University of Nottingham
UK | CHINA | MALAYSIA

An Ensemble Learning-Driven Pipeline for Alzheimer's Disease Classification Using Genetic based Multimodal ADNI Data

Dissertation submitted September 2025, to the University of Nottingham in partial fulfilment of the degree MSc Computer Science (Artificial Intelligence).

Shabbir Kutbuddin
20714352

Supervised by Dr. Armaghan Moemeni

School of Computer Science
University of Nottingham

I declare that this dissertation is all my own work, except as indicated in the text

Shabbir K

Signature

11th September 2025

Date

Abstract

This dissertation uses multimodal data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) to create and assess an end-to-end pipeline for the three-way classification of participants with Alzheimer's disease, mild cognitive impairment, and cognitively normal. The pipeline uses LD-pruned PCA for population-structure correction, an in-cohort GWAS for variant ranking, and rigorous PLINK-based quality control from whole-genome genotype data and cross-sectional FreeSurfer MRI measures. The final analysis set contains 794 persons with 1.43M high-quality autosomal variations and standardised structural MRI characteristics. A stacked "super-learner" design is used in predictive modelling: Level-1 uses a tailored LightGBM meta-learner on out-of-fold probability meta-features, while Level-0 combines heterogeneous base learners trained within leak-proof folds. Feature engineering yields compact, information-dense views while maintaining modality-specific signals. To prevent optimistic bias, the evaluation uses stratified nested cross-validation and reports macro-ROC-AUC, macro-F1, balanced accuracy, and overall accuracy.

Genetics+MRI models consistently outperform unimodal baselines in studies. The optimal setup, with χ^2-selected views and the entire four-model Level-0 stack, achieves balanced accuracy=0.904, accuracy=0.903, macro-AUC=0.978, and macro-F1 0.905. Clinically consistent error topology is evident, with misclassifications concentrated between adjacent stages (MCI↔CN/AD) and no direct CN↔AD leaps, suggesting well-calibrated boundaries. Strong association modelling and sufficient stratification control are supported by the APOE-centric signal with low inflation, as indicated by the Manhattan/QQ GWAS diagnostics.

The methodology offers a repeatable path from unprocessed genotypes and MRI to validated multiclass diagnosis, addressing common flaws in AD ML research. Combining morphometric markers and genome-wide signals improves over either modality alone while maintaining interpretability and calibration suitable for translational application. Limitations, deployment issues, longitudinal trajectories, external validation, and model interpretability are discussed in the conclusion.

Keywords- Alzheimer's disease; ADNI; multimodal learning; whole-genome data; FreeSurfer MRI; GWAS; population stratification; LD-pruned PCA; feature selection; Chi-Square; ANOVA-F; L1; mutual information; truncated SVD; stacked ensemble; LightGBM; Random Forest; XGBoost; MLP; nested cross-validation; ROC-AUC; macro-F1; PLINK; APOE.

Acknowledgements

I want to convey my sincere gratitude to Dr. Armaghan Moemeni, my supervisor, for her steadfast support, constructive criticism, and patient mentoring throughout my dissertation. I also want to express my gratitude to the academic and administrative team of the University of Nottingham's Department of Computer Science for fostering an environment that is conducive to study and research.

Without access to data from the Alzheimer's Disease Neuroimaging Initiative (ADNI), this dissertation would not have been feasible. To the ADNI investigators, study coordinators, and most importantly, participants and their families, whose dedication to research makes such studies possible, I express my sincere gratitude. I also value the input from the larger research community, whose conversations, documentation, and open-source tools influenced the analytical process created here.

Lastly, I want to express my sincere gratitude to my family for their unwavering support, tolerance, and love. Their confidence in me has consistently inspired me. Thank you to everyone who helped along the path, both directly and indirectly.

Table of Contents

List of Figures

List of Tables

1. Introduction

Alzheimer's disease (AD) is a global health issue with significant social and economic consequences. Early detection is crucial to stratify risk, design supportive interventions, and test disease-modifying tactics. However, the complex aetiology of AD, involving genetic predisposition and neurobiological alterations over time, complicates this goal. As biomarker science and observational platforms advance, single-gene hypotheses have been replaced by polygenic, systems-level accounts that encourage integrative techniques to accurately link genotype to phenotype.

From a genetic perspective, widespread variation in APOE modulates risk for late-onset disease in the general population, while uncommon, highly penetrant polymorphisms in APP, PSEN1, and PSEN2 account for a small percentage of early-onset family cases. Polygenic risk scores (PRS) seek to summarise genome-wide risk, but they have limitations due to ancestral composition, individual estimates, and the inability to capture rare or complex variation. Significant cost reductions in whole-genome sequencing (WGS) have enabled thorough variant discovery outside of array-based genome-wide association studies (GWAS), paving the way for richer genetic risk models.

The Alzheimer's Disease Neuroimaging Initiative (ADNI) provides a useful framework for pursuing models by offering standardised clinical, imaging, and genetic data spanning normal aging to dementia. However, determining AD status using WGS and MRI is challenging due to the "large-p, small-n" regime, which affects genome-scale data and can lead to overfitting and unstable generalisation.

To address this, the current dissertation creates and assesses a leak-resistant, end-to-end pipeline. It starts with WGS-grade QC, explicit population structure correction, disciplined feature reduction, and cautious validation to avoid leakage between training and assessment folds.

The pipeline combines an in-cohort GWAS, LD-pruned principle component analysis (PCA) for ancestry factors, and strict PLINK-based QC to give a biologically informed ranking of variants for machine learning. Four supervised selectors-ANOVA F, χ^2, L1, and mutual information-are followed by truncated SVD to produce compact, information-dense views as part of feature engineering, which progresses per-fold to prevent leaking. Out-of-fold probability features are generated at Level-0 by base learners, including Random Forest,

XGBoost, LightGBM, and a multilayer perceptron. At Level-1, these meta-features are combined by a tuned LightGBM meta-learner. Stratified K-fold cross-validation with nested tuning on inner folds to avoid optimistic bias is used to evaluate this design for genetics-only, MRI-only, and a joint Genetics+MRI experiment.

Methodologically, several decisions are worth noting. First, the validity of downstream inference depends on the pipeline's QC and stratification controls, which are crucial. These GWAS p-values function as a supervised, biologically grounded filter for ML feature space construction. The GWAS Manhattan and Q-Q diagnostics show a clear APOE-centric signal and minimal test-statistic inflation, consistent with robust association modelling and well-corrected population structure. Second, while maintaining modality-specific signal, per-fold, view-based feature learning plus SVD offers regularisation that reduces variance. Third, compared to single models, the stacked ensemble reduces variation and improves calibration by using the complementary error profiles of heterogeneous learners, a benefit noted in tabular biomedical tasks.

