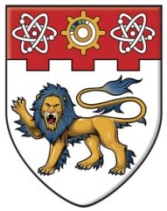


NANYANG TECHNOLOGICAL UNIVERSITY

Unlocking the prognostic potential of blood-based gene expression data for Alzheimer's Disease

Center for Biomedical Informatics
Lee Kong Chian School of Medicine
Nanyang Technological University

Han Wenhao

May/23/2023

Unlocking the prognostic potential of blood-based gene expression data for Alzheimer's Disease

Table of content

NANYANG TECHNOLOGICAL UNIVERSITY

Introduction

Dementia and Alzheimer's Disease

- Dementia is a cognitive impairment disease that mainly exists in middle-aged and elderly people.
- Alzheimer‘s disease is the most common type of dementia. It is a progressive disease beginning with mild memory loss and possibly leading to loss of the ability to carry on a conversation and respond to the environment.
- The disease progressively worsens with age, and early diagnosis is critical to its prevention.

Introduction

The significance of the project

AI-driven Clinical Diagnosis for Dementia

- Developing novel AI-driven pipeline for dementia diagnosis
- Evaluating the prognostic potential of blood-based gene expression data for dementia diagnosis

Dementia Biomarker identification

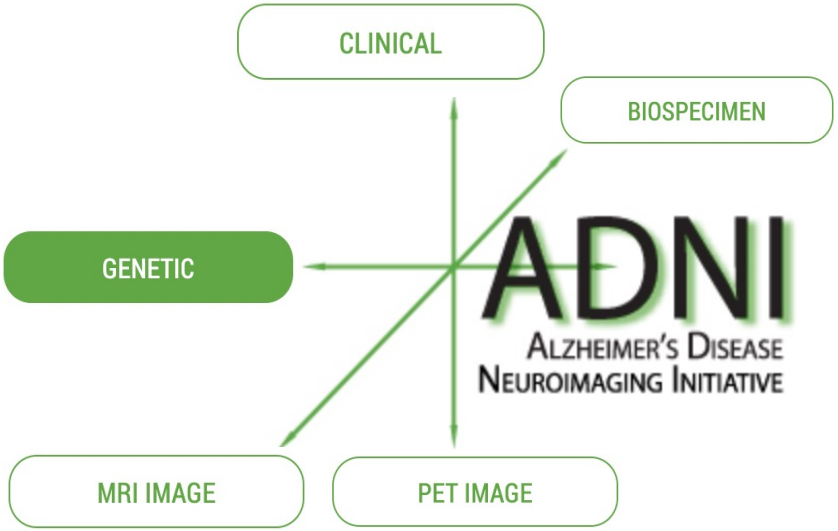
- Reveal the pathogenic factors of dementia
- Understand the onset and progression of dementia
- Identify new diagnostic and prognostic biomarkers

Material

ADNI (Alzheimer's Disease Neuroimaging Initiative)

- ADNI enrolls participants between the ages of 55 and 90 who are recruited at 57 sites in the United States and Canada.
- It is a multimodal dataset including a clinical evaluation, neuropsychological tests, genetic testing, lumbar puncture, and MRI and PET scans.
- This project will focus on the **Microarray gene expression of ADNI participants.**

	ADNI WGS
Genotyping Platform	Illumina Omni 2.5M (WGS Platform)
Number of SNPs	#SNPs: ~3.7 million #Indels: ~700,000 #SVs: ~3,500
Patients Diagnosis Groups	NC, MCI, AD
Number of Subjects	818
File format	VCF (version 4.1)



NANYANG TECHNOLOGICAL UNIVERSITY

Material

ADNI (Alzheimer's Disease Neuroimaging Initiative)

- The Microarray gene expression data of 811 ADNI participants from the ADNI WGS cohort are applied
- 67 samples didn't pass quality control.
- 14 samples without final diagnosis were excluded

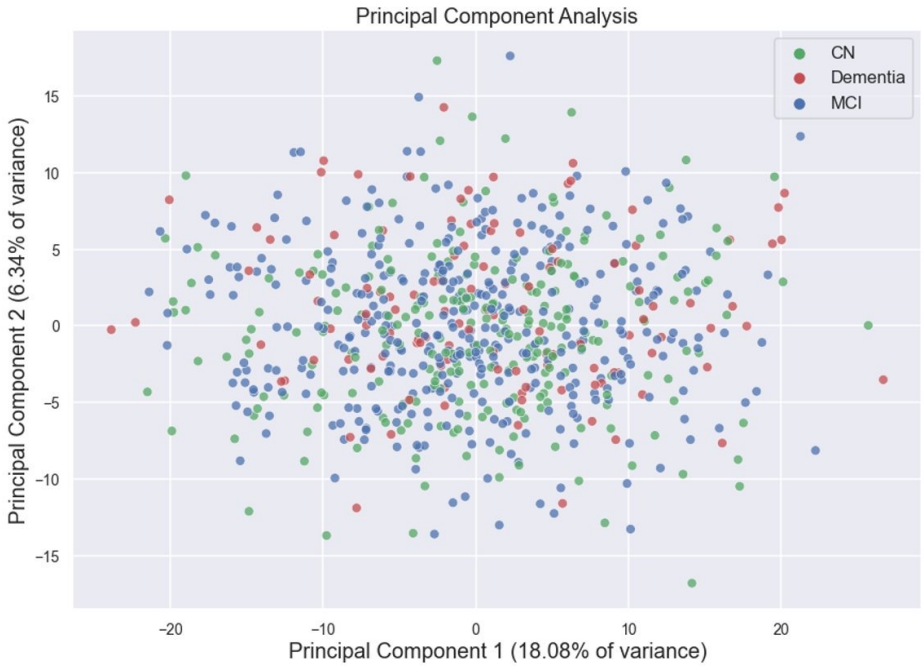
Table 1 Class distribution of Dataset

Class label	Size	Percentage
CN	235	32%
MCI	285	39%
Dementia	210	29%
total	730	100%

NC: Cognitively normal
MCI: Mild Cognitive Impairment

- RNA Integrity is main source of variance
- The effects was corrected by limma R. (removeBatchEffect)

- Dimension Reduction (PCA)



NANYANG TECHNOLOGICAL UNIVERSITY

Material

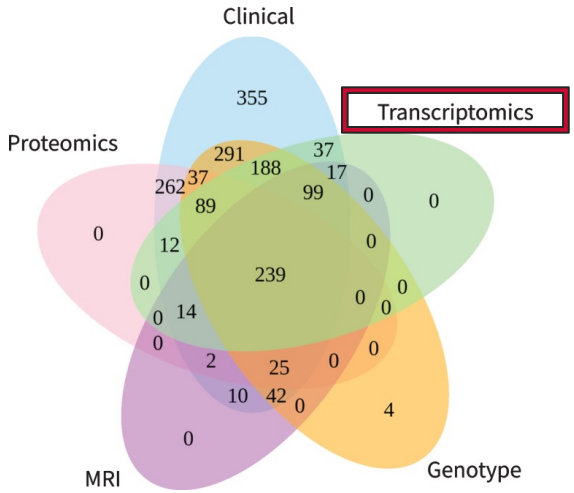
ANMERGE

- ANMerge is a highly preprocessed, multimodal, patient-level AD cohort dataset with the aim to discover AD biomarkers.

