

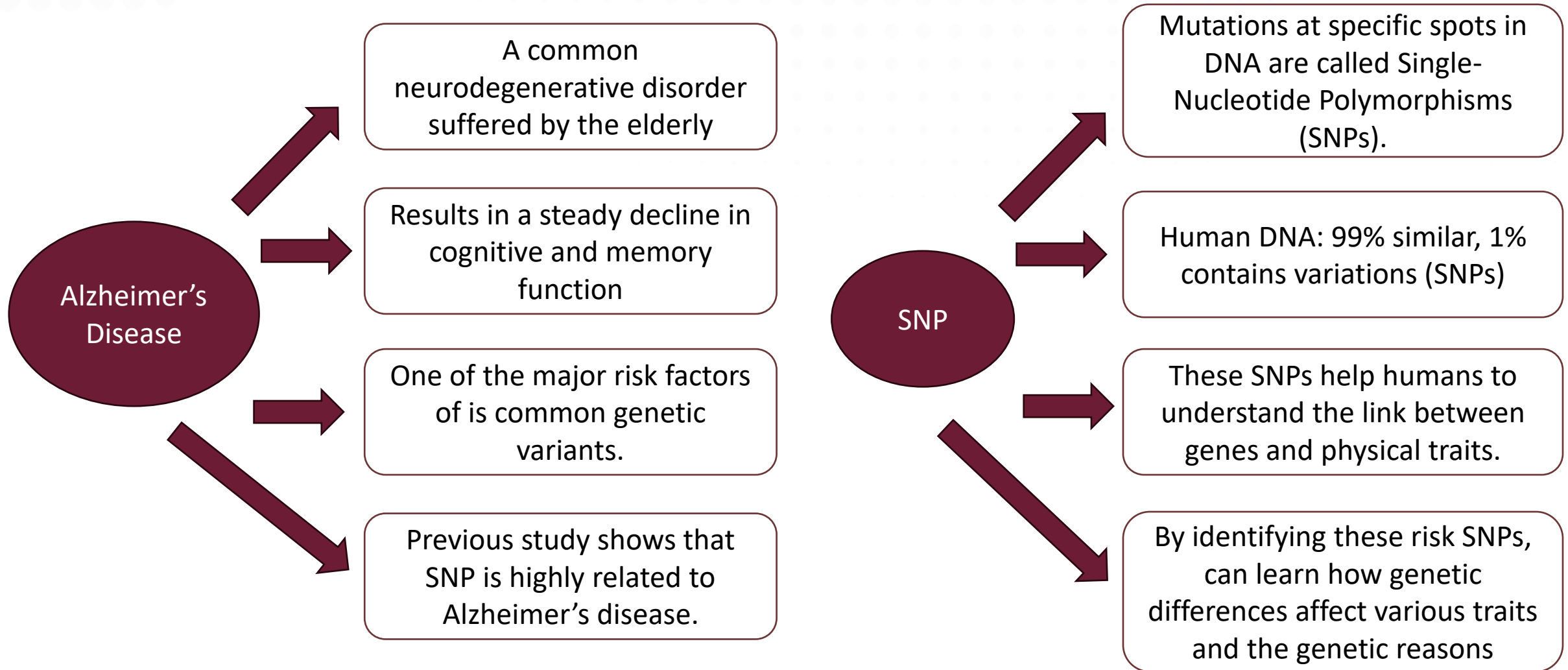
USE OF DEEP-LEARNING APPROACH TO CLASSIFY ALZHEIMER'S DISEASE OR MILD COGNITIVE IMPAIRMENT

LU QI YAN (A21EC0049)

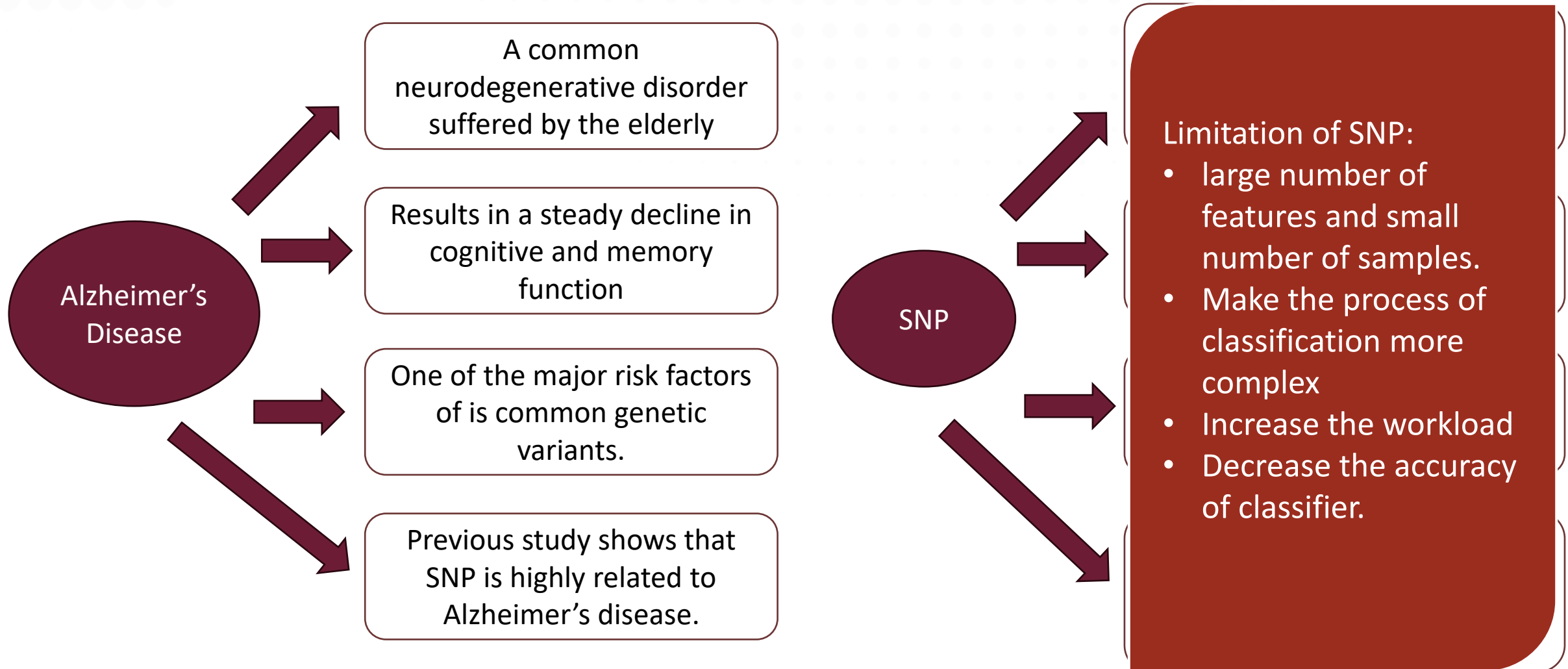
Video link: <https://youtu.be/V82iTmhq0Oo>

Demo Video: <https://youtu.be/VfdP4rfbCko>

Problem Background



Problem Background



Problem Statement

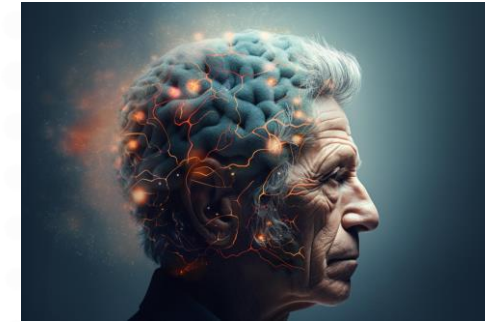
The high dimensionality of SNP dataset is one of the limitations for the deep learning model to accurately classify Alzheimer's disease also increase the risk of overfitted, affecting the model's ability to generalize.

