634	2.630427875
117_at	HSPA6	2.853698212	2.644438589	2.73239376	2.937016107	2.828659897	2.855519156
121_at	PAX8	4.124014879	4.167760266	4.139532772	4.376887057	4.244821195	4.142702246
1255_g_at	GUCA1A	2.492760389	2.531478917	2.57054294	2.586587305	2.542825427	2.526339277
1294_at	MIR5193 /// U	3.240299582	3.089198367	3.202215776	3.15715444	3.260548373	3.01494035
1316_at	THRA	2.777426822	3.017867719	2.797959644	2.733197265	2.866877814	2.654176542
1320_at	PTPN21	2.487138375	2.57054294	2.480006943	2.492760389	2.392696953	2.307496038

Gene symbol is added as it is needed for selecting the genes based on PPI data



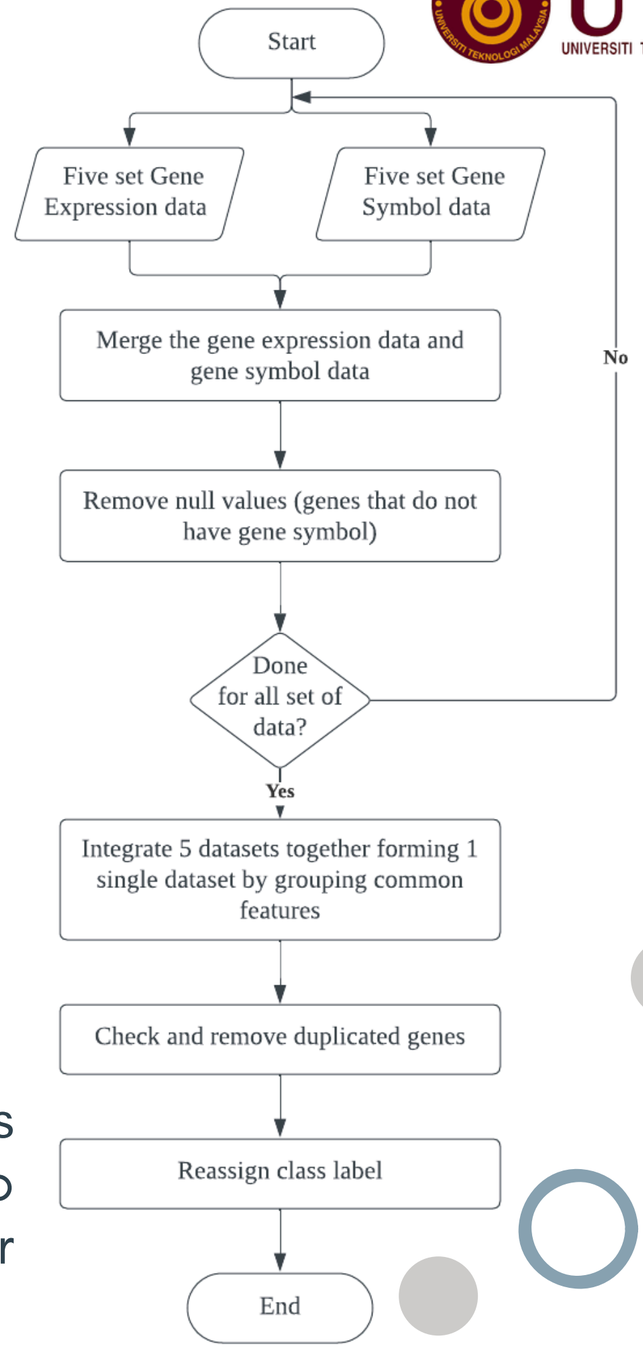
Data Pre-processing

Common Features

64432_at	MAPKAPK5	2.30103	2.499687	2.456366	2.721811	2.584331
65517_at	AP1M2	3.317854	3.215902	3.156549	3.359646	3.268812
65521_at	UBE2D4	3.2266	3.330819	3.142389	3.303844	3.2403
65884_at	MAN1B1	3.028571	3.004751	3.130334	3.111599	3.161368
78330_at	ZNF335	2.371068	2.798651	2.359835	2.660865	2.612784
823_at	CX3CL1	2.828015	2.800029	2.689309	3.025306	2.876218
90265_at	ADAP1	3.559548	3.513617	3.297323	3.734079	3.741546
90610_at	LRCH4	3.068557	3.0306	3.036629	3.037825	3.084576
91920_at	BCAN	3.140508	3.220108	3.269746	3.270912	3.216957
Class		1	1	1	1	1

Integrate 5 GE dataset

- only the common features (genes) will remain
- 8250 common features remain after the process
- Remove duplicated genes (only 8134 genes remaining)
- Combined as the **sample size** of each dataset is **small** and the number of normal control samples is too small. Wong *et al*. (2022) also used this method in their research to identify potential biomarkers.



Select Gene Expression Data Based on PPI Data

GE data

Gene Symbol	Sample#1	Sample#2	Sample#3
Gene#1			
Gene#2			
Gene#3			
Gene#4			
Gene#5			