- ANMerge can serve as a discovery and validation cohort for data-driven AD research, for example, machine learning and AI approaches.

- Number of assessed variables and participants per modality subtables

Modality	Participants	Variables
Clinical	1,702	40
Proteomics	680	1,016
MRI	453	136
Gene expression	709	56,701
Genotype	1,014	789,470

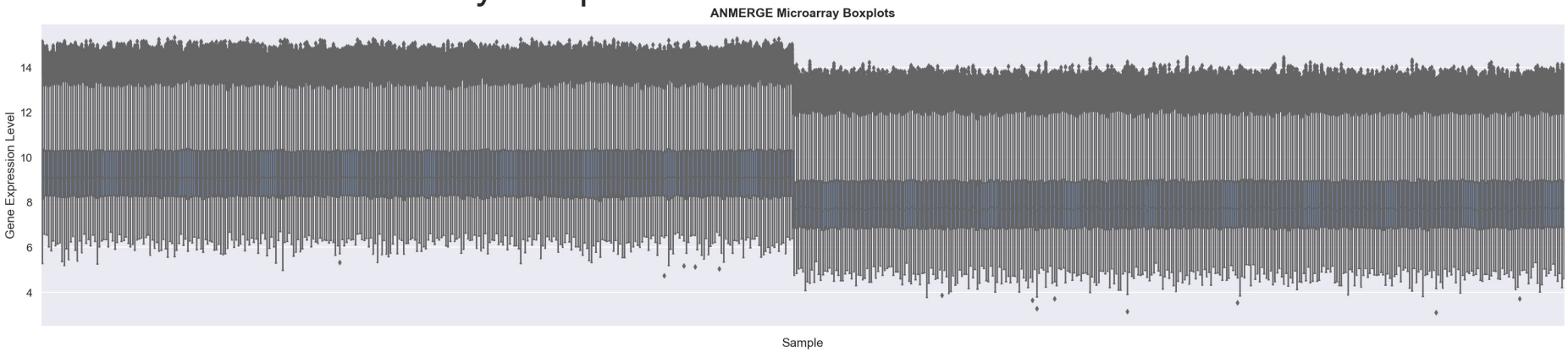


NANYANG TECHNOLOGICAL UNIVERSITY

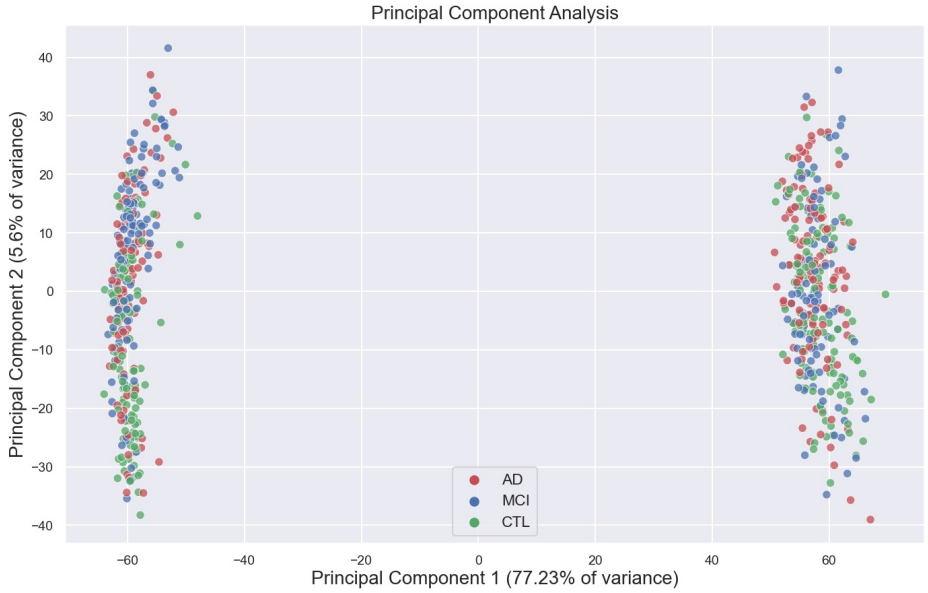
Material

ANMERGE: Data exploration

- ANMERGE Microarray Boxplot



- Dimension Reduction (PCA)



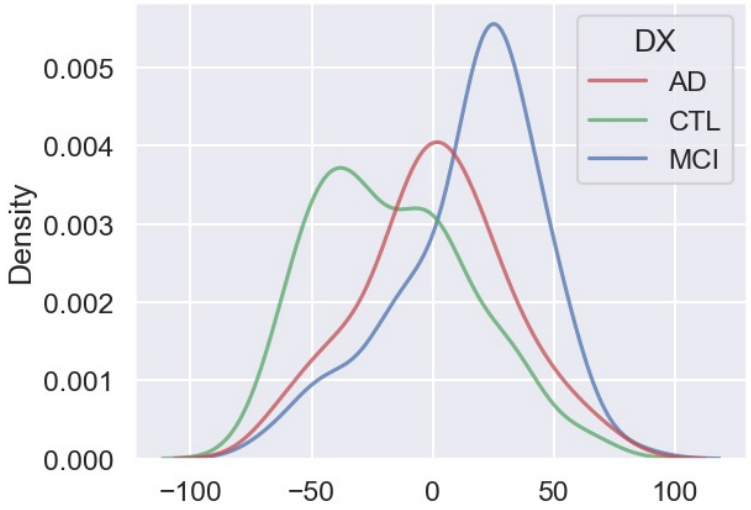
- Illumina Human HT-12 Expression BeadChips were used to analyze the whole transcriptome.
- The version of BeadChips applied for gene expression is the main source of variance. (Batch 1: BeadChip V3, Batch 2: BeadChip V4)

NANYANG TECHNOLOGICAL UNIVERSITY

Material

ANMERGE: Data exploration

- The distribution of class variance (Batch 1)



- The distribution of class variance (Batch 2)