These design choices are supported by empirical findings. Combining genetics and MRI consistently outperformed unimodal configurations on macro-averaged ROC-AUC and F1 across experiments, with MRI-only significantly weaker and genetics-only following closely behind. This is likely due to the small class-separating signal in single-visit morphometry compared to the variety of genome-wide markers. The top-ranked configuration achieved an out-of-fold macro-AUC of 0.978, macro-F1 of 0.905, balanced accuracy of 0.904, and overall accuracy of 0.903 over 793 subjects by combining χ^2-selected features with the whole four-model Level-0 stack. Error patterns were clinically intuitive, with no CN↔AD "jumps" and confusion centred between MCI and neighbouring classes, supporting the validity of the learned boundaries.

This dissertation provides a defined, repeatable route from raw WGS and harmonised MRI to a validated three-way diagnostic classifier. It connects cutting-edge ensemble learning, leak-proof evaluation, and rigorous statistical genetics (QC, stratification control, GWAS). This addresses persistent holes in the literature on AD ML, including excessive reliance on imaging or PRS-only summary, inadequate ancestry correction, and optimistic metrics from non-nested or incorrectly segmented validation. The next sections describe the dataset and phenotype construction, explain the QC, GWAS, and modelling pipeline, present comparison

trials between selectors and learners, and address implications, limitations, and potential extensions based on available data.

2. Literature Review

2.1. Introduction and Research Problem

Alzheimer's disease is still one of the biggest issues in modern medicine because of how common it is throughout the world and how much it costs governments to provide the necessary care. It is a neurodegenerative condition that progressively develops over time and may not even manifest clinical signs for several years. Additionally, it is the primary cause of dementia in older adults. Medical practitioners usually characterize it as a deterioration in cognitive ability and memory loss. According to the World Alzheimer Report [1], there were an estimated 50 million AD sufferers worldwide in 2019; this number is only projected to increase significantly as the population ages, as aging after the age of 50 is thought to be one of the strongest risk factors, and symptoms typically manifest after the age of 60 [2]. Because AD develops gradually and silently, there is a great need for technological advancements that can make predictions with enough accuracy to enable medical professionals to treat these patients long before AD symptoms appear or the neurodegeneration process begins. The foundation of such technology necessitates a very deep understanding of the disease's causes, which have been determined over decades of research to entail the complex interplay of genetic features and dynamic biological processes. In the past three decades, scientists have moved away from the idea of identifying a single significant genetic element and toward a more complex understanding of the illness that is impacted by several genes, frequently in combination with environmental factors. The development of biomarker technologies has played a significant role in this shift in perspective.

There has been a recent push to identify Alzheimer's disease early on, before it progresses to the dementia stage. In order to help patients and their families understand the situation and make lifestyle changes that may slow the progression from Alzheimer's to dementia, the Alzheimer Cooperative Valuation in Europe (ALCOVE) project suggests diagnosing Alzheimer's disease earlier, when patients and their families can notice a decline in cognitive function. (also known as "timely diagnosis") [3].

Early or timely Alzheimer's disease diagnosis is fraught with difficulties. Concerns about privacy, loss of status, being associated with a stigmatized label, losing one's work, and in

rare instances, depression are some of these problems [4]. Misdiagnosis presents another difficulty since it may lead to inappropriate treatment for diseases that can be cured or needless medication for Alzheimer's [5]. It can be expensive for society to set up mechanisms that offer early or timely diagnosis and intervention.

The hunt for genetic causes of AD has lasted for years. Three genes that can be altered to cause EOAD are presenilin 1 (PSEN1)[7], presenilin 2 (PSEN2)[8], and amyloid beta (A4) precursor protein (APP)[6]. While most of these mutations are dominantly inherited, some are not fully penetrant. APP, which is found on chromosome 21, was one of the first genes connected to Alzheimer's disease. PSEN1 is found on chromosome 14. PSEN1 mutations are responsible for a higher percentage of people with Early Onset Alzheimer's Disease (EOAD) (18–50%) than mutations in the other two genes. PSEN2 is found on chromosome 1. Mutations in PSEN2 that cause EOAD are less common than those in PSEN1.

One method used by researchers to conduct this inquiry and find potential genetic explanations is whole genome sequencing (WGS). SNP microarrays are used in whole genome association studies (GWAS) to identify genetic traits and the chance of disease. Although they have identified risk variations for a number of diseases, they frequently only explain a small percentage of the hereditary risk. This is caused by a number of variables, such as common mutations having little effect or differential penetrance due to epistatic (one gene influencing another) or epigenetic (gene expression changes lacking changes to the underlying DNA sequence) influences. Although copy number variants (CNVs) and rare variations have a major influence on disease symptoms, they are highly challenging to detect with the genotyping microarray technology available today. By gathering the fullest collection of rare variants and structural variation from sequenced individuals, whole genome sequencing offers a solution. As costs come down, I expect a paradigm shift in technology, even though it is currently too costly for widespread use [9].

Estimating the genetic effect is challenging due to the high dimensionality of genome-wide association study (GWAS) data, which contains many more markers (M) than persons (N). Due to a lack of statistical power, variations with small effect might not be discovered under a strict p-value threshold (~10-8) [10]. This challenge, which is more commonly known as the "large p small n" problem, is one of the numerous factors that lead to the well-known concept in machine learning called the "curse of dimensionality." The existence of numerous correlations between independent variables exacerbates this issue [11]. Since there are nearly two million markers (p) in WGS data compared to just 600k in GWAS data in the ADNI

Database, the "large p small n" problem is considerably more difficult to solve in WGS data than in GWAS data.

This is where machine learning can be useful. Machine learning and AI developments have resulted in the use of many classifiers and clustering approaches [12]. Medical informatics has gained attention from the data science research community in recent years due to the widespread adoption of computer-based technology in the health sector, including electronic health records and administrative data, and the availability of large health databases for researchers [13]. The curse of dimensionality could possibly be addressed through feature selection utilizing machine learning. Feature Selection is an effective preprocessing strategy in data mining that adds dimensions to data by finding essential risk features, which is critical in medical diagnosis. Identifying significant characteristics in medical datasets helps to remove non-contributing and duplicate features, resulting in faster and more accurate predictions [14].