+

PPI data

Gene Symbol 1	Gene Symbol 2
Gene#1	Gene#2
Gene#4	Gene#6

⬇ **Eliminate genes do not exist in PPI data**

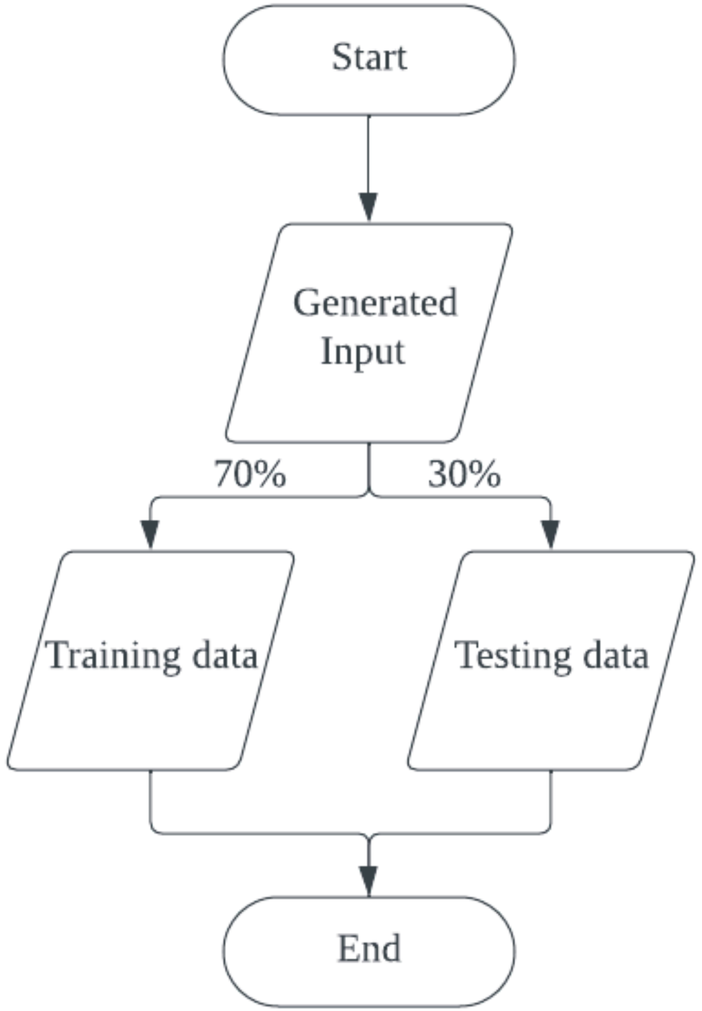
New GE data

Gene Symbol	Sample#1	Sample#2	Sample#3
Gene#1			
Gene#2			
Gene#4			

5729 features (genes) remain after the process

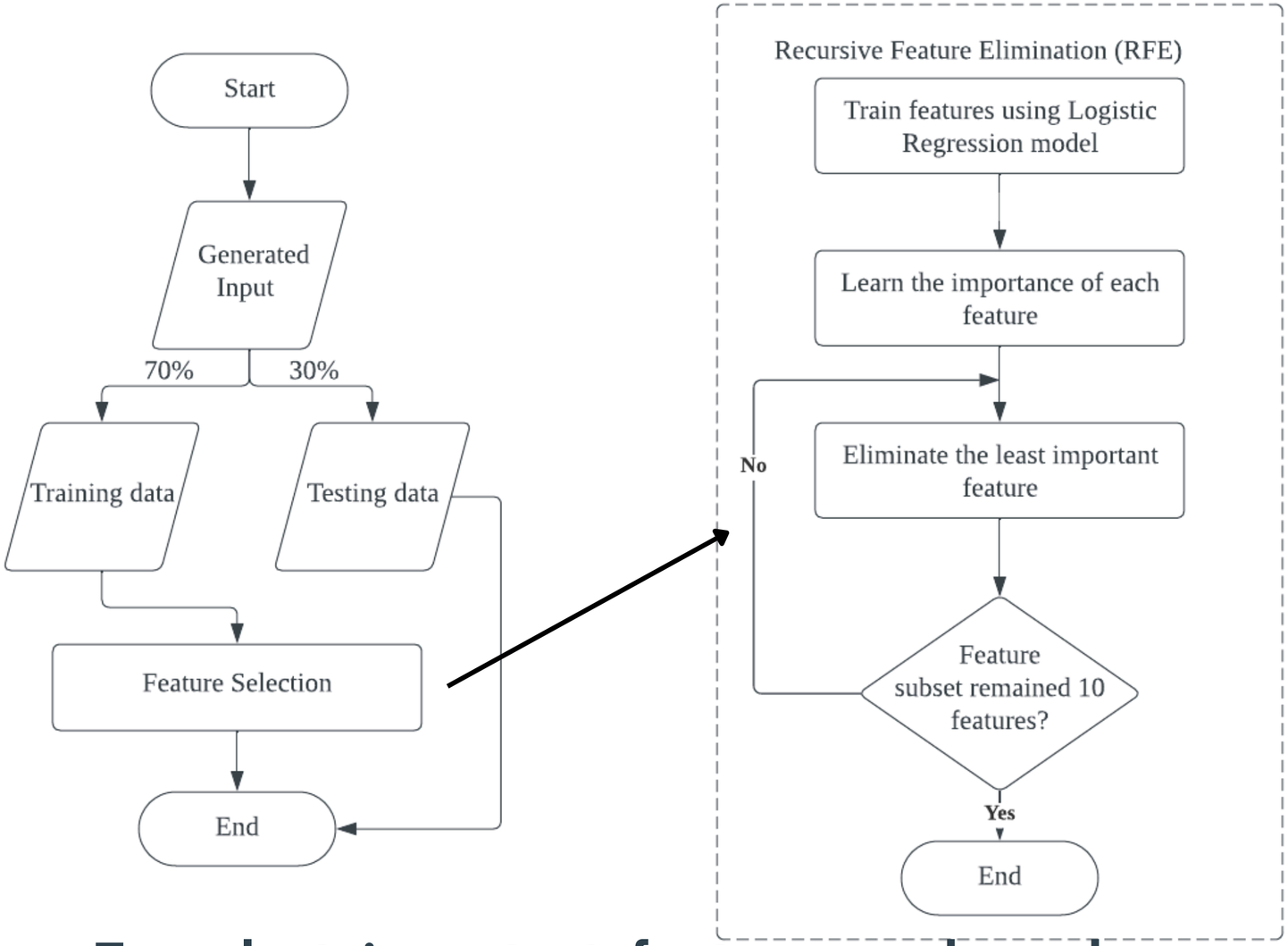
To ensure only genes that have interaction with other genes will be remained