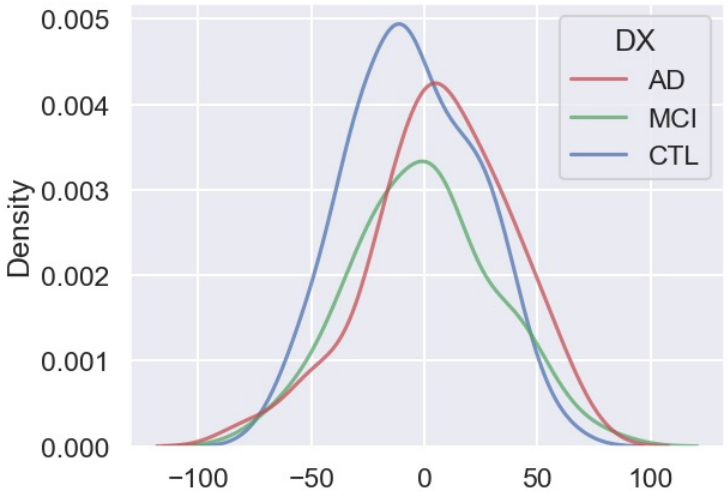


Table 3 Batch 1 class distribution

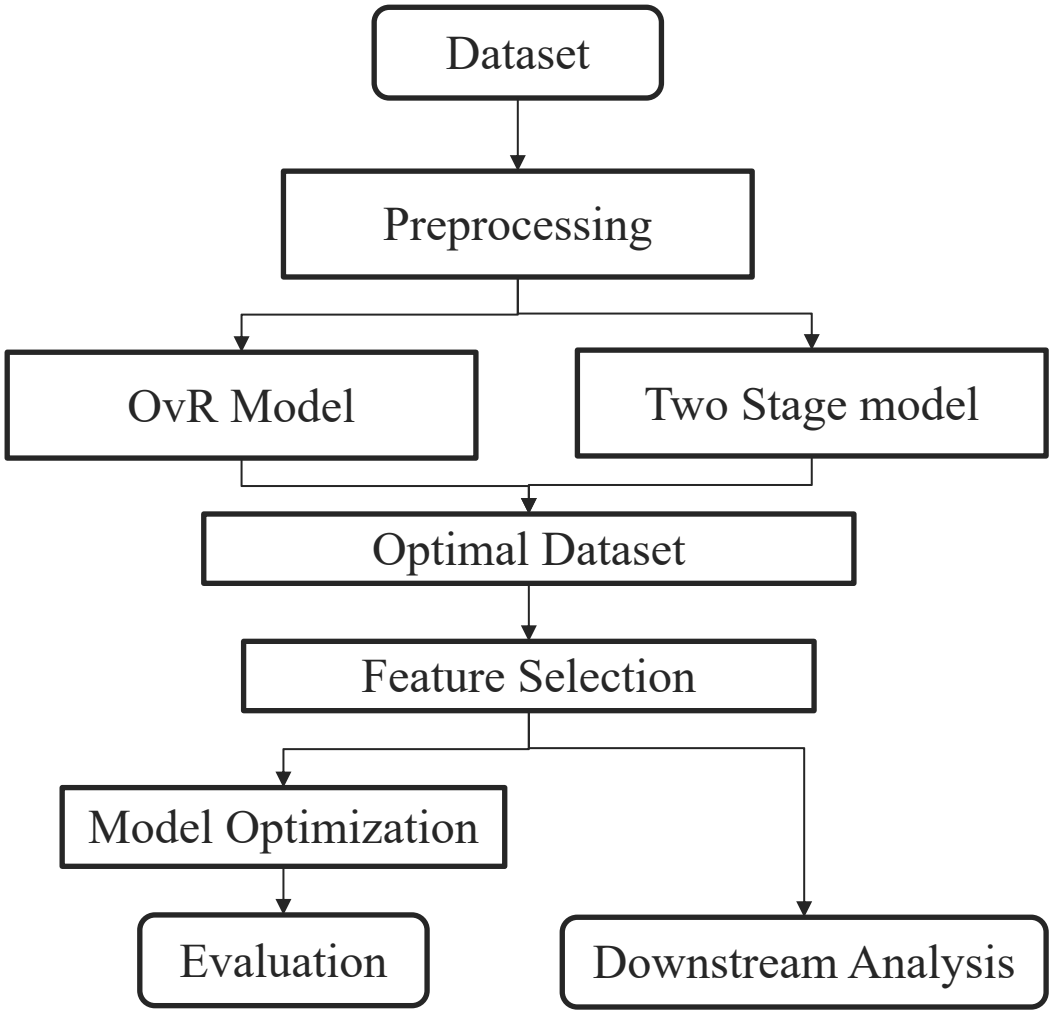
Class label	Size	Percentage
CTL	109	37%
MCI	84	27%
AD	147	36%
total	340	100%