Goal

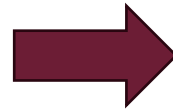
To implement feature reduction on a CNN-based classification model that classifies Alzheimer's Disease or Mild Cognitive Impairment using the SNPs dataset.

Research Questions	Research Objectives
a) What is the impact of feature reduction on SNP data for improving classification performance?	a) To investigate the impact of feature reduction on SNP data for improving classification performance.
b) How to implement a CNN-based classification model with feature reduction, including both feature selection (filter methods) and feature extraction (Autoencoder techniques) to classify Alzheimer's disease and Mild Cognitive Impairment?	b) To implement a CNN-based classification model with feature reduction, including both feature selection (filter methods) and feature extraction (Autoencoder techniques) to classify Alzheimer's disease and Mild Cognitive Impairment.
c) How to evaluate the performance of the CNN-based classification model?	c) To evaluate the performance of the CNN-based classification model in terms of accuracy, sensitivity, specificity, and Area Under Curve (AUC).

Research Importance



Alzheimer's disease is a chronic, irreversible brain illness that is currently no proven treatment



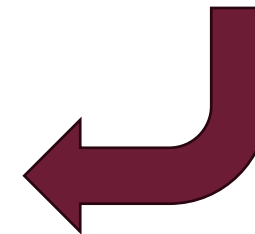
Advancement can be stop / slow down through medications



Stopping and managing the progression depends greatly on early detection



Important to have an accurate, sensitive and specific deep learning model that can help to classify Alzheimer's disease or mild cognitive impairment.



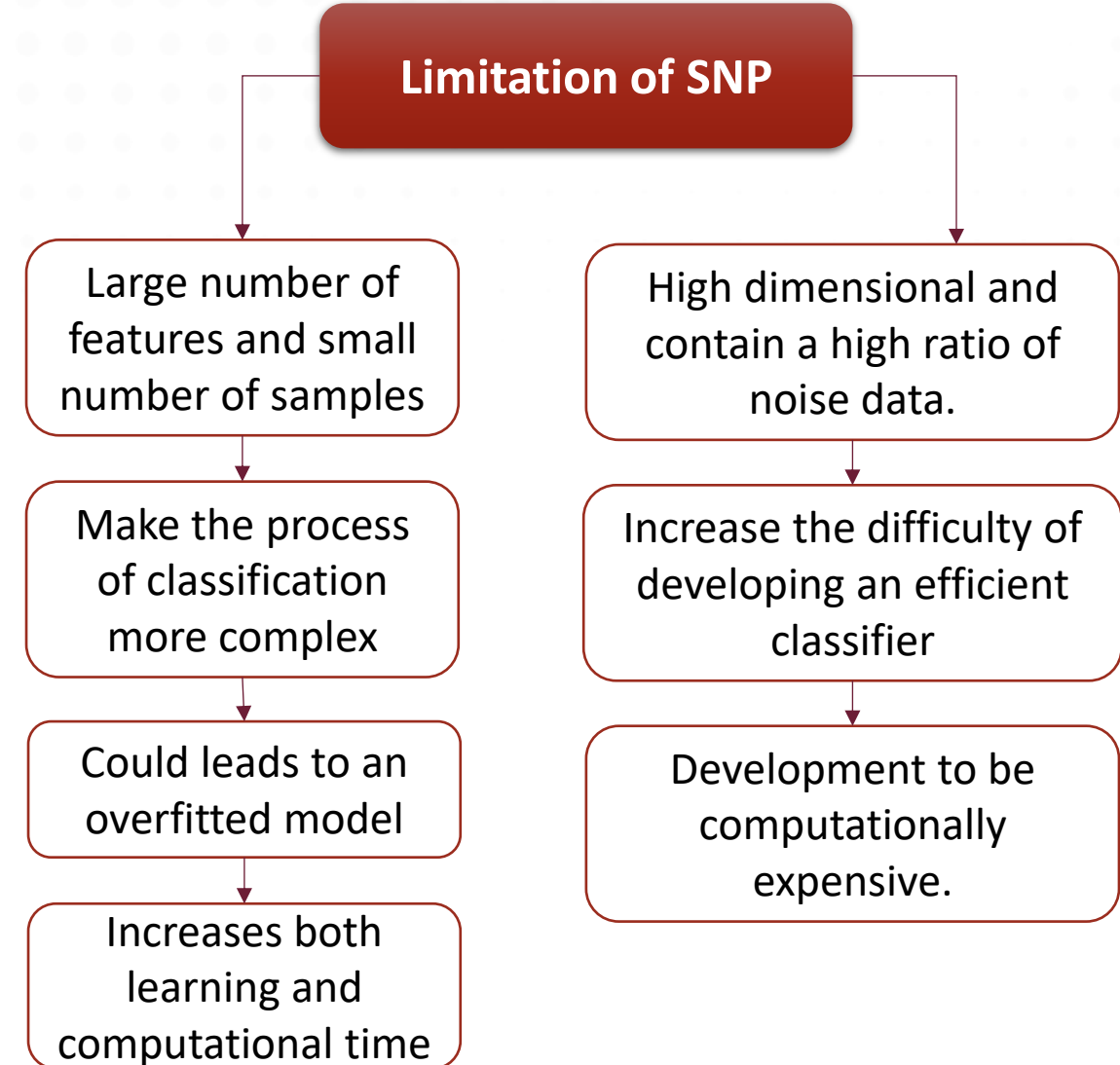
Chapter 2

Literature Review

USE OF DEEP-LEARNING APPROACH TO CLASSIFY ALZHEIMER'S
DISEASE OR MILD COGNITIVE IMPAIRMENT

SNP DATASET

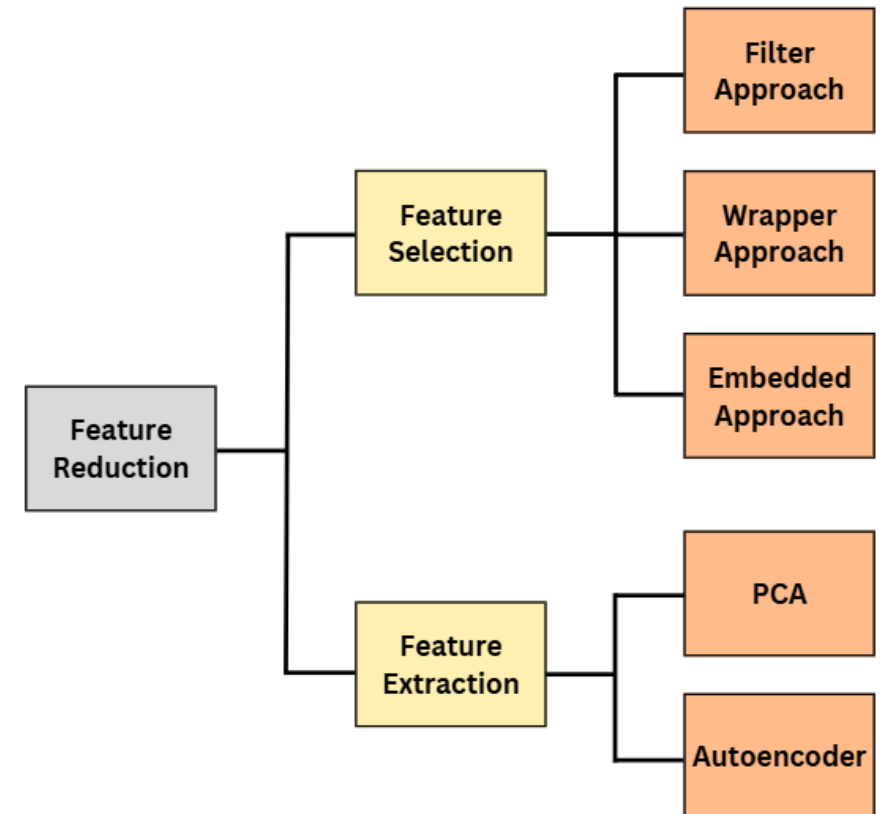
Importance of SNP	References
Help to identify the target heritability genes of disease and other polygenic traits.	(Liu et al., 2018)
Help in developing precision medicine as well as to improve molecular mechanisms.	(Liu et al., 2018)
Genomic prediction mainly relies on thousands of DNA markers, especially single nucleotide polymorphisms (SNPs). It is used to forecast disease risk and highly complex traits influenced by many genes.	(Lee et al., 2023)
Significantly improve the accuracy and quality of the findings.	(Muneeb et al., 2022)