Despite abundant work on AD prediction, most studies either rely on imaging alone or summarize genetics with polygenic risk scores that miss non-linear interactions; few present a fully leak-proof, multiclass pipeline that begins with Whole Genome Sequencing-grade quality control, corrects for population structure, and evaluates models under stringent out-of-fold validation. This project therefore formulates the following research problem: design and validate a reproducible pipeline that transforms raw ADNI whole-genome data and UCSF Cross-Sectional FreeSurfer measures into compact, well-behaved feature spaces and trains a stacked super-ensemble to classify CN, MCI and AD. The study asks whether joint Genetics+MRI learning significantly improves macro ROC-AUC and PR over unimodal models once sex/missingness/MAC/HWE filters and LD-pruned PCA covariates are enforced, and whether the resulting predictors are stable and interpretable across folds. Success would demonstrate a principled path from WGS QC to clinically meaningful, multi-modal staging of Alzheimer’s disease.

2.2 The Genetic and Genomic Landscape of Alzheimer's Disease

2.2.1 Established Genetic Risk Factors

Mutations in the genes APP on chromosome 21, PSEN1 on chromosome 14, and PSEN2 on chromosome 1 are the etiology of Alzheimer's disease. An further important genetic risk factor is the APOE-4 allele. Six genotypes are produced by the three polymorphic alleles (ε2, ε3, and ε4) of the APOE gene. ε2 is the least common allele (8%) whereas ε3 is the most

prevalent (77%) [15]. The frequency of the ε4 allele is approximately 15% in the general population, but it is closer to 40% in Alzheimer's patients. Alzheimer's disease is about three times more likely to strike people who carry one ε4 allele [16]. The majority of organs-liver, brain, spleen, lung, adrenal, ovary, kidney, and muscle-produce APOE [17]. Two-thirds to three-fourths of plasma APOE is produced in the liver, which also produces the highest APOE mRNA. About one-third of the APOE mRNA found in the liver is found in the brain, which has the second-highest quantity [18]. Apo-E is produced by a variety of cell types in organs and is probably what causes the RNA that is there. It is currently unknown what neurological mechanism underlies the APOE-4 allele's elevated risk of Alzheimer's disease. Prior studies have connected APOE-4 to increased levels of neurofibrillary tangles and brain plaques, indicating a quicker build-up of disease [19]. Despite having a substantial correlation with AD, APOE-4 has no function in the progression from AD to dementia since it does not contribute to cognitive deterioration. There is no relationship between the quantity of APOE-4 alleles and the amount of senile plaques and neurofibrillary tangles in the frontal cortex or hippocampus of people with vascular dementia, Alzheimer's disease, or control groups, according to studies [20]. Although APOE-4 affects a person's risk of developing AD, it has no effect on cognitive decline after dementia has already set in.

A child has a 50% chance of receiving a changed form of one of the three genes linked to Alzheimer's disease (APP, PSEN1, or PSEN2) if one of their biological parents has a genetic variant for that gene. The child has a significant chance of getting Alzheimer's disease before the age of 65, if it is inherited. Less than 10% of all Alzheimer's patients have early-onset Alzheimer's disease (EOAD), and 10% to 15% of those instances are caused by mutations or polymorphisms in APP, PSEN1, and PSEN2 [15].

2.2.2 Limitations of Traditional Approaches

2.2.2.1 Inadequacy of Polygenic Risk Score (PRS)

The Genetic Risk Score (GRS) is frequently used to quantify the effect of multiple genetic variables on a particular phenotype or disease. This method predicts an approximate of the likelihood of a desired outcome based on genetic variants. GRS aggregates the impacts of allelic combinations on phenotype by adding k separate genetic variants with a strong link, based on individual effect impact and P-value. The equation used for GRS is shown in Figure 1.

$$GRS_j = \sum_{i=1}^{k} \beta_i N_{ij},$$

Figure 1. Formula to calculate Genetic Risk Score (GRS)

where the effect size (βi) is estimated using log-odds ratios from logistic regression with additive genetic effects for binary traits or coefficients from linear models for quantitative traits associated with a single allele count, multiplied by the number of alleles (Ni) at a given locus in individual (j)[22]. To improve prediction, the GRS can be further improved upon to include loci with little effects and no significant genome-wide connections, resulting in a polygenic risk score(PRS). This technique is more useful for complicated features like schizophrenia, height, and primary open-angle glaucoma, which lack common risk variations with a big effect that is apparent. PRS can identify every variation that could be a cause for the phenotype/disease in the genotyping panel using single-marker association testing. The only difference between GRS and PRS score is the number of SNP-sites considered (k). By incorporating weak connections, the score becomes more 'poly-gene-informative' than the GRS, allowing for more exact identification of high-risk people. [23]. While drug dosage guidelines have been heavily based on single gene mutations (single nucleotide polymorphisms) over the last decade, polygenic risk scores (PRS) have emerged in recent years as a promising tool for taking into consideration the complex and dynamic polygenic nature of patients' genetic traits influencing drug response. [24]. PRSs provide individualized risk assessments, helping individuals understand their genetic vulnerability to diseases and make informed decisions based on it. PRS benefits healthcare and research by speeding up genetic research, reducing costs for risk assessment, and identifying disease risks at the population level. However, PRS comes with its set of limitations as well. Primarily is limited by the number of SNPs tested and the population on which the model is developed. More importantly, PRS can make biases in different ethnicities very apparent as the models are trained with genetic data from a group of one ethnicity and the model could show incorrect risk assessments when it is tested on a group of individuals from another ethnicity[22]. Another limitation is that present genetic prediction software lacks standard error metrics for individual PRSs, which limits the adaptation and usage in a clinical setting. Methodologies to estimate standard error measurements for individual PRSs are mainly unknown [25]. Furthermore, rare pathogenic alleles are often not considered in PRS produced from GWAS

summary data, as GWAS typically contain only "common" variants with a population frequency of 1% or above [26].

Single Nucleotide Polymorphisms (SNP) which are in a form of an array are biallelic genetic markers that are widely distributed throughout genomes, making them straightforward to examine and interpret. Biallelic means that both alleles of a gene (one inherited from each parent) are impacted by a genetic change or variation. SNPs can identify genes associated with human disorders by capturing linkage disequilibrium (LD) information present in the genome [27]. SNP arrays have its limitations as well.

Continuous regions of the genome where people are homozygous everywhere are known as runs of homozygosity (ROH). This means that for a particular place, they have two identical copies of a gene. These areas may be the result of inbreeding or other population genetic patterns and are inherited from a common ancestor. SNP arrays can identify ROH longer than 1 Mb because they encompass roughly 1.9 to 2.2 million SNPs. But just around 2% of all common SNPs in the human genome are represented by this. This hinders the use of SNP arrays to identify shorter ROH, which is crucial for expanding our understanding of human genetics [28]. In the majority of AD cases, common variations only make for 33% of the genetic variance, according to heritability estimates. This implies that the genetic makeup of AD is more intricate than the "common disease, common variant" theory that is examined by SNP arrays found in GWAS. ultra-rare, uncommon, and low-frequency variation, commonly referred to as 'rare variations', may be responsible for the missing heritability of Alzheimer disease [29]. Whole Genome Sequencing may be more helpful in this situation.