Data Splitting



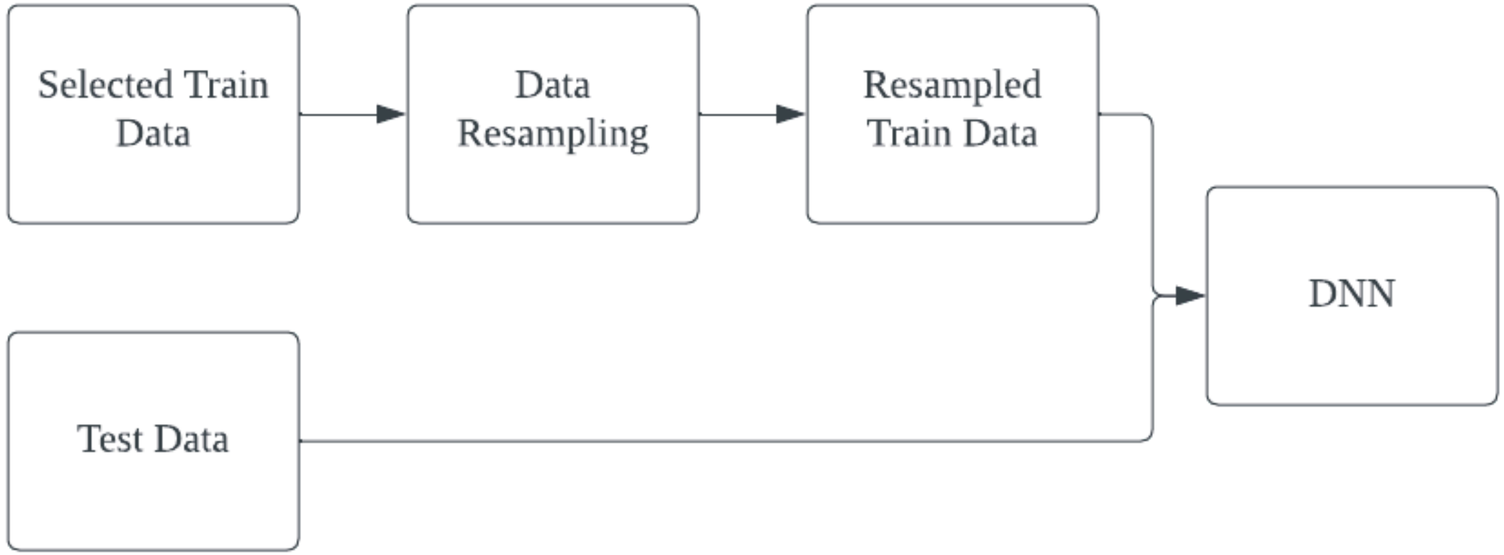
- The random state hyperparameter is set to ensure the data subsets produced are **reproducible**.
- This experiment uses **five different random states** to run the experiment five times to ensure the result of the experiment is not biased.

Feature Selection



To select important features and produce list of potential biomarker

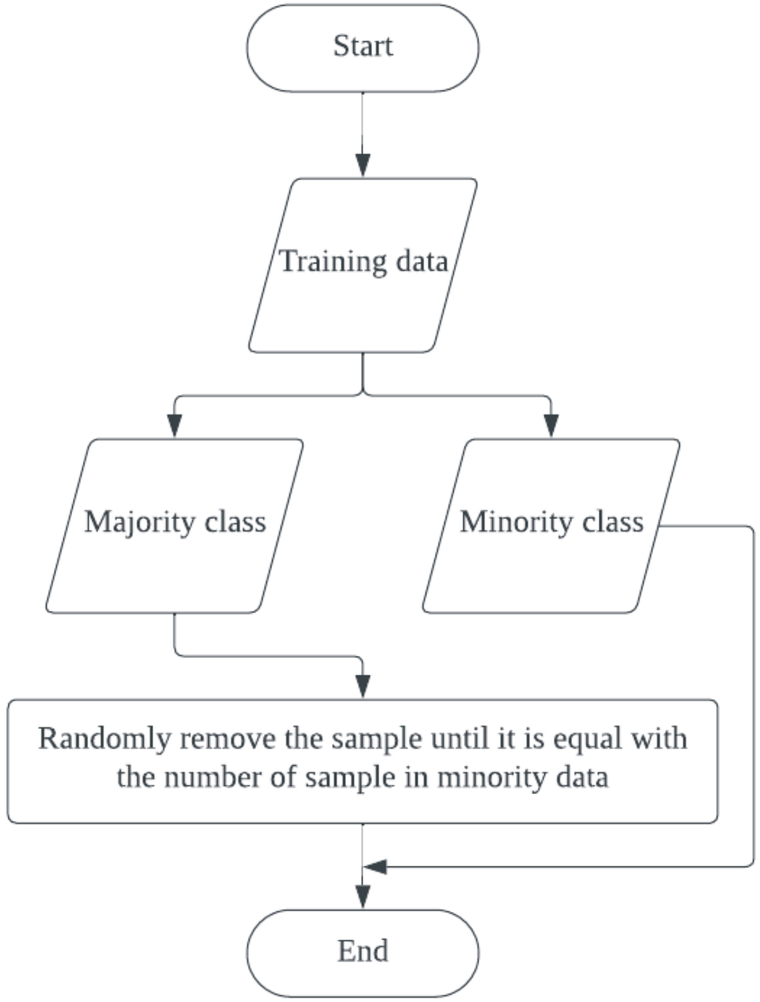
Data Resampling



- **Random Under-Sampling (RUS)**
- **Edited Nearest Neighbors (ENN)**
- **Random Over-Sampling (ROS)**
- **Synthetic Minority Over-Sampling Technique (SMOTE)**
- **SMOTE-RUS**
- **SMOTE-Tomek Link**

To balance the class distribution of gene expression data

Random Under-Sampling (RUS)