Feature Reduction

A common technique to address the issue of having too many irrelevant features.

Importance of Feature Reduction	References
Reducing Dimensionality and Noise	(Pirmoradi et al., 2020) (Arafa et al., 2023)
Improving Classification Performance <ul style="list-style-type: none"> • Increase efficiency and scalability by reducing training time and memory usage 	(Pirmoradi et al., 2020) (Li et al., 2022) (Qin et al., 2022) (Han et al., 2017)
Increase Efficiency and Scalability	(Jubair et al., 2021) (Chen et al., 2023) (Zhou et al., 2024) (Wang et al., 2025)
Improving Interpretability and Biological Relevance <ul style="list-style-type: none"> • Input can be more information and so-called more biologically meaningful 	(Muneeb et al., 2022)



Comparison of Feature Selection Methods

Method	Description	Advantages	Disadvantages	Use Cases
Filter Methods	Rank features based on statistical tests (e.g., Chi-square, correlation). Features scoring below a threshold are removed.	<ul style="list-style-type: none"> • Computationally efficient • Reduces risk of overfitting • Easy to implement 	<ul style="list-style-type: none"> • Ignores feature dependencies • May miss interactions • Less accurate compared to wrapper/embedded methods 	<ul style="list-style-type: none"> • Initial screening of features • High-dimensional datasets like SNP genotype datasets
Wrapper Methods	Selects feature subsets based on classifier performance (e.g., forward/backward selection).	<ul style="list-style-type: none"> • Considers feature interactions • Often more accurate than filter methods 	<ul style="list-style-type: none"> • Computationally intensive Prone to overfitting • Can be slow for large datasets 	<ul style="list-style-type: none"> • Small to medium-sized datasets • When computational resources are available
Embedded Methods	Integrates feature selection into the training process (e.g., LASSO, decision tree-based methods).	<ul style="list-style-type: none"> • Efficient with classifier • Can handle feature interactions • Generally balances accuracy and efficiency 	<ul style="list-style-type: none"> • Specific to the learning algorithm • May require more complex implementation 	<ul style="list-style-type: none"> • Real-time applications • When a specific classifier is preferred

Comparison of Feature Extraction Methods

Method	Pros	Cons	References
PCA	<ul style="list-style-type: none"> ● Captures major variance in data- ● Useful for population stratification correction 	<ul style="list-style-type: none"> ● May discard phenotype-relevant SNPs ● Components not interpretable ● Sensitive to outliers 	Zhang et al., 2022 Price et al., 2006
Autoencoder	<ul style="list-style-type: none"> ● Learns nonlinear pattern ● Dimensionality reduction without label requirement ● Denoises input 	<ul style="list-style-type: none"> ● Hard to interpret features ● Requires large sample size ● Sensitive to missing data or scaling 	Nguyen et al., 2020 Le et al., 2017

Summarization

“ Feature reduction is a common technique to address the issue of having too many irrelevant features. Typically, feature reduction methods **reduce the dimensionality** of training data by excluding SNPs that either have low or negligible predictive power for the phenotype class, or are redundant. This approach **enhances learning efficiency, improves predictive accuracy, and simplifies the results.**

Consequently, selecting the most relevant features can **aid researchers in understanding the biological processes underlying the disease.** ”

Highest performance achieved by benchmark paper

Li, L., Yang, Y., Zhang, Q., Wang, J., Jiang, J. and Neuroimaging Initiative, 2021. Use of deep-learning genomics to discriminate healthy individuals from those with Alzheimer's disease or mild cognitive impairment. Behavioural Neurology, 2021.

Deep Learning Algorithms	Performance
CNN	<ul style="list-style-type: none">● Accuracy = 92.45%● Sensitivity = 93.87%● Specificity = 90.00%● Area under the curve (AUC) = 0.915

Highest performance achieved by benchmark paper

Li, L., Yang, Y., Zhang, Q., Wang, J., Jiang, J. and Neuroimaging Initiative, 2021. Use of deep-learning genomics to discriminate healthy individuals from those with Alzheimer's disease or mild cognitive impairment. *NeuroImage*. 2021.

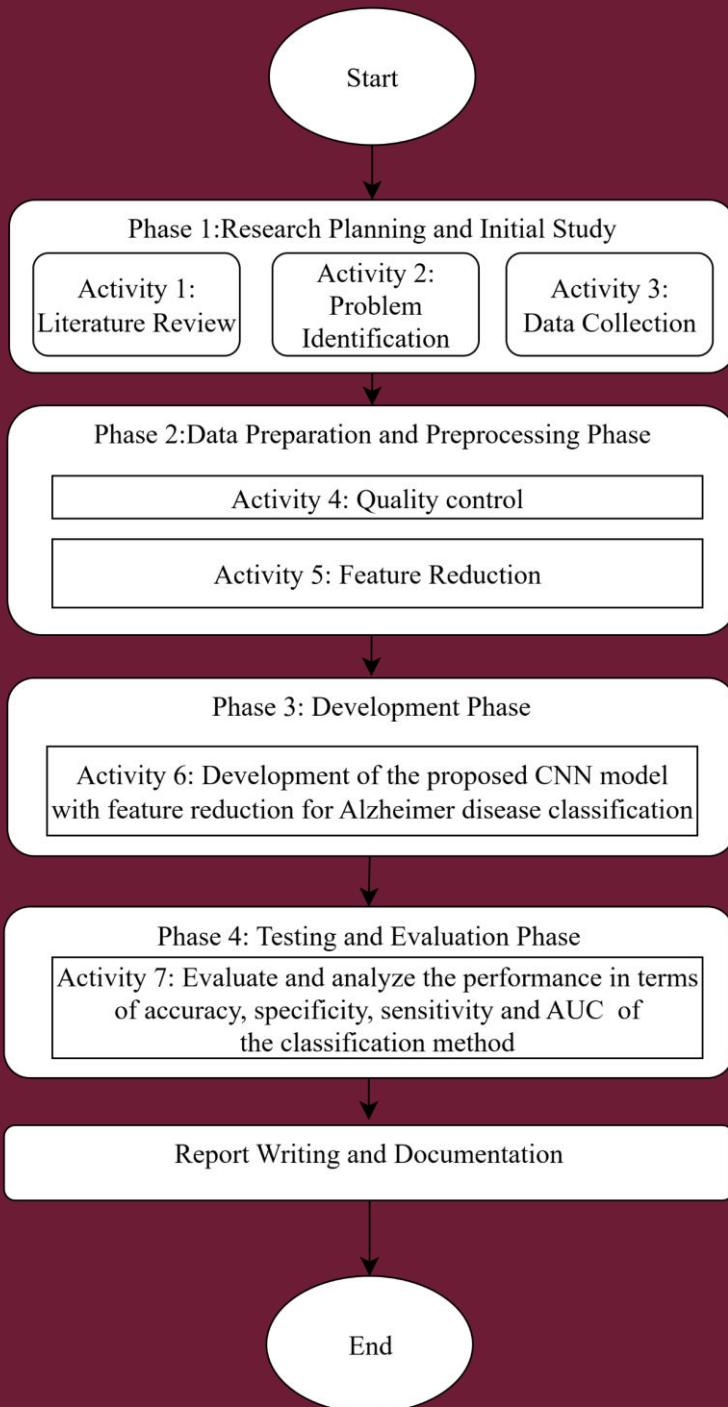
- Did not include feature reduction
- Did not interpret significance features