Table 4 Batch 2 class distribution

Class label	Size	Percentage
CTL	128	37%
MCI	95	27%
AD	125	36%
total	348	100%

Methods

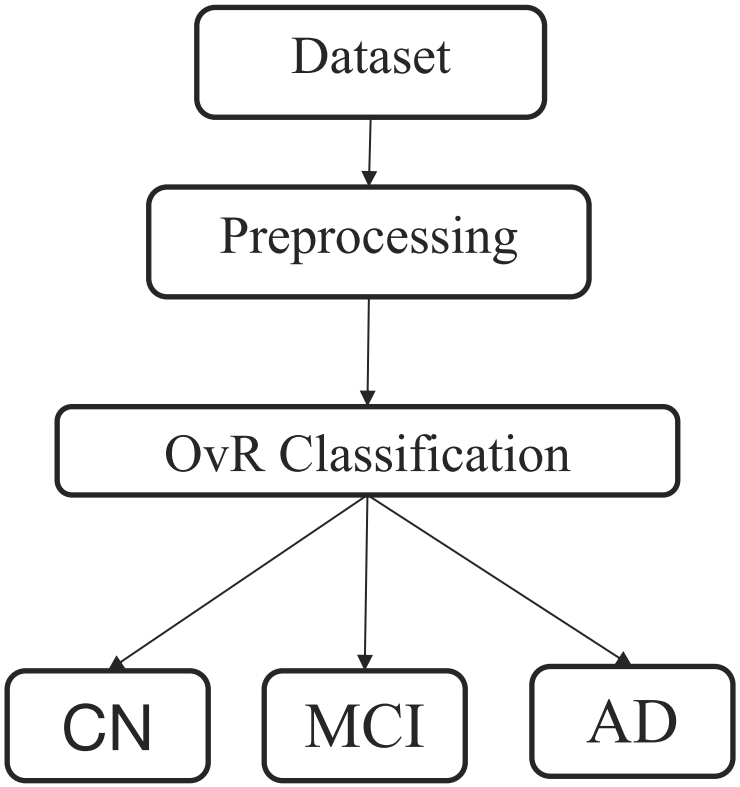
The workflow of the study

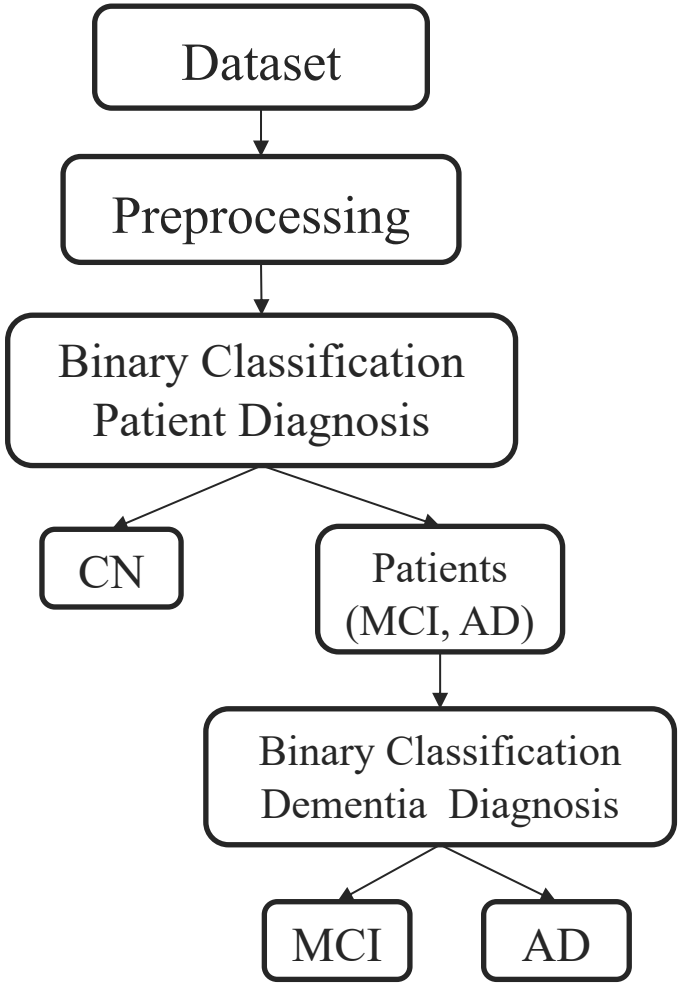


NANYANG TECHNOLOGICAL UNIVERSITY

Methods

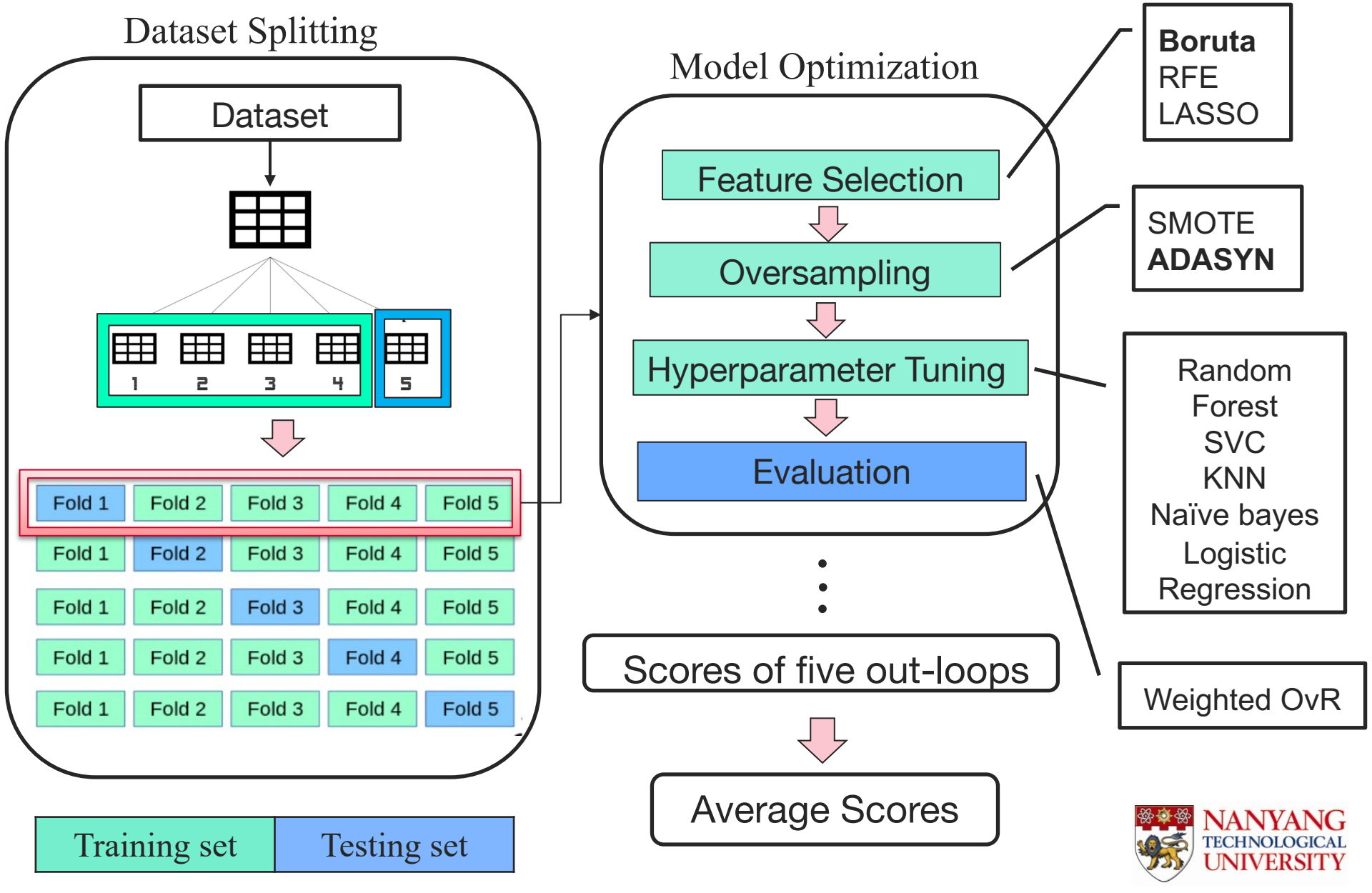
OvR multi-class modelling VS. Two stage binary modelling





Methods

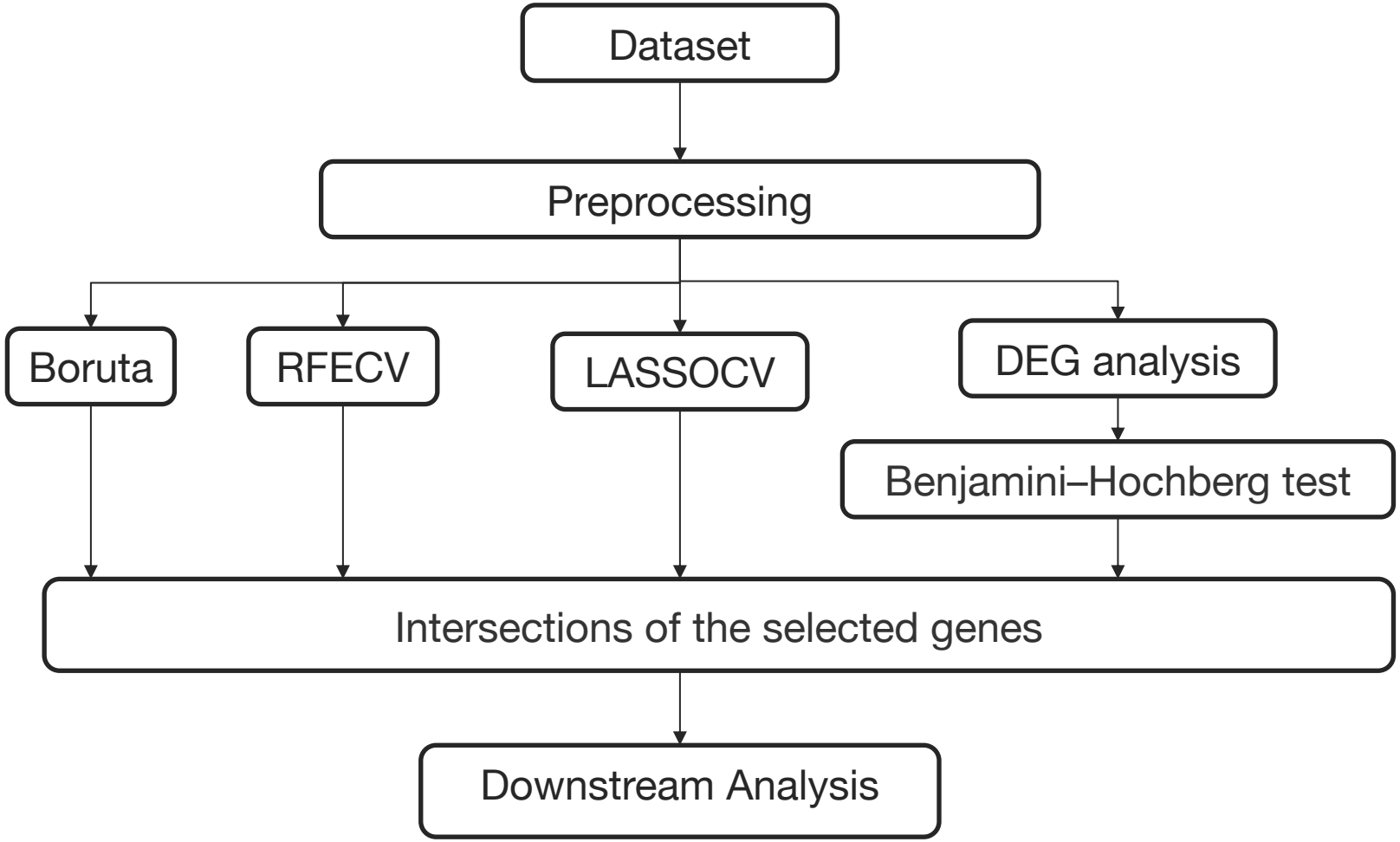
Two STAGE Model Optimization and Nested Cross Validation