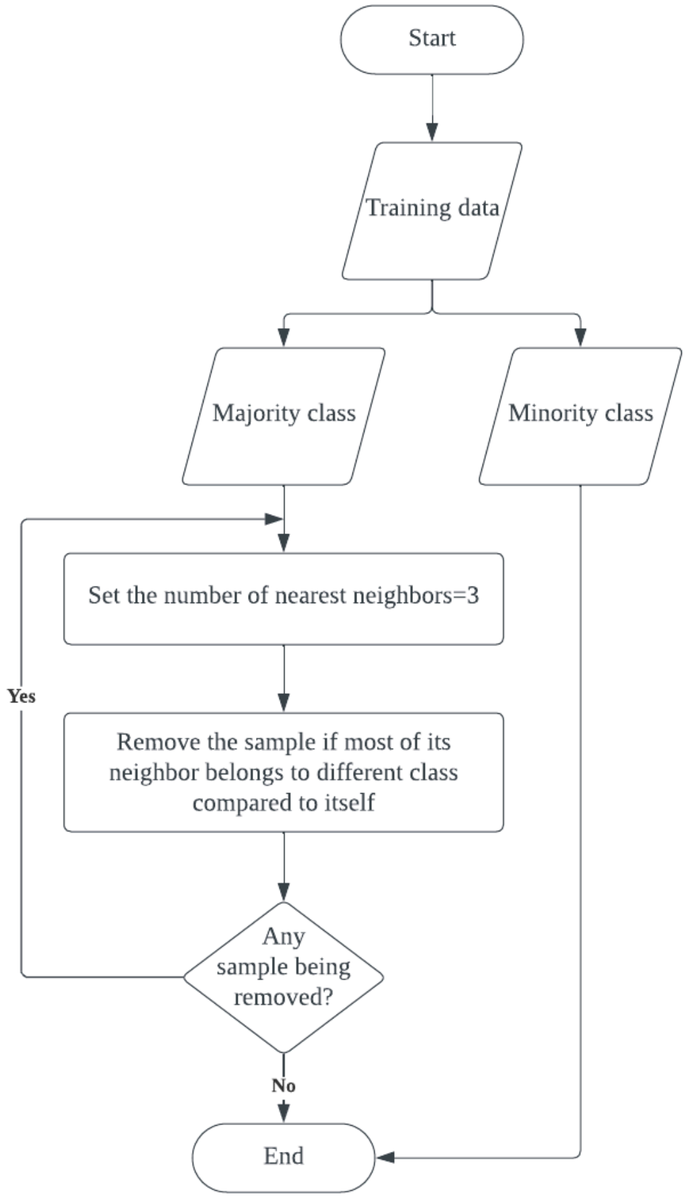
Before RUS

Normal Control Sample	42	minority
Cancerous Sample	253	majority

After RUS

Normal Control Sample	42	minority
Cancerous Sample	42	majority

Edited Nearest Neighbors (ENN)



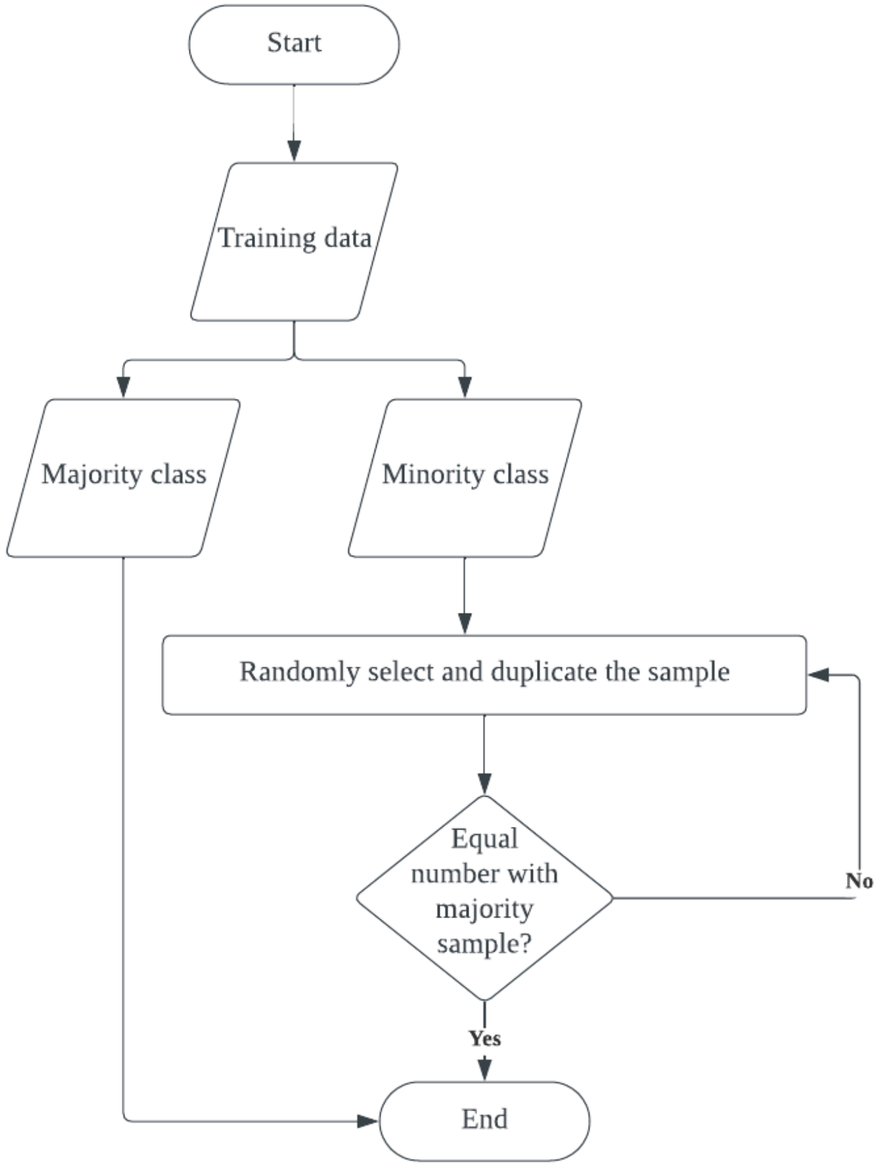
Before ENN

Normal Control Sample	42	minority
Cancerous Sample	253	majority

After ENN

Normal Control Sample	42	minority
Cancerous Sample	240	majority

Random Over-Sampling (ROS)