2.2.3 The Promise of Whole Genome Sequencing

The cost per human genome carried out by Whole Genome Sequencing(WGS) has decreased dramatically in recent years, as shown in figure 2[30], and has even fallen below the benchmark target of $USD 1000, which had been set since the beginning of genome sequencing. As of 2022, the cost per human genome was $525, which is over 180000 times less than the cost in 2001.

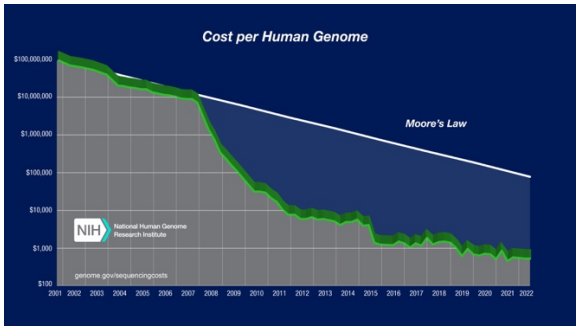

Figure 2. Cost for Whole Genome Sequencing[30] has gone down drastically since 2007 and it costs less than $1000 USD as of 2022.

With the costs per genome exponentially reducing, it has sparked an increased level of attention on WGS. It investigates current advancements in the intersection of fundamental genomic research and the clinical implications of the presence or lack of specific genes. WGS is now one of the most used applications, producing massive amounts of genome sequences in comparison to previous public and commercial human genome sequencing initiatives around the world[31]. Massive volume of data can identify disease-causing alleles that were previously undetectable.

Affordable genomic technology has the potential to enhance health care by accelerating application adoption[32]. One benefit of an increased rate of individuals carrying out WGS on their Genome is that genomics provides the benefit of storing and analysing data as more individuals with similar traits are identified, even if initial analysis does not yield significant signals for the phenotype being tested[33]. Targeted assays cannot achieve this. With the number of samples increasing, pattern recognition within genomes with individuals having a phenotype could be made easier and efficient. Another benefit of WGS is that it can be very useful in identifying possible genetic traits that can contribute or lead to very rare diseases. A research from the 100,000 Genomes Project found that Whole Genome Sequencing (WGS) helped detect 25% of rare diseases[34]. Alzheimer's is not a rare illness, but Early Onset Alzheimer's is. Early-onset Alzheimer's is an uncommon form of the illness that affects those under the age of 65 at diagnosis. Approximately 5% of the 6.5 million Alzheimer's patients in the United States have the early-onset type of the illness [35].

2.3 Machine Learning Applications in Alzheimer’s Disease Prediction

2.3.1 Early and Traditional Machine Learning Models

Significant advances have been made in the study of Alzheimer's disease in recent decades, while early machine learning (ML) applications concentrated on clinical and neuroimaging data employing rather basic classifiers. One such straightforward model is called Support Vector Machines (SVM), a supervised machine learning technique that uses the best line or hyperplane that optimizes the distance between each class to classify data with two classes (e.g., Male/Female, Yes/No) in an N-dimensional space [36]. Another well-liked machine learning model for categorization is logistic regression. A supervised machine learning classifier called logistic regression evaluates input real-valued characteristics, multiplies them by a weight, adds them up, and computes probability using a sigmoid function. A decision is made based on a threshold. Multinomial logistic regression is used for text classification and part-of-speech identification, and logistic regression can be used to two or more classes, such as positive and negative sentiment. The softmax function is used to calculate the probabilities in multinomial logistic regression. Using a loss function, like the cross-entropy loss, which needs to be minimized, the weights (vector w and bias b) are learned from a labelled training set [37]. Decision trees are another well-known machine learning model that is used for supervised classification. Its purpose is to categorize a single distinct target feature. Every internal node runs a Boolean test on an input feature, with a yes/no response possible. It is possible to convert a test with multiple possible outcomes into a set of Boolean tests. The values of the input feature are used to identify edges. A value for the target characteristic is represented by each leaf node [38]. To lessen connection between feature data, the machine learning algorithm Random Forest employs several decision trees. By selecting samples and features at random, Random Forest (RF) lowers correlation between decision trees. In the original training data, an equal number of data points are first chosen at random from the training sample. Additionally, the decision tree is constructed using a random subset of the features. By reducing the correlation between each decision tree, these two randomization techniques help to improve the accuracy of the model by reducing the potential mistake caused by overfitting [39].

There has been a paradigm shift in the quantity of research conducted with the aid of technology with the usage of basic and conventional machine learning models. Observational studies carried out extensively and over an extended period of time have had a significant influence on and accelerated this technological research evolution. The Alzheimer's Disease Neuroimaging Initiative (ADNI) is one such significant observational study. The ADNI initiative provides active support for research and development of treatments that delay or stop the progression of Alzheimer's disease (AD). Researchers collected data from over 60

clinical locations in the USA and Canada to examine how AD develops in the human brain spanning normal aging, dementia, Alzheimer's disease, and moderate cognitive impairment (MCI)[40]. The ADNI project began collecting data in 2004 and has continued to do so throughout its many phases. The current trial phase is ADNI4 (2022-2027). The ADNI study assesses the structure and function of the brain across three disease states (cognitively normal/unimpaired, mild cognitive impairment, and dementia) using clinical measures (cognitive and neuropsychological tests) and biological markers (biomarkers; such as chemicals in blood or changes to the brain seen in MRI and PET scans). By providing biospecimens (samples) and study data to qualified researchers worldwide, ADNI has established itself as a global resource for AD progression research.

Since the database's creation, academics have utilized basic machine learning classification methods on the ADNI Database. In one such study [41], the progression from mild cognitive impairment (MCI) to Alzheimer's disease (AD) was predicted using logistic regression on baseline data from the Alzheimer's Disease Neuroimaging Initiative (ADNI). Sparse logistic regression with stability selection was utilized in the study to find predictive biomarkers in 319 MCI patients by combining MRI, genetic, demographic, and cognitive data. High prediction accuracy (AUC = 0.8587) was shown for a set of 15 parameters, including APOE genotyping, cognitive test scores, and MRI-derived brain sizes. Another researcher [42] accurately identified Alzheimer's disease (AD) in many clinical centers by using SVM with FDG-PET and MRI data. 21 AD patients and 13 controls from their own data collecting center in Leipzig were compared to data from 28 AD patients and 28 controls from ADNI. SVM outperformed single modalities with integrated imaging using volumes of interest (VOIs) based on meta-analyses. Cross-cohort validation hit 91%, and accuracy rates in the Leipzig data cohort and ADNI were 100% and 88%, respectively. According to the study, the combination of SVM and multimodal imaging improves the generalizability and reliability of the diagnosis across various data points and patients. Using Random Forests, a research team [43] was able to win an international challenge for automated MCI prediction from MRI data. This study used structural and multimodal neuroimaging data from ADNI to try to categorize AD and forecast the course of moderate cognitive impairment (MCI). In addition to its inherent feature selection capabilities, RF is renowned for its resilience to noise, outliers, and non-linear data. They demonstrated the efficacy of RF in binary and multi-class classification tasks using MRI, PET, DTI, and fMRI data. Their RF-based ensemble model classified AD with 76% accuracy in real-world subjects.