Deep Learning Algorithms	
CNN	<ul style="list-style-type: none">• Accuracy = 92.45%• Sensitivity = 93.87%• Specificity = 90.00%• Area under the curve (AUC) = 0.915

Chapter 3

Research Methodology

USE OF DEEP-LEARNING APPROACH TO CLASSIFY ALZHEIMER'S
DISEASE OR MILD COGNITIVE IMPAIRMENT



Research Framework

The dataset used in this study is obtained from the ADNI database (<http://adni.loni.usc.edu/>).

- Data from a total of 1461 participants was collected and included in the ADNI database. These participants came from specific groups within ADNI, called ADNI 1, ADNI 2, and ADNI GO.
- Dataset used contained 622 subjects with AD, 473 subjects with MCI, and 366 normal subjects.
- In this dataset, there are total of 620,901 genotype markers.

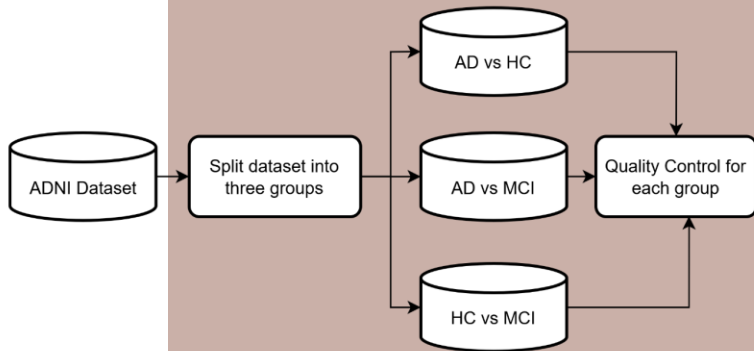
- AD = Alzheimer's Disease
- MCI = Mild Cognitive Impairment
- HC = Healthy Control

Performance Measure

In this research, the performance measurement for Alzheimer's disease is measured based on accuracy, specificity, sensitivity, and area under the curve (AUC). Significance features (SNP) that are related to Alzheimer's disease will be verify through biological verification.

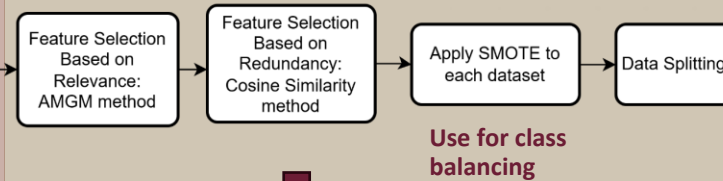
Research Workflow

Data Preprocessing: Standard procedure following existing paper

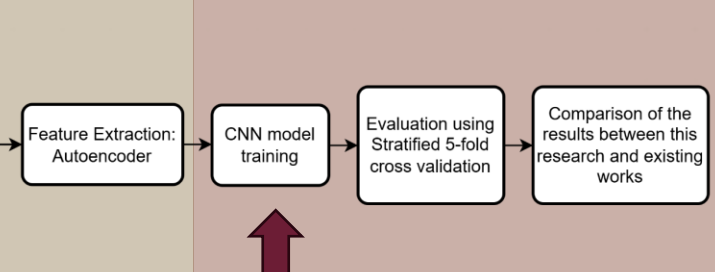
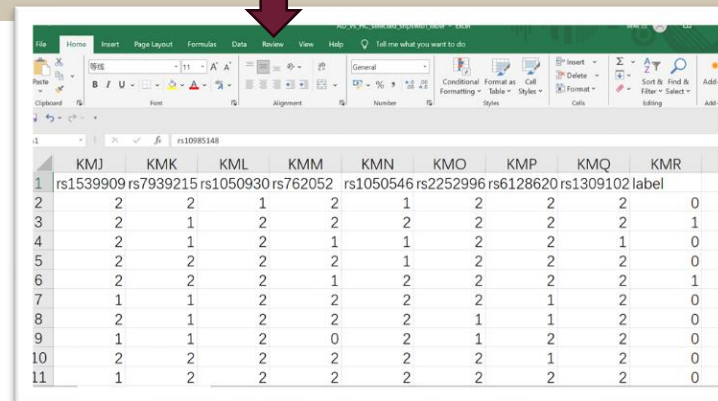


Perform quality control (QC) process to preprocess the data such as handle missing data and inconsistent records.

Feature Reduction

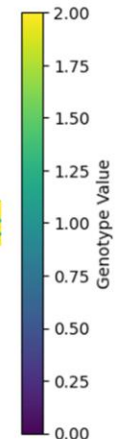


Model Training & Evaluation

	KMJ	KMK	KML	KMM	KMN	KMO	KMP	KMQ	KMR
1	rs1539909	rs7939215	rs1050930	rs762052	rs1050546	rs2252996	rs6128620	rs1309102	label
2	2	2	1	2	1	2	2	2	0
3	2	1	2	2	2	2	2	2	1
4	2	1	2	1	1	2	2	1	0
5	2	2	2	2	1	2	2	2	0
6	2	2	2	1	2	2	2	2	1
7	1	1	2	2	2	2	1	2	0
8	2	1	2	2	2	1	1	2	0
9	1	1	2	0	2	1	2	2	0
10	2	2	2	2	2	2	1	2	0
11	1	2	2	2	2	2	2	2	0