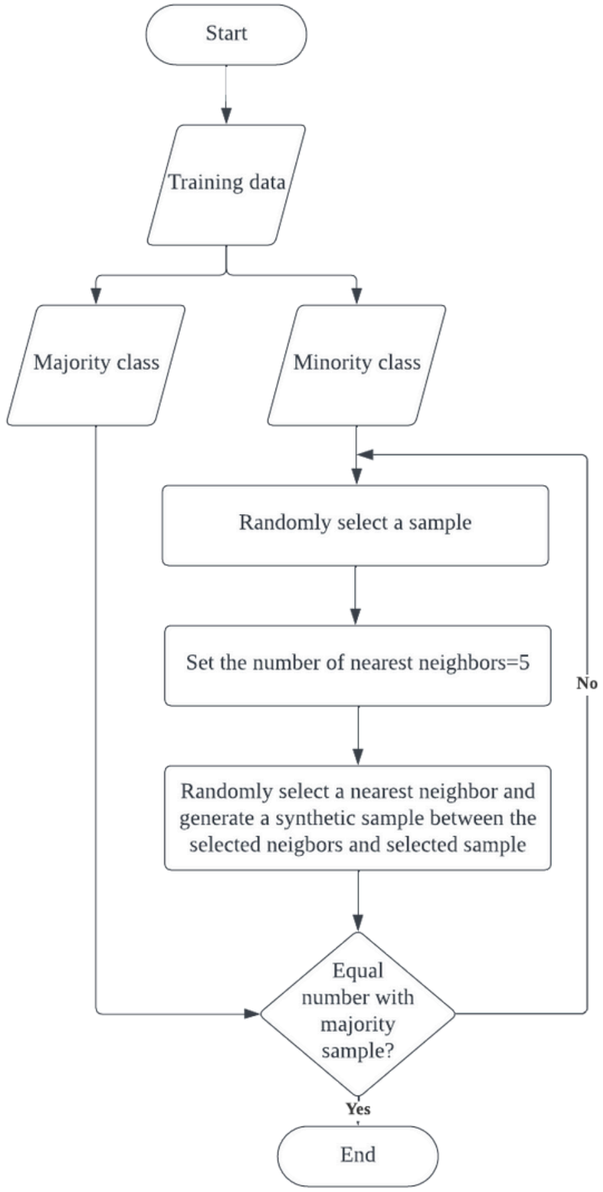
Before ROS

Normal Control Sample	42	minority
Cancerous Sample	253	majority

After ROS

Normal Control Sample	253	minority
Cancerous Sample	253	majority

Synthetic Minority Over-Sampling Technique (SMOTE)