2.3.2 Transition to Genomic Data

In addition to image and clinical data, ADNI's database contains genomic data. From the standpoint of AD research, genomic data in ADNI has been one of the most important sources of data fueling the surge in the usage of WGS data for medical research in recent years. Using information from 374 non-Hispanic Caucasian ADNI patients, one study [44] examined the genome-wide association study (GWAS) of biomarkers linked to AD in cerebrospinal fluid (CSF). Genetic influences on Aβ1-42, total tau (t-tau), phosphorylated tau (p-tau181p), and their ratios were examined. The genome-wide significance of four SNPs-APOE, TOMM40, LOC100129500, and EPC2-was demonstrated, with EPC2 emerging as a putative tau disease candidate gene. Other associations with CCDC134, ABCG2, SREBF2, and NFATC4 were found. The results highlight the significance of replication in larger cohorts, support the significant impact of APOE, highlight the relevance of TOMM40, and suggest other pathways. In order to find quantitative trait loci (QTLs) connected to moderate cognitive impairment (MCI) and Alzheimer's disease (AD), another study [45] investigated a genome-wide association analysis of brain-wide imaging characteristics in the ADNI cohort. Grey matter density, cortical thickness, and brain volumes were among the 142 characteristics that researchers examined using 1.5T MRI data and more than 530,000 SNPs from 733 patients (AD, MCI, and healthy controls). Strong correlations with APOE and TOMM40 were verified by the results, and new loci affecting hippocampus and global grey matter density were found close to NXPH1, EPHA4, and TP63. According to research, genomic imaging may reveal novel molecular mechanisms that underlie the development of AD.

Combining genomic data with machine learning has been used for research in other fields as well. One such study [46] looked into the potential applications of genetic information and machine learning (ML) in forecasting the efficiency of anaerobic digestion, a process that turns organic waste into methane for sustainable energy. Six machine learning approaches were used to analyze data from eight research groups, including digester operating conditions and microbial DNA sequencing. Neural networks predicted methane yields the most accurately, whereas Random Forest generated the highest classification accuracy (82%), when genetic and operational data were combined. The aforementioned "large p small n" problem, which arises when there are a million or even more features and only a relatively small number of samples, is one of the major issues with genomic data. Because of this, choosing a significant subset of features is essential to reducing the data's dimensions. The likelihood of overfitting, which happens when a model learns the training data too well by

incorporating noise and irrelevant data points, is quite high if feature selection is done incorrectly. This results in poor performance on fresh and unknown data.

2.3.3 Ensemble Learning and Advanced Models

Several machine learning methods are used in ensemble learning to provide mediocre predictions based on data projections. Voting procedures are then added to these outcomes to enhance performance over separate algorithms [47]. Figure 3[48] depicts the flow of ensemble learning models.

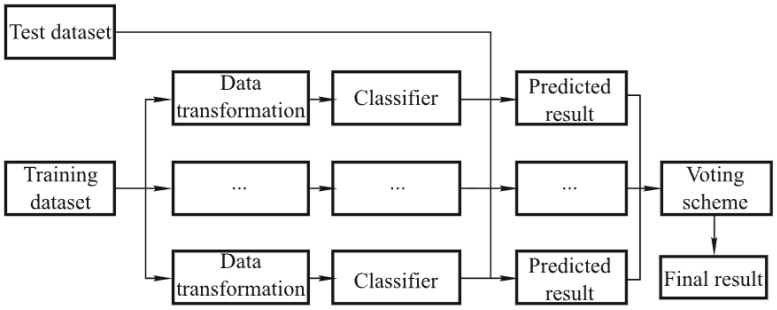

Figure 3. Ensemble Learning Model flow

The three primary categories of ensemble learning machine learning techniques are bagging, boosting, and stacking. Using random sampling and replacement, Bagging (Bootstrap Aggregating) first creates a number of distinct subsets of the training data. A different model is then used to train each subset. The final forecast is then produced by combining the output of all models, either by majority vote for classification or by averaging the results for regression jobs. Bagging has two main advantages: it reduces variance, which reduces the likelihood of overfitting, and it improves accuracy because integrating many models usually performs better than using just one. Unlike bagging, boosting trains models in a sequential fashion, with each new model concentrating on fixing the shortcomings of the previous one. Every data point is given a weight during training. A point gains weight and receives more attention in the subsequent model if it is misclassified. Ultimately, the conclusion is reached by integrating the predictions of all models, typically through weighted voting or averaging. Because it eliminates bias and turns weak learners into powerful forecasters, boosting increases accuracy. A large number of basic models (level-0) that have been trained on the same dataset are the starting point for stacking (stacked generalization). A new model called the meta-model (level-1) then uses their predictions as input features to determine how to optimally integrate these outputs to get the final prediction. The employment of many models gives stacking its strength by enabling it to identify a broad variety of patterns in data.

Comparing this approach to utilizing only one model, accuracy and performance are usually improved [49].

Another cutting-edge ML idea that has gained a lot of traction in the last 20 years or so is deep learning. By using neural networks, which are modelled after the human brain, Deep Learning (DL) techniques are a type of Representation Learning techniques that allow a machine to automatically identify the representations needed for classification tasks just by feeding it raw data. DL uses several layers of simple, nonlinear modifications to process input. It begins with unprocessed data and works its way up to more meaningful and abstract representations. If enough layers are added, deep learning models can learn incredibly complex functions. The higher layers focus on the important components that aid in differentiating between groups while eliminating extraneous information in tasks like categorization. Deep learning can automatically extract patterns and insights from data thanks to this layered approach, which makes it perfect for applications like speech understanding, photo identification, and natural language processing [50]. The input layer accepts raw data, the hidden layers process it, and the output layer outputs the outcome. Neural networks are composed of nodes, or neurons. To improve accuracy, a weight is assigned to each connection and changes throughout training. Using a method called backpropagation with gradient descent, the model learns by comparing its predictions to the right answers and adjusting weights.