NANYANG TECHNOLOGICAL UNIVERSITY

Methods

Downstream Analysis



Results

Dataset evaluation

- ANMERGE Batch 1 possesses more predictive information than the other two

Table 5 ADNI

	Accuracy	Precision	Sensitivity	Specificity	F1-score	ROC_AUC
Two Stage	0.5831	0.3055	0.3945	0.6356	**0.4868**	**0.5415**
OvR	**0.5864**	**0.3967**	**0.3973**	**0.6481**	0.3477	0.5299

Table 6 ANMerge Batch 1

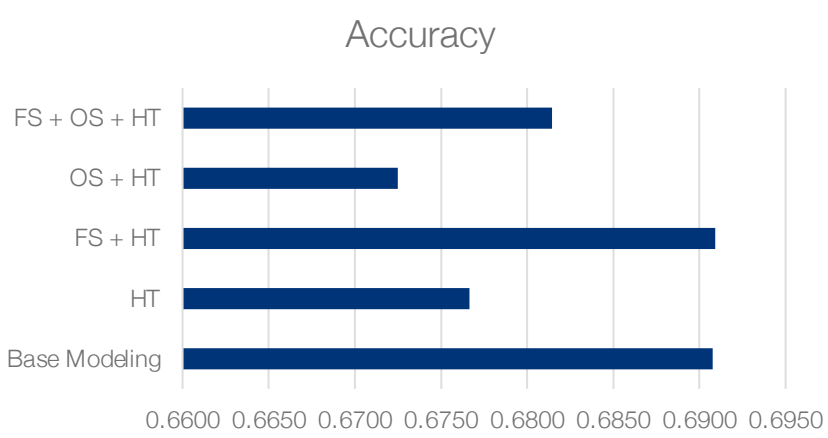
	Accuracy	Precision	Sensitivity	Specificity	F1-score	ROC_AUC
Two Stage	0.6909	0.5721	0.5503	**0.7444**	**0.5488**	0.7133
OvR	**0.7006**	**0.5973**	**0.5710**	0.7254	0.5475	**0.7190**

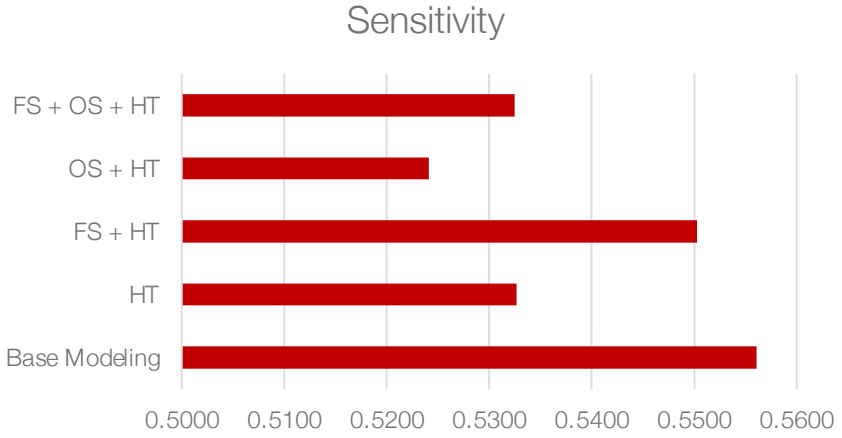
Table 7 ANMerge Batch 2

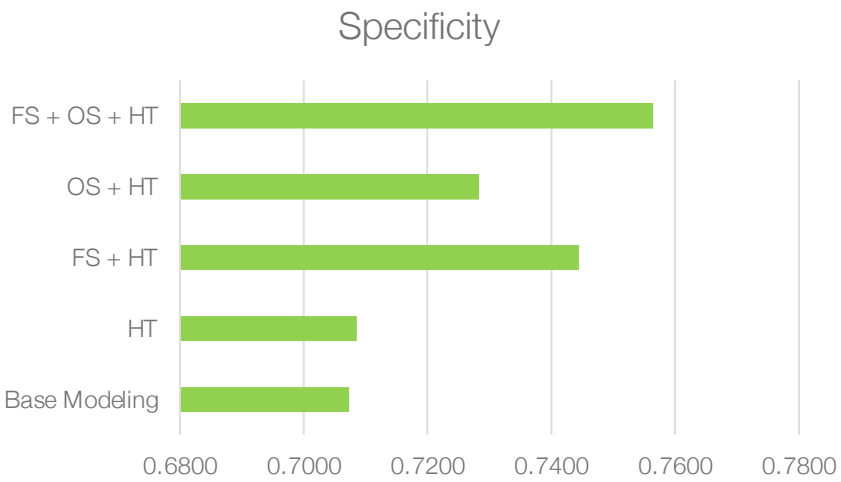
	Accuracy	Precision	Sensitivity	Specificity	F1-score	ROC_AUC
Two Stage	0.5958	0.4009	0.3938	0.6931	0.3456	0.6215
OvR	**0.6430**	**0.4571**	**0.4717**	**0.7087**	**0.4320**	**0.6278**