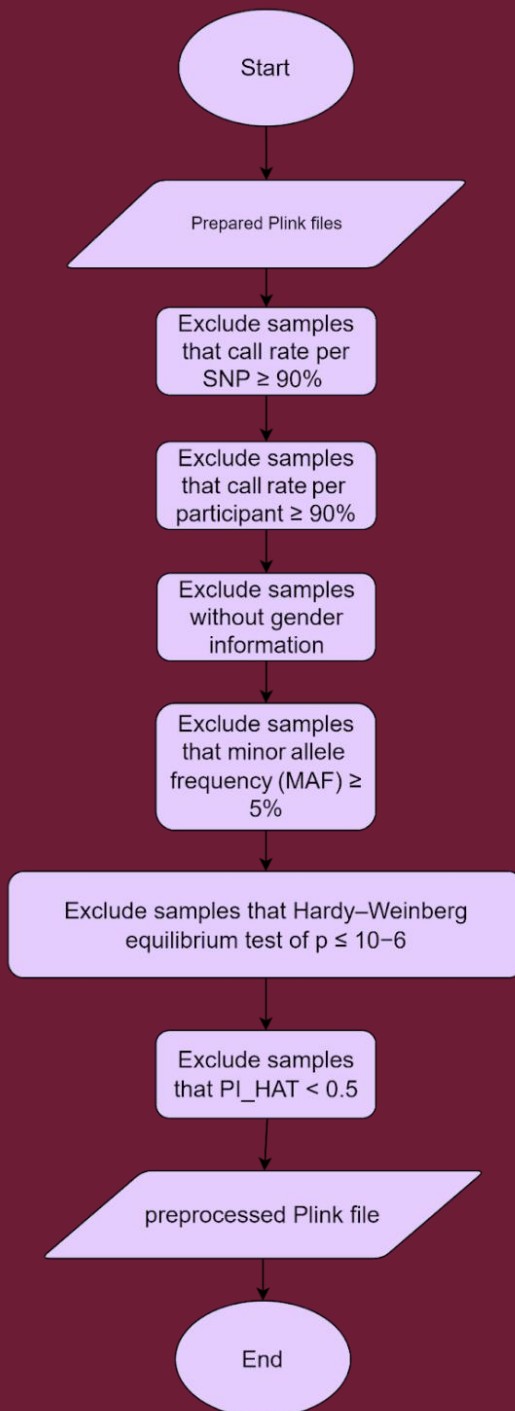
Reshaped SNP Input (21x149)



What does the model learn

CNN model learn patterns in allele burden across SNPs for AD classification

Genotype	Encoded Value	Meaning
AA	0	Homozygous major
AG	1	Heterozygous
GG	2	Homozygous minor



Quality Control

SNPs and participants were excluded from the analysis if they failed to meet any of the following criteria:

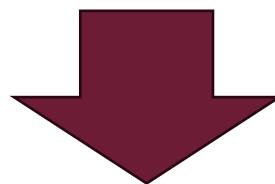
- call rate per SNP $\geq 90\%$;
- call rate per participant $\geq 90\%$
- gender check
- minor allele frequency (MAF) $\geq 5\%$;
- Hardy–Weinberg equilibrium test of $p \leq 10^{-6}$
- $PI_HAT < 0.5$.

Based on paper
(Li et al., 2021)

Quality Control

After the QC steps, the number of features for each subject in the paired groups were as follows: 302,585 in the HC and MCI groups, 302,631 in the HC and AD groups, and 302,149 in the MCI and AD groups. The data is now cleaned and processed and it is now ready for the further steps. The table below shows the number of variants and samples left after quality control.

Groups	Number of Variants	Number of Samples
AD	620,901	622
HC		473
MCI		366



Groups	Number of Variants	Number of Samples
AD vs HC	302,631	746
MCI vs HC	302,585	890
AD vs MCI	302,149	440

Chapter 4

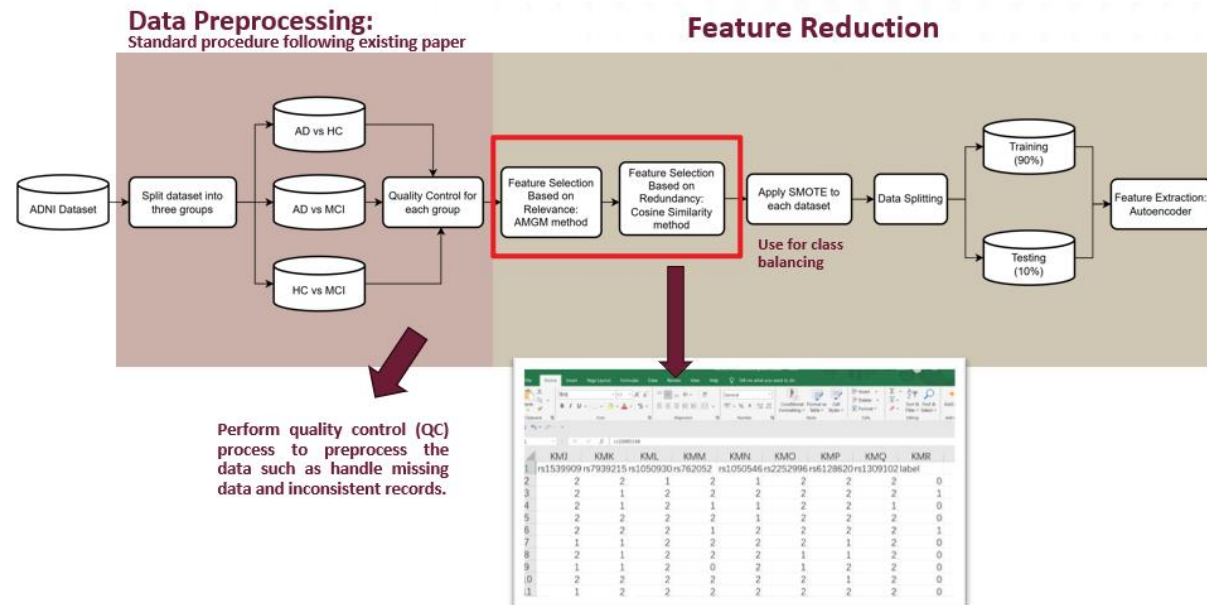
Research Design and Implementation

USE OF DEEP-LEARNING APPROACH TO CLASSIFY ALZHEIMER'S
DISEASE OR MILD COGNITIVE IMPAIRMENT

Feature Reduction

Feature Selection

In this project, the feature selection approach used is the filter method as the filter method is classifier dependent which decreases time and computational cost, especially in high dimensional data. The data is first filtered based on **relevance (related)** and top features are selected in this phase. Then, the features are filtered again based on **redundancy (not/less related)** to create a subset of top SNPs which is the top features selected.



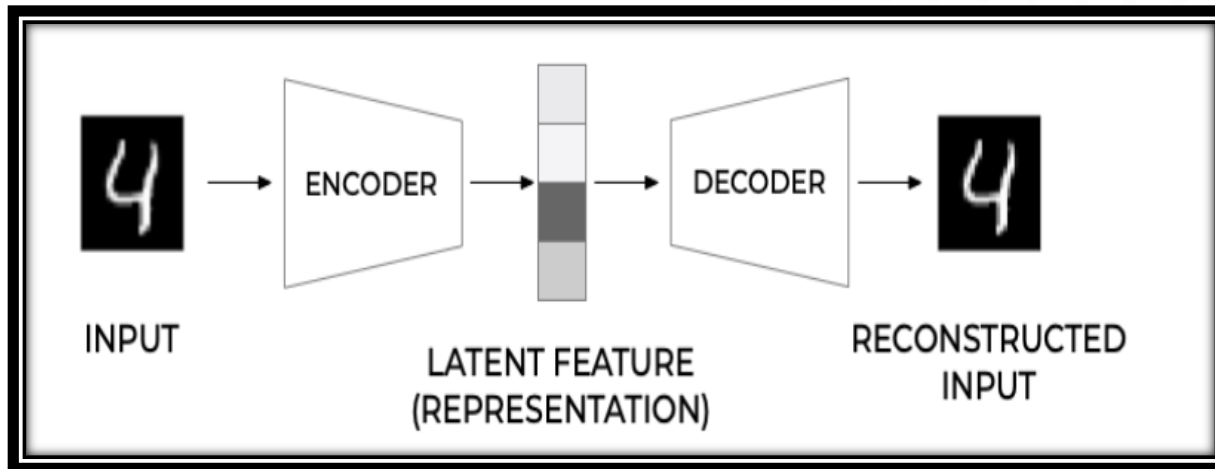
- **Relevance Calculation:** The relevance of each SNP to the target is computed using dispersion measures, particularly the Arithmetic Mean-Geometric Mean (AMGM). This measure is effective for sparse data and helps identify the top features with the highest relevance scores.