Figure 4[51] provides a visual representation of this process.

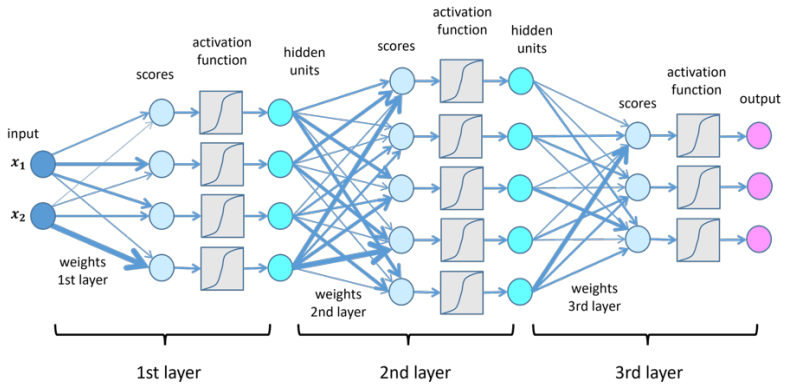

Figure 4. How deep learning works in summary.

Using the ADNI Database, several intricate methods have been used to forecast Alzheimer's disease. One study [52] implemented a multistage classifier-based method for early Alzheimer's disease (AD) prediction using MRI images from the ADNI Dataset. FreeSurfer was used to retrieve MRI features like brain volume and cortical thickness, and Particle Swarm Optimization (PSO) was used to select features. SVM, KNN, and Gaussian Naive

Bayes were used in a two-stage classification process. The results demonstrated that PSO-enhanced cortical thickness attributes provided the largest bias, with the system outperforming single classifiers and achieving over 96% accuracy. Another study [53] used resting-state fMRI (R-fMRI) brain networks in conjunction with clinical information (APOE genotype, age, and gender), which was available in ADNI, in an effort to make an early diagnosis of AD. R-fMRI signals from 91 people with mild cognitive impairment (MCI) and 79 participants with normal control (NC) ADNI were used to construct functional link networks. Compared to traditional classifiers (LDA, LR, and SVM), a multilayer autoencoder was trained to collect discriminative features with significantly higher precision and consistency. With improved sensitivity (0.92) and specificity (0.81), the model's accuracy was 86.47%.

2.4 Overcoming Methodological Weakness in Genomic Machine Learning

2.4.1 Perils of Overfitting

From a machine learning standpoint, overfitting occurs when a model has a tendency to retain all of the training data rather than comprehending the patterns that impacted the target variable. Three causes are largely responsible for this [54].

The first reason is thought to be noise learning on the training set. The model is more likely to learn from noise, which might affect predictions, when the training set is small, has few data points that show patterns, or contains a lot of noise. The second issue is hypothesis complexity (bias-variance trade-off). The ratio of accuracy to consistency is balanced. While highly complex models (with an excessive number of inputs, features, or assumptions) may perform well on training data, their performance varies greatly on new datasets. Although simpler models are more reliable, they could miss some significant patterns. The third factor is multiple comparisons, which involves choosing winners at random. Chance can make certain options seem great when I test out multiple models or features and choose the one with the highest evaluation score. On fresh data, the "winner" might not be superior and might even lose accuracy.

There are three primary strategies to prevent or lessen the likelihood of overfitting. Dimensionality Reduction is the most popular. It is composed of two techniques, namely Feature Selection and Feature Extraction. When raw inputs do not transfer cleanly to what a model can use, feature extraction helps reduce the original high-dimensional data to a smaller

collection of new features (linear or nonlinear mixes of the originals). Principal Component Analysis is one type of feature extraction (PCA). These additional features, however, lose their obvious practical use. Rather, feature selection-a technique employed in domains like text mining and genetics-maintains a more manageable, more pertinent subset of the initial features, preserving their physical significance and simplifying the interpretation of models. Both approaches are useful even when there aren't many features because they can increase accuracy, reduce overfitting, speed up training, and conserve memory. You can retain the most valuable inputs for a model by using feature selection techniques. This is accomplished via a variety of techniques. Features that seldom change across samples are eliminated by the variance threshold, and features with too many gaps are eliminated by the missing value ratio. ANOVA ascertains whether the values of a feature are connected to distinct outcomes; While LASSO applies a penalty that lowers irrelevant feature weights to zero and subsequently excludes them, Recursive Feature Elimination (RFE) repeatedly trains a model and prunes the least valuable features. Reducing noise, maintaining signal, and customizing any penalties to eliminate irrelevant features without sacrificing crucial information are the objectives in each situation [55].

2.4.2 Rigorous Validation Frameworks

A labelled dataset, consisting of a set of characteristics designated as X and a target variable designated as y, is a prerequisite for any supervised machine learning model. For instance, a dataset pertaining to home prices might include features such as the house's square footage, number of rooms, etc., and the feature column would include the home's price. After that, this information is separated into training datasets, which are used to train the model for either a regression or classification task. The performance of the trained model is evaluated on an unknown dataset known as the testing or validation set after it has been trained on a subset of samples from the full dataset. This set has the remaining samples after the training set is separated from the full dataset. Assessing the model's ability to classify or predict using fresh data that wasn't used for model training is the aim of cross-validation, which may help detect model problems like selection bias or overfitting early on [56]. K-fold Cross Validation is a widely used cross-validation technique that splits a dataset into K partitions of roughly similar size. The model is constructed using (K-1) folds, and the validation is done using the excluded sample. Each of the K folds is assigned as validation data one after the other in this K-times-repeated process. When implementing, a K value of 5 to 10 is typically selected [57]. Despite being a far superior cross-validation technique to a static train-validation set

split, K-fold cross validation is not without its drawbacks. The main drawback is that K-Fold cross validation can occasionally be excessively optimistic because it use the same folds for assessment and hyperparameter tuning, which can result in selection bias and produce an unrealistic estimate of how well the tuned model will perform on actual unseen data.
An effective method for combining feature selection and hyperparameter adjustment to train an optimal predictive model is nested cross validation. Data is separated into K outer folds in a general nested cross validation technique, and a set of inner folds is generated for each outer training fold. These inner folds are utilized for model training, feature selection, and hyperparameter tuning. The model is tested using an inner nest, which minimizes excessively biased and positive outer-fold CV by limiting data leakage between outer folds that may arise from feature selection [58]. This is the main justification for choosing nested cross validation over K-fold cross validation since it provides a more realistic albeit harsh generalization estimate of the model's performance with unseen data.