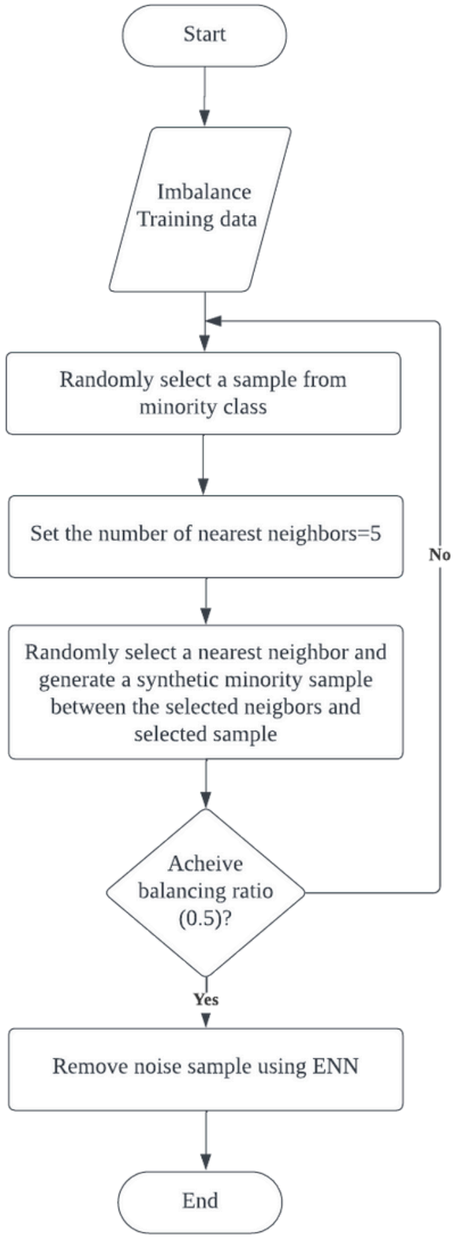
Before SMOTE

Normal Control Sample	42	minority
Cancerous Sample	253	majority

After SMOTE

Normal Control Sample	253	minority
Cancerous Sample	253	majority

SMOTE - ENN



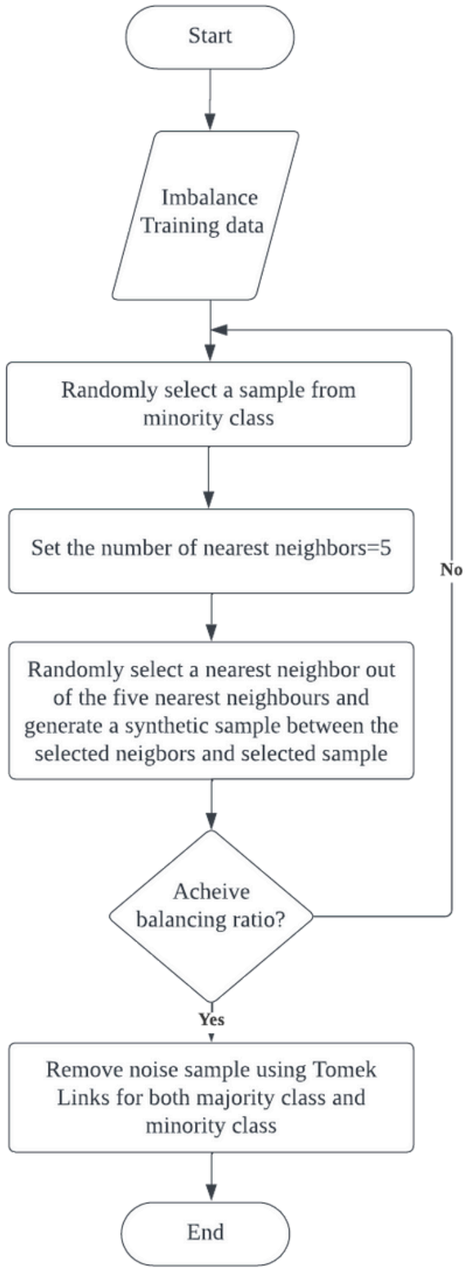
Before SMOTE-ENN

Normal Control Sample	42	minority
Cancerous Sample	253	majority

After SMOTE-ENN

Normal Control Sample	247	minority
Cancerous Sample	239	majority

SMOTE - Tomek Link



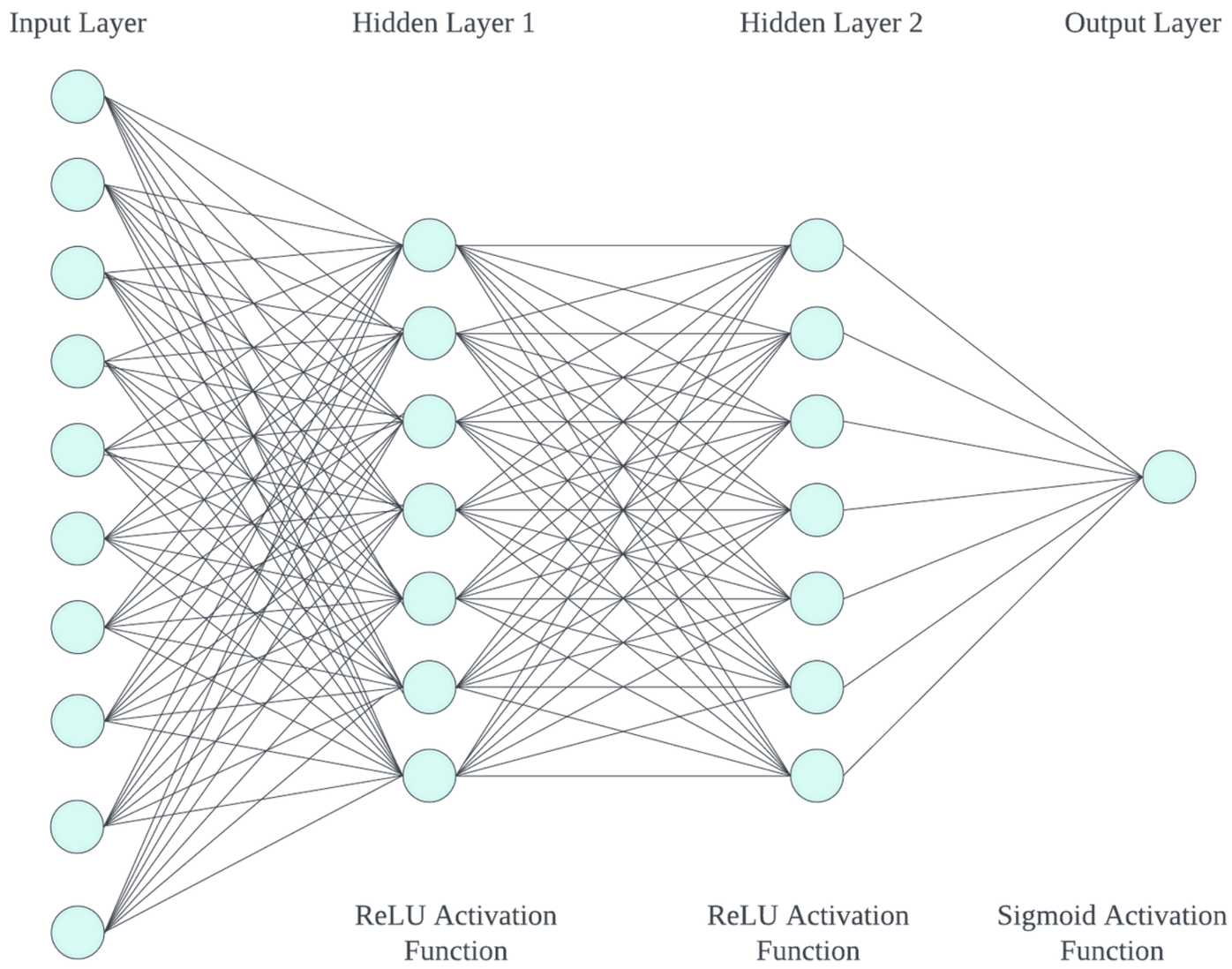
Before SMOTE-Tomek Link

Normal Control Sample	42	minority
Cancerous Sample	253	majority

After SMOTE-Tomek Link

Normal Control Sample	252	minority
Cancerous Sample	252	majority

DNN Model



Epochs:250
Batch size:32

Effect of Integrating GE Data with PPI Data

Data	Runs	Accuracy (Acc)	Specificity (Sp)	Sensitivity (Ss)	Precision (Pre)	F1 score
GE Only	1	90.55%	29.41%	98.18%	90.00%	93.91%
	2	88.98%	26.32%	100.00%	88.52%	93.91%
	3	90.55%	36.84%	100.00%	90.00%	94.74%
	4	89.76%	35.00%	100.00%	89.17%	94.27%
	5	91.34%	35.29%	100.00%	90.91%	95.24%
Average		**90.24%**	**32.57%**	**99.64%**	**89.72%**	**94.41%**
GE + PPI	1	86.02%	58.82%	97.27%	93.86%	95.54%
	2	90.55%	57.89%	96.30%	92.86%	94.55%
	3	89.91%	42.11%	100.00%	90.76%	95.15%
	4	95.28%	70.00%	100.00%	94.69%	97.27%
	5	91.34%	35.29%	100.00%	90.91%	95.24%
Average		**90.62%**	**52.82%**	**98.71%**	**92.62%**	**95.55%**

There is a significant **increase** in the **specificity** by integrating GE data with PPI data. This shows that **more normal control samples are predicted correctly** although it is the **minority** class.