A higher AMGM score indicates greater dispersion, suggesting that the SNP is more variable and potentially more informative in differentiating between healthy and disease states.

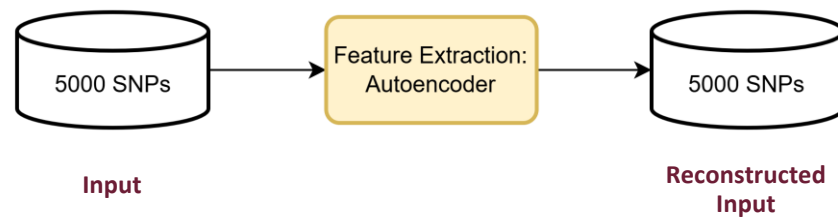
- **Redundancy Calculation:** For the top features, redundancy is assessed with cosine similarity method to remove redundant SNPs. If two SNPs exhibited a similarity above a threshold of 0.95, the latter SNP in the comparison was removed to retain only one representative from the redundant pair.

Feature Reduction

Feature Extraction

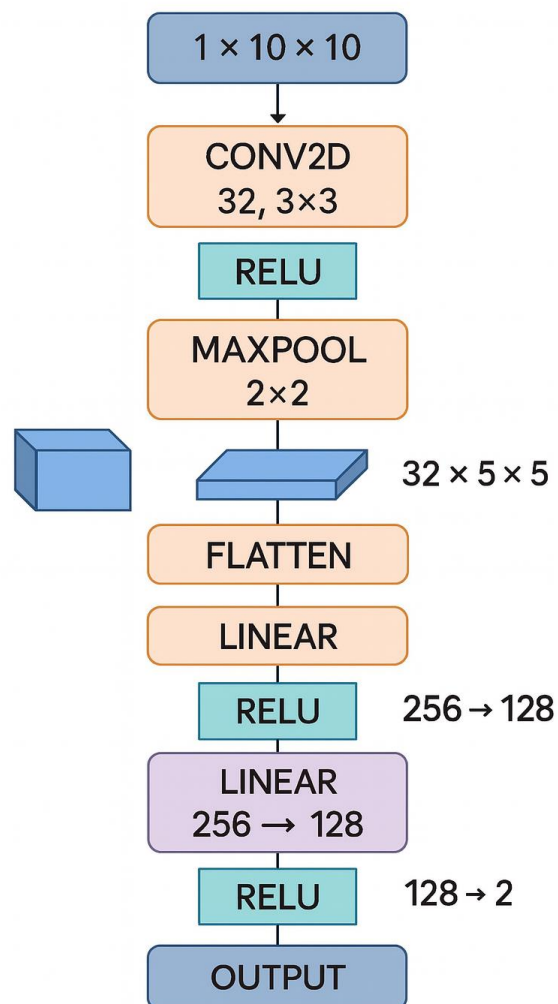


Next, the clean SNPs features is passed into a denoising autoencoder to further denoise the input features. The autoencoder is built with a fully connected encoder and decoder network. Encoder is used to compress the high dimensional input into a low dimensional latent space which will later be reconstruct through decoder.



Component	Hyperparameter	Value
Encoder	Layer 1 Neurons	1024
	Activation Function (L1)	ReLU
	Layer 2 Neurons	512
	Activation Function (L2)	ReLU
Decoder	Layer 1 Neurons	1024
	Activation Function (L1)	ReLU
	Layer 2 Neurons	Original input dimension
	Activation Function (L2)	ReLU
Training	Optimizer	Adam
	Learning Rate	0.001
	Loss Function	Mean Squared Error (MSE)
	Batch Size	32
	Epochs	20
	Framework	PyTorch

CNN Architecture



Training Configuration	
Optimizer	Adam
Learning Rate	$1e-2$
Batch Size	8
Epochs	30
Validation	5-fold cross-validation

Based on paper
(Li et al., 2021)

Chapter 5

Results, Analysis And Discussion

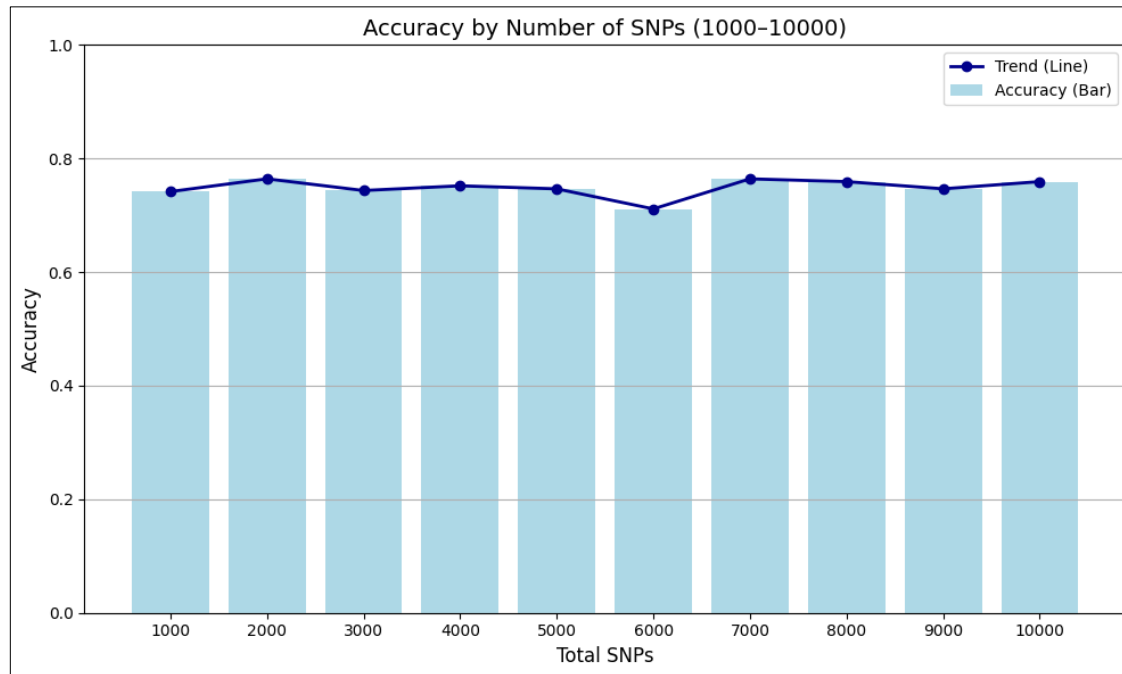
USE OF DEEP-LEARNING APPROACH TO CLASSIFY ALZHEIMER'S
DISEASE OR MILD COGNITIVE IMPAIRMENT