3. Dataset

This study uses a large, multi-modal dataset from the Alzheimer's Disease Neuroimaging Initiative (ADNI). Established in 2003, ADNI is a historic public-private collaboration led by Principal Investigator Michael W. Weiner, MD, whose goal is to understand the progression of Alzheimer's disease (AD). It accomplishes this by integrating serial data from positron emission tomography (PET), magnetic resonance imaging (MRI), comprehensive clinical and cognitive evaluations, and a variety of biological markers, such as genetics and cerebrospinal fluid (CSF) studies. An unmatched resource for examining the intricate genesis of AD, monitoring its course from preclinical stages to dementia, and creating reliable biomarkers for diagnosis and treatment trials is the ADNI cohort's longitudinal and multifaceted character [59]. This study intends to develop integrative models that reflect the combined effects of neuroanatomical alterations and genetic predisposition on disease state by utilizing this vast dataset. For up-to-date information and data access procedures, please https://adni.loni.usc.edu/ , the official ADNI website.

3.1 Participant Cohort and Phenotype Definition

3.1.1 Cohort Description and Diagnostic Criteria

812 participants from the ADNI-1, ADNI-GO, and ADNI-2 stages of the trial made up the first cohort chosen for this analysis. In order to train and validate a three-way classification

model, these subjects must reflect a range of cognitive health. Following established guidelines, ADNI clinicians at various study locations developed the clinical diagnoses that serve as the foundation for our phenotypic. These diagnoses can be divided into three main groups:

- Cognitively Normal (CN): This cohort functioned as a robust control group. They showed no signs of moderate cognitive impairment, depression, or dementia. Their Mini-Mental State Examination (MMSE) scores ranged from 24 to 30 (inclusive), their Clinical Dementia Rating (CDR) was 0, and their Wechsler Memory Scale Logical Memory II subscale scores were all within the normal range.
- Mild Cognitive Impairment (MCI): People with a CDR of 0.5, an MMSE score between 24 and 30, subjective memory issues, and objective memory loss as assessed by standardized testing are classified as having mild cognitive impairment (MCI). Additionally, they preserve regular life activities in general. Because people with MCI have a much higher risk of developing Alzheimer's disease (AD), it is a topic of intense scientific study.
- Alzheimer’s Disease (AD): People who fit the probable AD criteria set by the National Institute of Neurological and Communicative Disorders and Stroke and the Alzheimer's Disease and Related Disorders Association (NINCDS-ADRDA) fall into this group. They have a CDR of 0.5 or 1.0, an MMSE score between 20 and 26, and significant memory impairment.

After applying the rigorous quality control procedures described in Section 3.2, the final analytic cohort was whittled down to 794 people, 440 of whom were men and 354 of whom were women. The individuals' average age was 73 at baseline (the initial examination visit) and 80 at their last examination visit (the diagnostic label was also utilized for the final assessment visit).

2.1.2 Phenotype Assignment and Processing

Participants are evaluated at various intervals due to the longitudinal nature of ADNI, and their diagnosis may evolve over time (for example, from CN to MCI or MCI to AD). It is essential to give each person a singular, distinct phenotype for a cross-sectional study like this one. In order to achieve this, the most current diagnostic status for every participant was extracted from the clinical diagnosis summary file (DXSUM_13Aug2025.csv).

Using the pandas package and a custom Python script, this assignment was completed. The logic of the script went like this:

1. Data Loading: A pandas DataFrame was loaded using the diagnosis summary CSV.
2. Temporal Sorting: A standard datetime format was applied to the EXAMDATE column. Since raw data files may have conflicting date formats, this step is essential to enabling correct chronological sorting.
3. Latest Diagnosis Extraction: The DataFrame was arranged first by examination date and then by patient identifier (PTID). Next, the groupby('PTID').last() function was used. This robust method selects the final record, which, because of the previous sorting, corresponds to the most recent clinical visit, by grouping all of the records for a single patient. This guarantees that the participant's most recent cognitive state is reflected in the phenotype.
4. Numerical Mapping: To ensure interoperability with subsequent statistics and machine learning applications, the categorical diagnoses were converted to numerical integers. The script's mapping was as follows: CN → 1, MCI → 2, and AD → 3. To preserve the integrity of the three target classes, people with unclear or different diagnoses (such as "Dementia other") were not included in the analysis.
5. Integration with Genetic Data: The PLINK.fam file was combined with the final list of patient IDs and their numerical diagnoses. One line per individual, comprising columns for Family ID, Individual ID (IID), Paternal ID, Maternal ID, Sex, and Phenotype, makes up the.fam file, a fundamental part of the PLINK fileset. The newly determined diagnostic code was added to the sixth column (Phenotype) by the script.
6. Machine Learning Label Conversion: For the final stage of machine learning, these phenotype codes were converted to a zero-indexed format (**CN → 0, MCI → 1, AD → 2**), as this is the standard input format for most machine learning libraries, including scikit-learn.

This systematic and automated approach to phenotype definition ensures reproducibility and accuracy, providing a solid foundation for the subsequent predictive modelling.

3.2 Genetic Data

3.2.1 Data Source and Technology

Whole-genome sequencing (WGS) of blood-derived DNA samples from ADNI participants provides the genetic data for this investigation. The PLINK binary fileset format (WGS_Omni25_BIN_wo_ConsentsIssues) contained the raw data. According to the file name, the Illumina Omni 2.5M microarray may have been used for the initial genotyping,

followed by imputation and sequencing to provide whole-genome coverage. In contrast to targeted SNP arrays, WGS offers a comprehensive perspective of genetic variation, collecting uncommon variants and other structural differences throughout the entire genome in addition to typical single nucleotide polymorphisms (SNPs) [60].

There were 2,379,855 genomic variations for the 812 individuals in the first, pre-QC dataset. Prior to applying our more strict filters, the raw dataset's overall genotyping rate, also known as the call rate, was a high 99.74%, suggesting good initial data quality.

2.2.2 Data Format

The PLINK binary format, the de facto standard for computational efficiency in large-scale genetic investigations, was used to organize the genetic data. There are three linked files in this format:

- .bed file: A compressed binary file that contains the genotyping data. It keeps track of each person's genotype (AA, AG, GG, etc.) at every variation site. Because it is binary, it can be loaded and processed quickly, which is crucial for datasets with millions of variations.
- .bim file: A text file called that acts as a map for a ".bed" file. The chromosome, identifier (such as an rsID), position in centimorgans (often set to 0), base-pair position, and the two alleles (such as A and G) are all included in each line, which represents a single genetic variant. There were 2,379,855 records in this file in our original dataset.
- .fam file: A text file with each person's phenotype and pedigree details, as previously mentioned. There were 812 rows in this file at first, one for each participant.