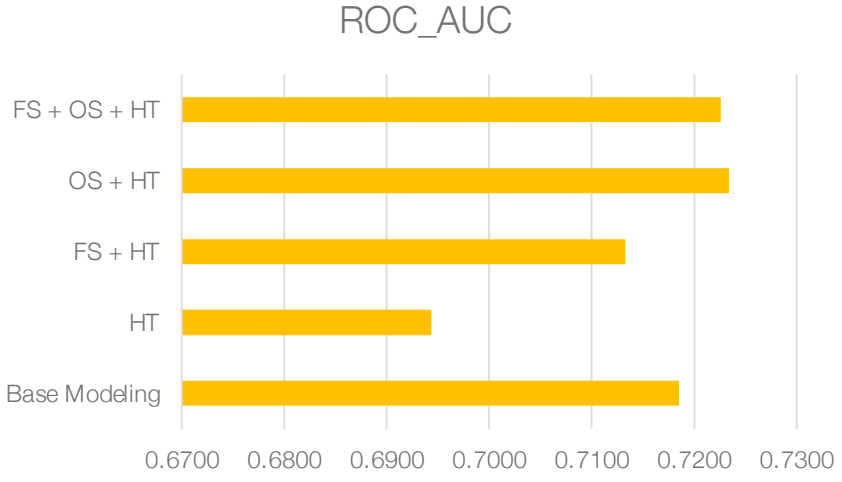
Results

2 Stage modelling optimizations









- FS: Feature selection (Boruta)
- OS: Oversampling (ADASYN)
- HT: Hyperparameter tuning
- Base model: Training model with default setting (Random Forest)

NANYANG TECHNOLOGICAL UNIVERSITY

Results

Two stage Model Evaluation

- Random Forest performs well in patients diagnosis in stage 1

Table 8 Phase 1 model

	Accuracy	Precision	Recall	Specificity	F1-score	ROC_AUC
RF	**0.794118**	0.787832	**0.956614**	0.450216	**0.771023**	**0.784198**
SVC	0.711765	0.801594	0.766698	0.596104	0.714366	0.717369
KNN	0.773529	0.808004	0.874561	0.559307	0.766911	0.750003
NB	0.711765	0.794459	0.775301	0.578355	0.713773	0.705435
LGR	0.735294	**0.815326**	0.788437	**0.623377**	0.737712	0.750262

- Random Forest performs well in AD diagnosis (AD vs (CN and MCI)) in stage 2

Table 9 Overall model performance（Random Forest in both stage）

	Accuracy	Precision	Recall	Specificity	F1-score	ROC_AUC
CN	0.794118	0.843333	0.450216	0.956614	0.575933	/
MCI	0.720588	0.372601	0.214706	0.886652	0.259706	/
AD	0.597059	0.521138	**0.829655**	0.419973	0.639837	/
Weighted	0.690753	0.587733	0.556083	0.707311	0.525435	0.718512

Results

OvR Model Evaluation

Table 10 CN VS Rest

	Accuracy	Precision	Recall	Specificity	F1-score	ROC_AUC
RF	0.785294	0.717791	0.587013	0.879001	0.637696	0.718969
SVC	0.720588	0.55993	**0.623377**	0.766512	0.588613	0.666541
KNN	0.732353	0.593182	0.587013	0.800925	0.584224	0.686961
NB	0.720588	0.573665	0.596537	0.779556	0.581494	0.679079
LGR	0.741176	0.59908	**0.62381**	0.796855	0.60868	0.684385

Table 11 MCI VS Rest

	Accuracy	Precision	Recall	Specificity	F1-score	ROC_AUC
RF	0.741176	0.538022	0.216176	0.914253	0.283308	0.718969
SVC	0.664706	0.280736	0.2375	0.804525	0.254338	0.666541
KNN	0.694118	0.343585	**0.2625**	0.835973	0.290445	0.686961
NB	0.697059	0.332784	0.215441	0.855581	0.256614	0.679079
LGR	0.652941	0.256803	0.225735	0.792986	0.237952	0.684385

Table 12 AD VS Rest

	Accuracy	Precision	Recall	Specificity	F1-score	ROC_AUC
RF	0.614706	0.541765	**0.761839**	0.503644	0.631613	0.718969
SVC	0.626471	0.566658	0.571724	0.668421	0.567594	0.666541
KNN	0.620588	0.557574	0.625747	0.617139	0.588524	0.686961
NB	0.6	0.531708	0.612874	0.590823	0.567102	0.679079
LGR	0.629412	0.567365	0.591724	0.6583	0.57817	0.684385

Results

OvR Model Optimization

Table 13 OvR Model with Default Setting

Class	Method	Model	Accuracy	Precision	Recall	Specificity	F1-score	ROC_AUC
CN	/	**LGR**	0.7412	0.5991	0.6238	0.7969	0.6087	0.6844
MCI	/	**KNN**	0.6941	0.3436	0.2625	0.8360	0.2904	0.6870
AD	/	**RF**	0.6147	0.5418	**0.7618**	0.5036	0.6316	0.7190
Weighted	/	/	0.6749	0.5112	0.5942	0.6797	0.5400	0.7000

Table 14 OvR Model after Optimization

Class	Method	Model	Accuracy	Precision	Recall	Specificity	F1-score	ROC_AUC
CN	**OS + FS**	**NB**	0.7735	0.6568	**0.6784**	0.8185	0.6629	0.7346
MCI	**HT**	**NB**	0.6147	0.3005	**0.5257**	0.6449	0.3772	0.6882
AD	/	**RF**	0.6147	0.541765	**0.7618**	0.5036	0.6316	0.7190
Weighted	/	/	0.6656	0.5190	**0.6767**	0.6395	0.5788	**0.7164**

- FS: Feature selection (Boruta)
- OS: Oversampling (ADASYN)
- HT: Hyperparameter tuning
- Base model: Training model with default setting (Random Forest)