AD vs HC group

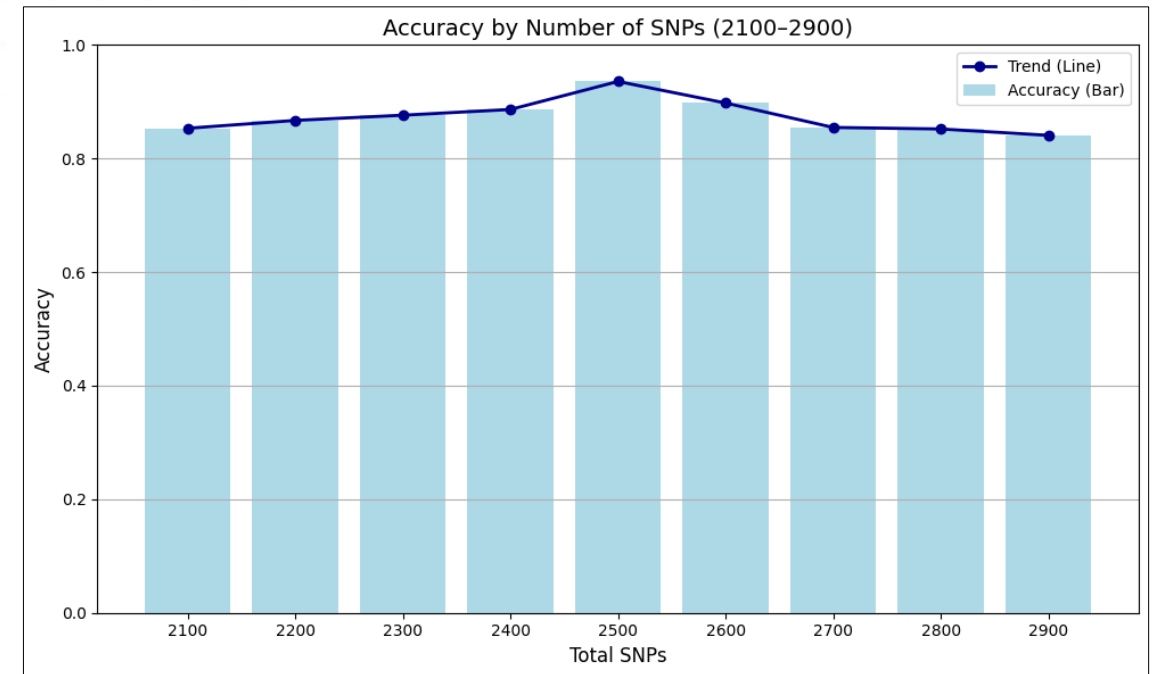
Method	Selected SNPs	Accuracy	Specificity	Sensitivity	AUC
CNN (Publish result)	302,631 (Use all)	92.45%	93.87%	90.00%	0.9150
FS + CNN	2000	76.42%	76.32%	76.52%	0.8412
FS + FE + CNN	2500	93.56%	92.14%	94.99%	0.9831

- Method with both feature selection and feature extraction has the best performance.
- The results are better than benchmark paper.
- Method with only feature selection also achieved quite good results with only 2000 SNPs selected.

AD vs HC group



FS + CNN



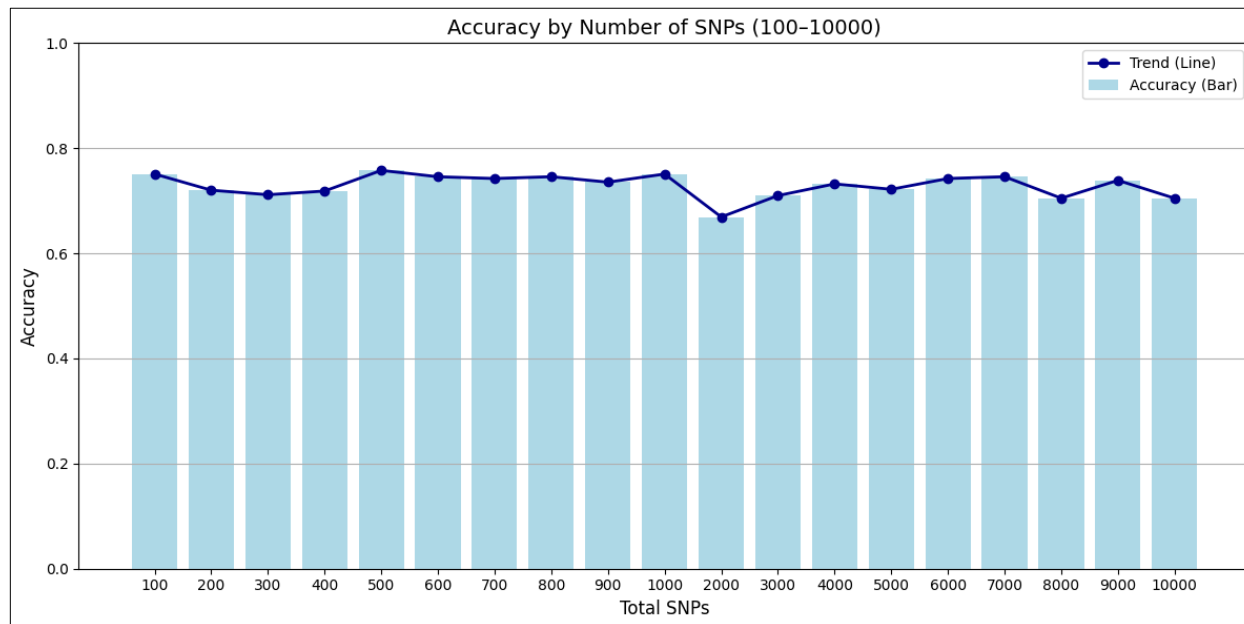
FS + FE + CNN

AD vs MCI group

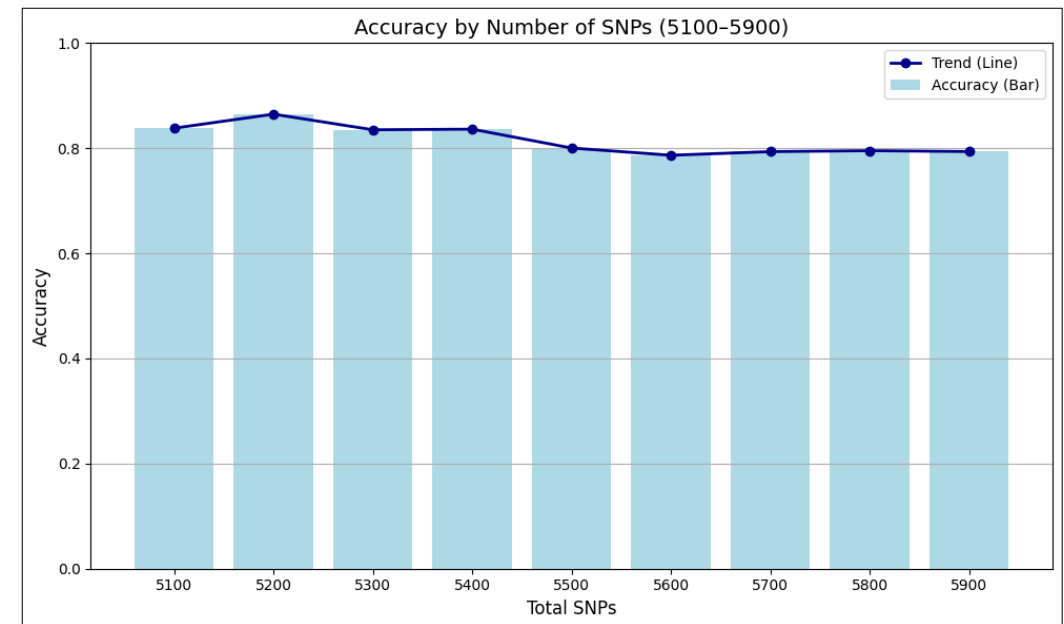
Method	Selected SNPs	Accuracy	Specificity	Sensitivity	AUC
CNN (Publish result)	302,585 (Use all)	86.42%	97.42%	71.91%	0.8400
FS + CNN	1000	75.09%	69.32%	80.90%	0.8154
FS + FE + CNN	5200	86.48%	85.34%	87.59%	0.9385

- Method with both feature selection and feature extraction has the best performance.
- The results are better than benchmark paper.
- Method with only feature selection also achieved quite good results with only 1000 SNPs selected.