Effect of Feature Selection

Method	Runs	Accuracy (Acc)	Specificity (Sp)	Sensitivity (Ss)	Precision (Pre)	F1 score
GE + PPI	1	86.02%	58.82%	97.27%	93.86%	95.54%
	2	90.55%	57.89%	96.30%	92.86%	94.55%
	3	89.91%	42.11%	100.00%	90.76%	95.15%
	4	95.28%	70.00%	100.00%	94.69%	97.27%
	5	91.34%	35.29%	100.00%	90.91%	95.24%
Average		**90.62%**	**52.82%**	**98.71%**	**92.62%**	**95.55%**
GE + PPI + RFE	1	88.98%	58.82%	93.64%	93.64%	93.64%
	2	90.55%	73.68%	93.52%	95.28%	94.39%
	3	94.49%	73.68%	98.15%	95.50%	96.80%
	4	92.13%	65.00%	97.20%	93.69%	95.41%
	5	88.19%	58.82%	92.73%	93.58%	93.15%
Average		**90.87%**	**66.00%**	**95.05%**	**94.34%**	**94.68%**

The **specificity** has improved around 13%. This indicates that through the application of RFE, more significant features are identified and the **noisy features remaining from the previous process are removed** now.

Effect of Data Resampling

	Method	AVERAGE				
		Accuracy (Acc)	Specificity (Sp)	Sensitivity (Ss)	Precision (Pre)	F1 score
	Without Data Resampling	90.87%	66.00%	95.05%	94.34%	94.68%
Under-sampling	RUS	87.40%	80.15%	89.89%	96.47%	92.99%
	ENN	90.39%	80.32%	92.08%	94.29%	94.21%
Over-sampling	ROS	93.54%	93.54%	93.57%	98.86%	95.85%
	SMOTE	94.65%	87.96%	95.77%	97.96%	96.83%
Hybrid-sampling	SMOTE-ENN	93.70%	90.24%	94.31%	98.31%	96.24%
	SMOTE-Tomek	95.12%	90.14%	95.96%	98.32%	97.11%

The implementation of data resampling methods has successfully **solved the class imbalance issue** and led to the yield of improved results. Overall, the **specificity shows huge improvement** after applying the data resampling method.

Comparison of Under-Sampling Results

Method	Runs	Accuracy (Acc)	Specificity (Sp)	Sensitivity (Ss)	Precision (Pre)	F1 score
RUS	1	88.98%	76.47%	90.91%	96.15%	93.46%
	2	84.25%	89.47%	83.33%	97.83%	90.00%
	3	84.25%	84.21%	90.74%	97.03%	93.78%
	4	95.28%	80.00%	98.13%	96.33%	97.22%
	5	84.25%	70.59%	86.36%	95.00%	90.48%
Average		**87.40%**	**80.15%**	**89.89%**	**96.47%**	**92.99%**
ENN	1	90.55%	82.35%	91.82%	97.12%	94.39%
	2	86.61%	84.21%	87.04%	96.91%	91.71%
	3	88.98%	89.47%	88.88%	97.96%	93.20%
	4	93.70%	75.00%	97.20%	95.41%	96.30%
	5	92.13%	70.59%	95.45%	95.45%	95.45%
Average		**90.39%**	**80.32%**	**92.08%**	**94.29%**	**94.21%**

ENN has **strong overall performance** and reliability in identifying both normal control samples and ovarian cancer samples despite its slightly lower precision

Comparison of Over-Sampling Results

Method	Runs	Accuracy (Acc)	Specificity (Sp)	Sensitivity (Ss)	Precision (Pre)	F1 score
ROS	1	95.28%	94.12%	95.45%	99.06%	97.22%
	2	90.55%	100.00%	88.88%	100.00%	94.12%
	3	96.06%	89.47%	97.22%	98.13%	96.47%
	4	96.07%	90.00%	97.20%	98.11%	97.65%
	5	89.76%	94.12%	89.09%	98.99%	93.78%
Average		**93.54%**	**93.54%**	**93.57%**	**98.86%**	**95.85%**
SMOTE	1	94.49%	94.12%	94.55%	99.05%	96.74%
	2	92.91%	89.47%	93.52%	98.06%	95.73%
	3	94.49%	94.74%	94.44%	99.03%	96.68%
	4	96.85%	85.00%	99.07%	97.25%	98.15%
	5	94.49%	76.47%	97.27%	96.40%	96.83%
Average		**94.65%**	**87.96%**	**95.77%**	**97.96%**	**96.83%**

SMOTE perform well overall and is **good at identifying ovarian cancer samples**. However, the low specificity of SMOTE caused **some of the normal control samples to be classified incorrectly** and led to a lower reliability in the negative predictions.

Comparison of Hybrid-Sampling Results

Method	Runs	Accuracy (Acc)	Specificity (Sp)	Sensitivity (Ss)	Precision (Pre)	F1 score
SMOTE-ENN	1	92.91%	94.12%	92.73%	99.03%	95.77%
	2	93.70%	100.00%	92.59%	100.00%	96.15%
	3	94.49%	94.74%	94.44%	99.03%	96.68%
	4	95.28%	80.00%	98.13%	96.33%	97.22%
	5	92.13%	82.35%	93.64%	97.17%	95.37%
Average		**93.70%**	**90.24%**	**94.31%**	**98.31%**	**96.24%**
SMOTE-Tomek	1	94.49%	88.24%	95.45%	98.13%	96.77%
	2	95.28%	100.00%	94.44%	100.00%	97.14%
	3	96.06%	84.21%	98.13%	97.25%	97.70%
	4	96.85%	90.00%	98.15%	98.13%	98.13%
	5	92.91%	88.24%	93.64%	98.10%	95.81%
Average		**95.12%**	**90.14%**	**95.96%**	**98.32%**	**97.11%**