Results

OvR Modelling VS Optimized Two Stage Modelling（AD diagnosis）

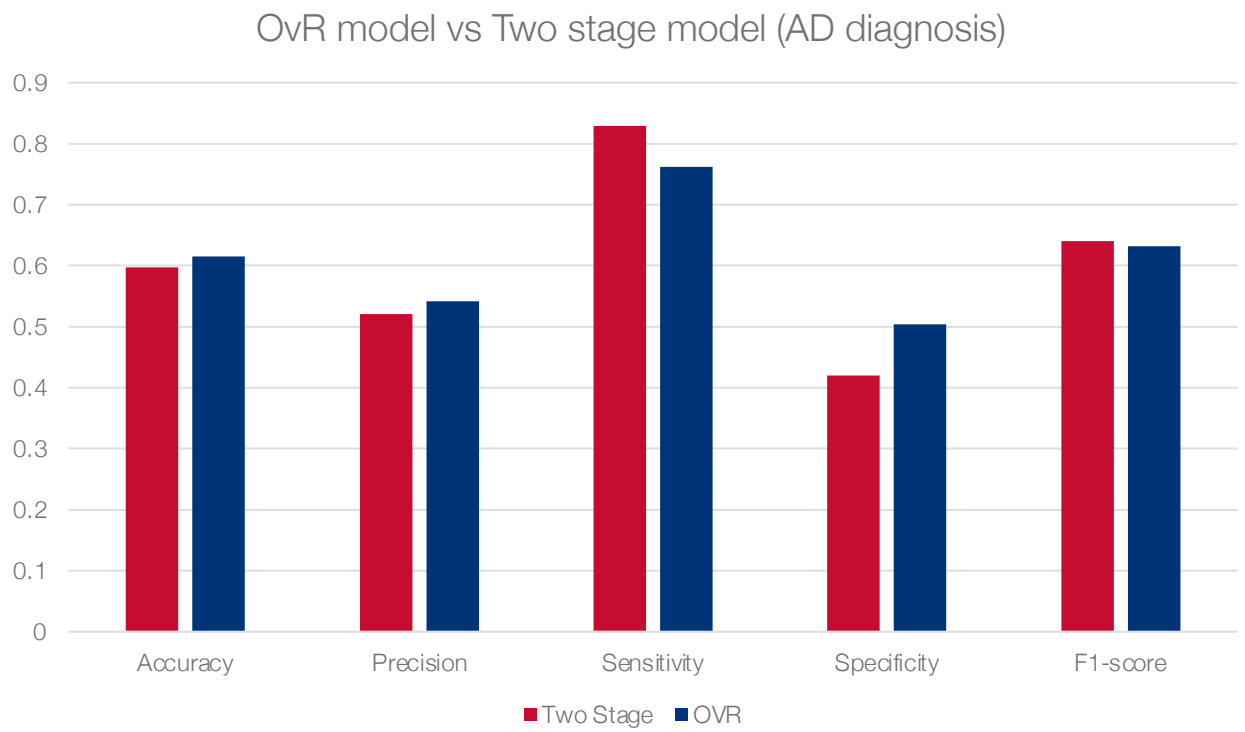


Table 15 OvR Modelling VS Optimized Two Stage Modelling（AD diagnosis）

	Accuracy	Precision	Recall	Specificity	F1-score	ROC_AUC
Two Stage	0.597059	0.521138	**0.829655**	0.419973	**0.639837**	0.7185
OvR	**0.6147**	**0.541765**	0.7618	**0.5036**	0.6316	**0.7190**

NANYANG TECHNOLOGICAL UNIVERSITY

Results

OvR Modelling VS Optimized Two Stage Modelling (overall)

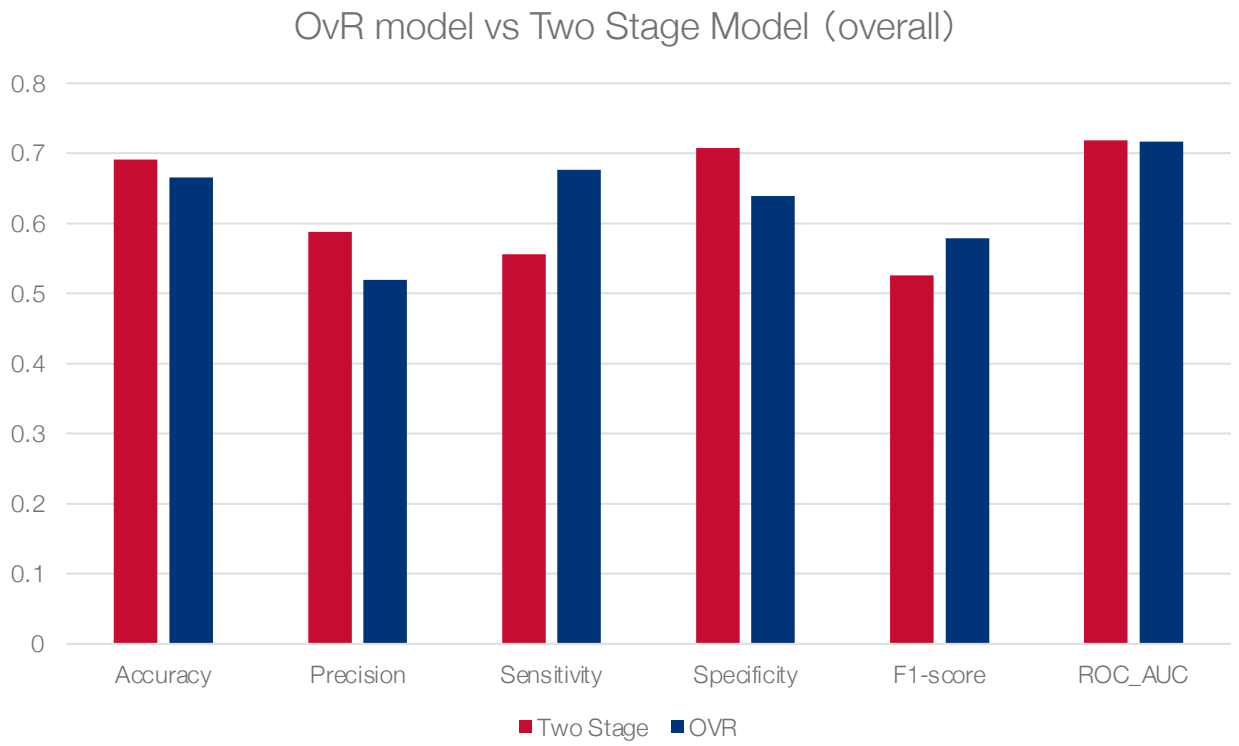


Table 16 OvR VS Optimized Two Stage Modelling（weighted score over three classes）

	Accuracy	Precision	Recll	Specificity	F1-score	ROC_AUC
Two Stage	**0.6907**	**0.5877**	0.5561	**0.7073**	0.5254	**0.7185**
OvR	0.6656	0.5190	**0.6767**	0.6395	**0.5788**	0.7164

NANYANG TECHNOLOGICAL UNIVERSITY

Results

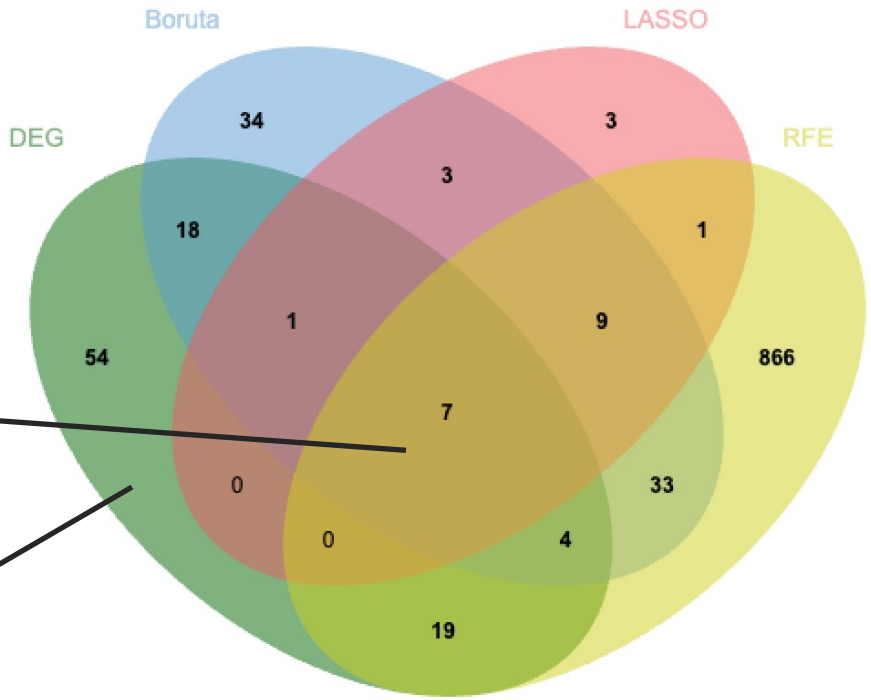
Blood-Based biomarker for Alzheimer's patients Diagnosis