AD vs MCI group



FS + CNN



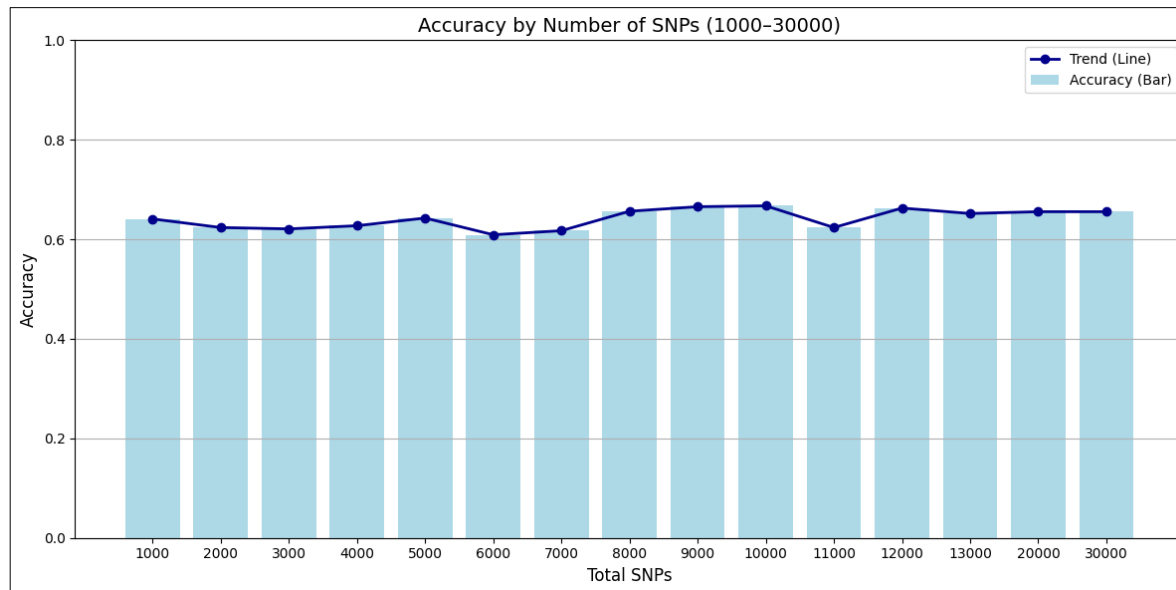
FS + FE + CNN

HC vs MCI group

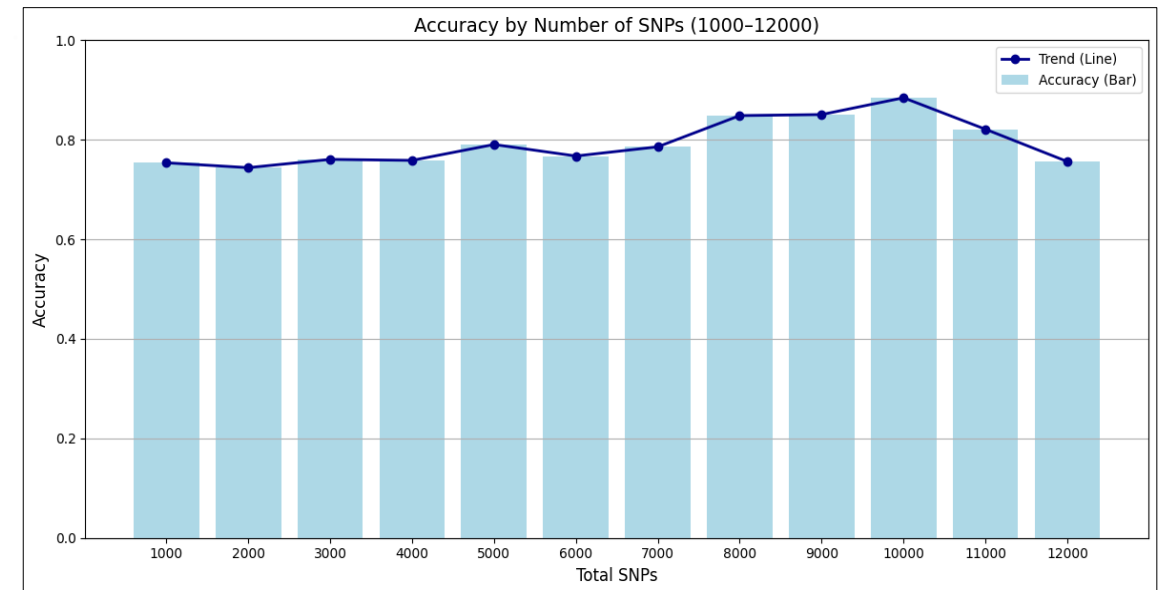
Method	Selected SNPs	Accuracy	Specificity	Sensitivity	AUC
CNN (Publish result)	302,149 (Use all)	87.47%	99.57%	71.67%	0.8520
FS + CNN	10000	66.73%	65.82%	67.64%	0.7432
FS + FE + CNN	10000	88.44%	96.51%	85.93%	0.9094

- Method with both feature selection and feature extraction has the best performance.
- The results are better than benchmark paper.
- Method with only feature selection does not achieve very well performance compare to the other groups.

HC vs MCI group



FS + CNN



FS + FE + CNN

Results, Analysis And Discussion

The performance comparison across different classification tasks and model configurations reveals that integrating feature selection (FS) and autoencoder-based denoising significantly improves classification outcomes across all groups: AD vs HC, AD vs MCI, and HC vs MCI.

Chapter 6

Conclusion

USE OF DEEP-LEARNING APPROACH TO CLASSIFY ALZHEIMER'S
DISEASE OR MILD COGNITIVE IMPAIRMENT

Conclusion

Research Objectives	Achievements
a) To investigate the impact of feature reduction on SNP data for improving classification performance.	<ul style="list-style-type: none"> Feature selection techniques significantly reduce the dimensionality of the SNPs dataset while remove irrelevant and redundant features. Feature extraction method used which is autoencoder successfully extract important biological information of the SNPs
b) To implement a CNN-based classification model with feature reduction, including both feature selection (filter methods) and feature extraction (Autoencoder techniques) to classify Alzheimer's disease and Mild Cognitive Impairment.	<ul style="list-style-type: none"> Feature reduction is implemented to the ADNI dataset before input the data to a CNN model to classify AD, MCI, and HC classes in this research.
c) To evaluate the performance of the CNN-based classification model in terms of accuracy, sensitivity, specificity, and Area Under Curve (AUC).	<ul style="list-style-type: none"> .The performance of CNN-based classification model is better compared to benchmark paper.

Future Works

- a) In future, this research can continue with finding out a better solution to handle the missing genotype values which can increase the accuracy of the model by using less features.
- b) The final features that are ranked can be validate and verify by including biological validation and verification.

THANK YOU



univteknologimalaysia



utm.my



utmofficial



UTM
UNIVERSITI TEKNOLOGI MALAYSIA

***Innovating Solutions
Menginovasi Penyelesaian***