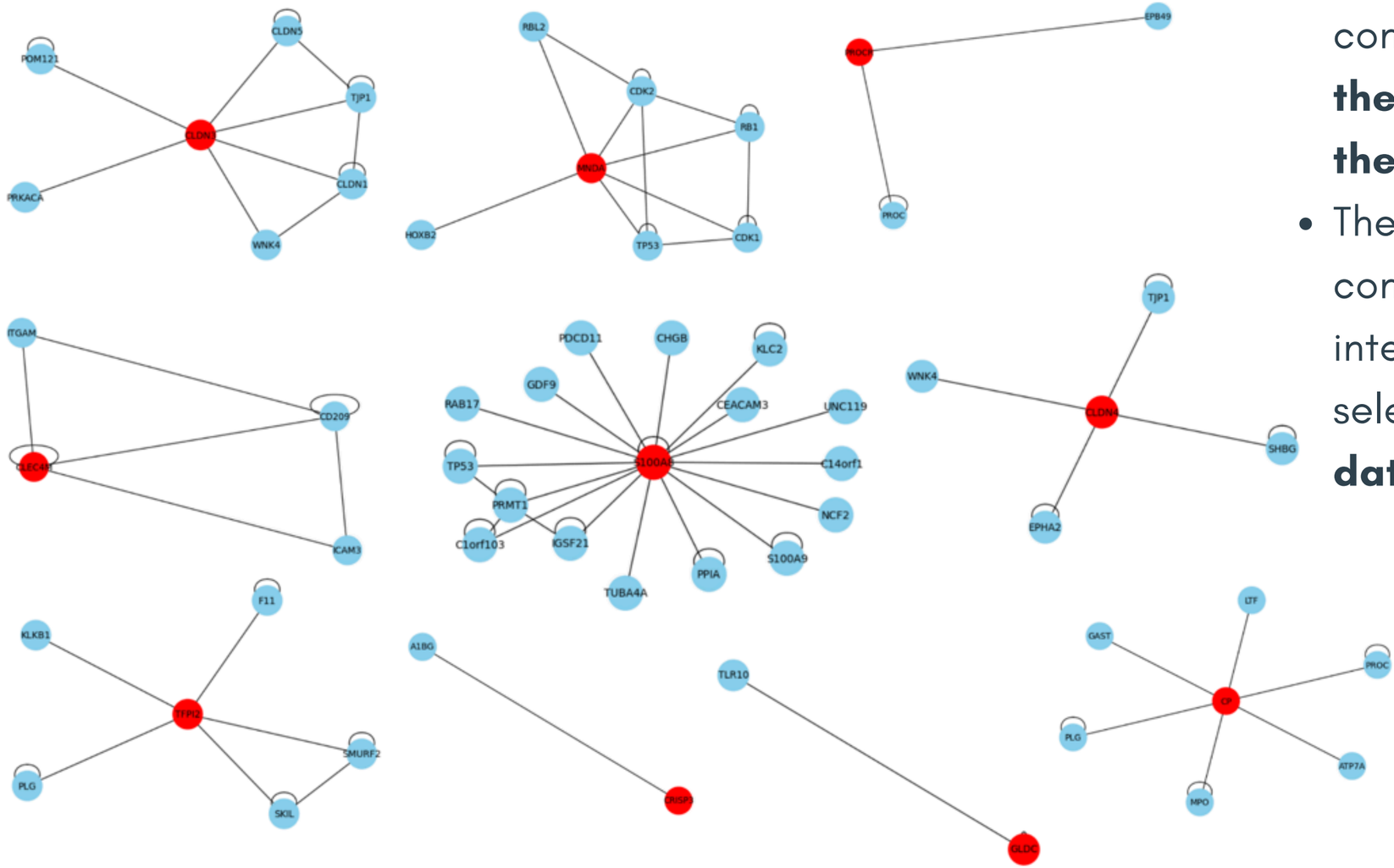
SMOTE-Tomek **outperforms** overall, especially in **identifying** the **ovarian cancer samples**. However, its lower specificity indicates that its capability to identify the normal control samples is lower than the SMOTE-ENN

Biological Context Validation

Gene Symbol	Gene Name	Descriptions	Publications
CLDN3	Claudin 3	positively correlated with the development of ovarian cancer	Hao *et al*., 2023; Yuan *et al*., 2020
MNDA	Myeloid Cell Nuclear Differentiation Antigen	-	-
CLEC4M	C-Type Lectin Domain Family 4 Member M	play an important role in oncogenesis in ovarian cancer	Wang *et al*., 2024; Li *et al*., 2022
S100A8	Calcium-binding Protein A8	prognostic biomarker of ovarian cancer	Muqaku *et al*.,2020; Xu *et al*., 2020
CP	Ceruloplasmin	ovarian cancer patient usually has a higher level of CP	Trifanescu *et al*., 2023; Chen *et al*., 2021
CRISP3	Cysteine Rich Secretory Protein 3	secreted at increased levels in women with ovarian tumours and cancer	Gasiorowska *et al*., 2018; Yu *et al*., 2022
PROCR	Protein C Receptor	increased level of PROCR caused poor prognosis for ovarian cancer patients	Yuan *et al*., 2020; Torabian *et al*., 2023
TFPI2	Tissue Factor Pathway Inhibitor 2	preoperative biomarker for ovarian cancer	Li *et al*., 2023; Miyagi *et al*., 2021
CLDN4	Claudin 4	predictive biomarker of ovarian cancer	Hu *et al*., 2023; Wang *et al*., 2021

9 out of 10 of the selected features have been proved their relatedness with ovarian cancer

PPI Network Diagram



- PPI network diagrams are constructed **to visualise the selected features on their interacting genes.**
- The network diagrams are constructed based on the interaction of the selected features with the **data in HPRD**.

Conclusion

UTM
UNIVERSITI TEKNOLOGI MALAYSIA

Research Outcomes

1 Findings from PPI

The integration of **PPI data with GE data improve** the result of classifier in comparison with using GE data only

2 Findings from Resampling

SMOTE-Tomek performs the **best** out of all of the data resampling strategies

3 Findings from Biological Context Validation

Obtain **9 verify potential ovarian biomarkers**

Conclusion

Suggestions for Improvement and Future Work

Explore **different feature selection** to see the relations within feature selection and data resampling in dealing with imbalanced data.

Use a **larger** imbalanced GE data with more samples

PPI network can be **further analysed** in detail whether their interactions with those genes will lead to ovarian cancer

PSM 1 PRESENTATION

THANKS

THANK YOU