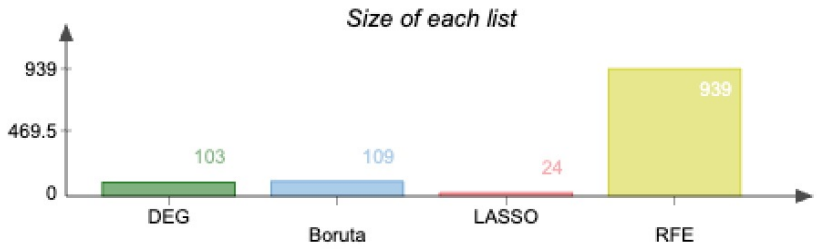
Table 17 Filtering thresholds

	UP	Down
adj.P.Val	< 0.05	< 0.05

DOWN	NOT	UP
1	4346	102

NDUFA1
MRPL51
RPL36AL
RPA3
ING3
LOC653658
CMTM2





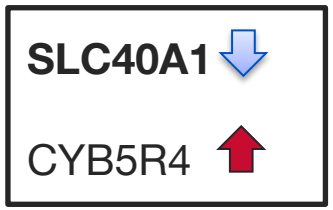
NANYANG TECHNOLOGICAL UNIVERSITY

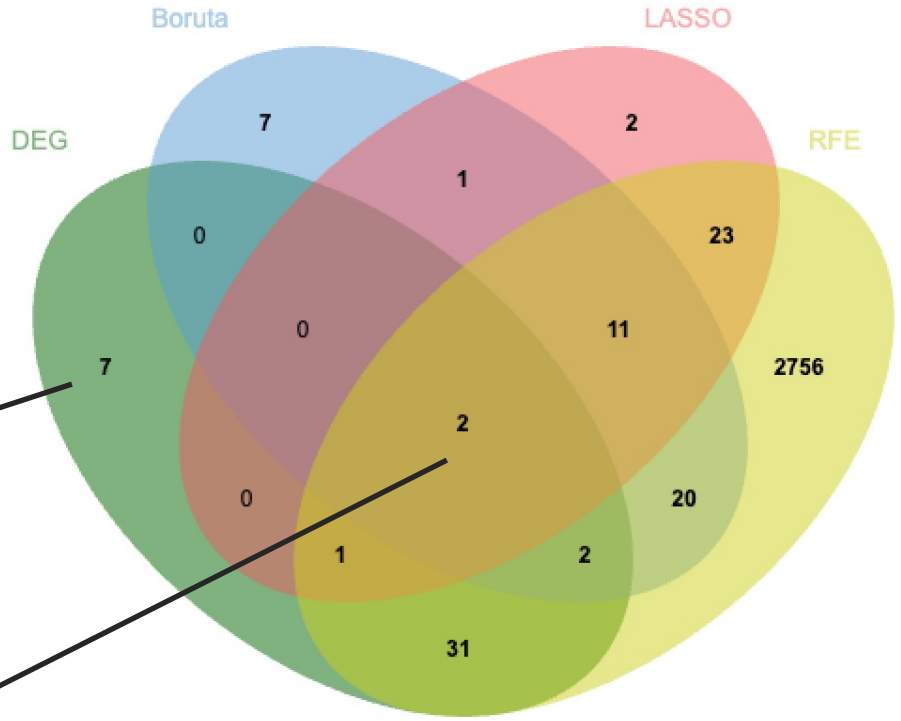
Results

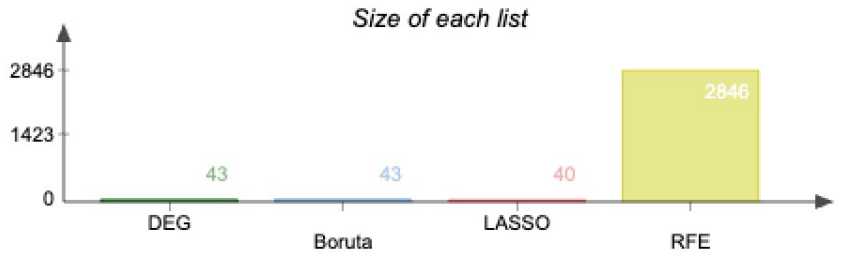
Blood-Based biomarker for Alzheimer's Disease early Diagnosis

Table 17 Filtering thresholds

	UP	Down
adj.P.Val	< 0.05	< 0.05







NANYANG TECHNOLOGICAL UNIVERSITY

Discussion

The prognostic potential of Blood-based Genetic Data

- Blood-based gene expression has the prognostic potential for cognitive patients diagnosis, the prognostic potential for this data may associate with the sequencing methods
- The microarray genetic data in ADNI didn't contain enough predictive information, the prognostic potential of this data still needs more assessment
- The blood-based genetic data in ANMerge contain predictive information for AD diagnosis, but may not contain enough information for MCI diagnosis

Discussion

Classification

- The two-stage model got good potential for AD diagnosis, the random forest classifier performs well in both cognitive impairment patients(MCI and AD) diagnosis (recall: 95.66%) in stage 1 and AD diagnosis (recall: 83.12%) over the two-stage classification

- The diagnosis of the two-stage model is convincing because the AD patients selected have gone through two classification

- The overall classification performance of the OvR model for the three classes is better than that of the two-stage model, the OvR is more flexible in both model selection and model optimization

- Neither the two-stage model nor the OvR model performs well in MCI diagnosis. Early diagnosis is the main challenge to be tackled for both the two-stage model and the OvR model

Discussion

The significant Biomarker for Alzheimer's Disease Diagnosis

Confirmed Biomarkers:

- Mutations in NDUFA1gene may lead to neurodegenerative diseases like dementia, MRPL51、RPL36AL are associated with Ribosome dysfunction is an early event in Alzheimer's disease.

- With aging, iron will accumulate in the brain, catalyzing oxidative radicals that damage brain neurons and induce Alzheimer's disease. SLC40A1 gene is associate with the function of iron excretion. Its downregulated expression can lead to the progression of Alzheimer's disease.

Potential biomarkers:

- RPA3 is a protein-coding gene mainly involved in DNA repair and DNA replication. It has been shown that disruption of DNA repair may lead to increased DNA damage in AD patients and increase the risk of AD.

- CMTM2, ING3 and LOC653658 are potential biomarkers for Alzheimer's disease prediction. CYB5R4 gene is potential biomarker for Alzheimer's development and progression.

Conclusion

1. Blood-Based Gene expression data possess the ability for Alzheimer's disease diagnosis

2. Both the two-stage model and the OvR model possess the ability of AD identification, while the diagnosis of the two-stage model is strict and convincing

3. Compared to the two-stage model, the OvR model is more flexible than the two-stage model in model selection and optimization

4. The combination of ML-based (RFE, LASSO, Boruta) and statistic-based methods (DEG analysis) can make the result of feature selection more robust

5. The biomarkers detected show that Alzheimer's disease is associated with the dysfunction of ribosomes and mitochondria in multiple cortical areas, the progression of Alzheimer's disease is associated with the iron excretion of the brain.

Thank you!

NANYANG
TECHNOLOGICAL
UNIVERSITY

Q & A