By separating genotype, variation, and sample data, this threefold structure offers a versatile and computationally effective framework for the exacting quality control and analytical workflow outlined in the approach.

3.3 Neuroimaging Data

3.3.1 Image Acquisition and Processing with FreeSurfer

T1-weighted structural MRI was the neuroimaging technique employed in this investigation. To reduce scanner-specific variability, these pictures were obtained utilizing similar techniques at several ADNI facilities. At UCSF, the extensively tested FreeSurfer software suite (v7.1.1) was used to process all raw MRI scans [61]. A comprehensive morphometric investigation of the human brain is carried out by the robust, automated toolset FreeSurfer,

which generates precise quantitative measurements of brain structure.

A series of complex image analysis procedures are part of the FreeSurfer cross-sectional processing pipeline. These include:

- Motion Correction and Averaging: To improve the signal-to-noise ratio, several T1-weighted images for a subject are registered and averaged.
- Intensity Normalisation: By correcting for intensity variations brought on by magnetic field irregularities, intensity normalization makes sure that the intensity of a particular tissue type-like white matter-remains constant throughout the image.
- Skull Stripping: An essential procedure that isolates the brain for additional examination by removing the skull and other non-brain tissue from the picture.
- Volumetric Segmentation: The brain is divided into many tissue classes, mainly cerebrospinal fluid (CSF), white matter (WM), and gray matter (GM). Additionally, it names many subcortical structures, such as the thalamus and hippocampal regions.
- Cortical Surface Reconstruction and Parcellation: The pial surface, or outer boundary of the gray matter, and the gray-white matter boundary are recreated. Then, using well-known atlases such as the Desikan-Killiany atlas, the cerebral cortex is automatically divided into discrete anatomical regions of interest (ROIs) according to gyral and sulcal patterns [62].

A high-resolution brain image is efficiently converted into a structured table of numerical data appropriate for statistical analysis and machine learning by this automated workflow, which generates a sizable set of quantitative features for every subject.

3.3.2 Neuroanatomical Feature Set

FreeSurfer's output offers a wealth of features that measure the morphometry of hundreds of different brain structures. A wide range of these characteristics, including cortical and subcortical measures, were used in this investigation. These fall into the following general categories:

- Global and Subcortical Volumes: These features, which are obtained from the aseg.stats output, show the volume of different structures in cubic millimeters. The thalamus, amygdala, putamen, and left and right hippocampal volumes are important characteristics. Hippocampal atrophy is a characteristic biomarker of AD, and these areas are known to be impacted early in the disease's progression [63]. The volume of white matter hypointensities, a measure of the burden of cerebrovascular disease, and

total intracranial volume (ICV), which is essential for normalizing other volumetric measurements, are also included in the dataset.

- Cortical Thickness: This is the mean separation between the white matter surface and the pial surface within a specific cortical ROI, as determined by the aparc.stats output. One well-known early sign of AD-related neurodegeneration is cortical thinning in particular areas, such as the precuneus, temporal lobe, and entorhinal cortex. The dataset comprises the mean and standard deviation of cortical thickness for each of the 68 cortical areas found in the Desikan-Killiany atlas [64].
- Cortical Surface Area: This quantifies each cortical ROI's surface area in square millimeters. Surface area alterations can be sensitive to neurodegenerative effects and represent different biological processes than thickness changes.
- Cortical Volume: This is the amount of gray matter in each cortical ROI, measured in cubic millimeters. It offers a composite assessment of a cortical region's size and is mathematically connected to thickness and surface area.

Each participant had access to several hundred structural MRI features in total. To guarantee that the brain structure measurements appropriately reflect the participant's cognitive state at the time of categorization, the MRI scan that was temporally closest to the final ascribed diagnosis was chosen for each individual, just like with the clinical data. This high-dimensional imaging feature set provides a detailed snapshot of brain health and atrophy patterns, serving as a powerful complementary data source to the genetic information.

4. Methodology

A thorough, multi-stage pipeline was used as the analytical method in this study to carefully integrate high-dimensional genetic and neuroimaging data in order to build and assess reliable predictive models for the classification of AD. To guarantee data quality, account for potential confounders, and use cutting-edge methods for multi-modal data integration, this methodology combines best practices from statistical genetics, bioinformatics, and machine learning. The PLINK v1.9 software package for genetic data manipulation [65], the R programming language for statistical visualization, and Python with its extensive ecosystem of scientific computing libraries (scikit-learn, pandas, numpy) and specialized machine learning frameworks (lightgbm, xgboost) for constructing the predictive super-ensemble were all used to methodically implement the entire workflow.

4.1 Overview of Analytical Pipeline

In order to ensure a natural development from raw data to final predictive insights, the study's methodology can be thought of as a sequential workflow, with each stage's output acting as the input for the subsequent one. This can be seen in Figure 5.

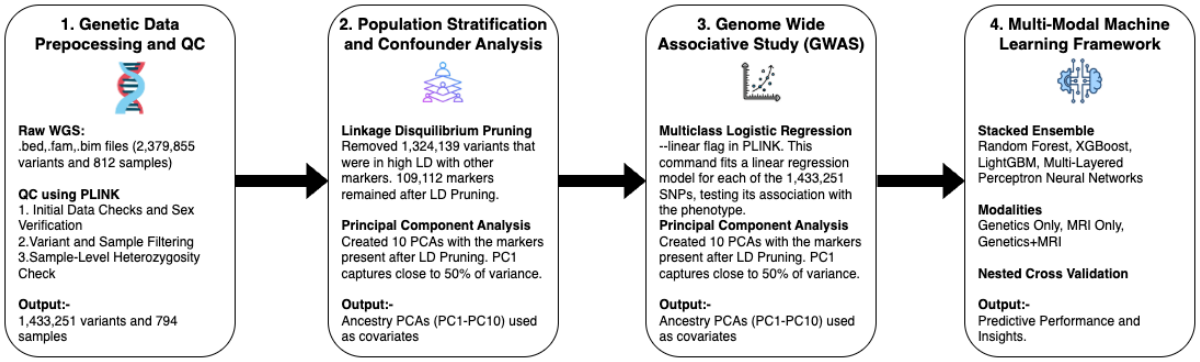

Figure 5. Pipeline for Multi-Modal AD prediction with Ensemble Learning

The following are the crucial phases:

1. Quality Control (QC) and Preprocessing of Genetic Data: In this initial phase, the raw Whole-Genome Sequencing (WGS) data is subjected to a rigorous, multi-step quality control methodology. Finding and eliminating inaccurate or low-quality data at the individual sample and genetic variation levels is the main goal. To guarantee consistency and reproducibility, this technique is automated using a BASH script.
2. Population Stratification and Confounder Analysis: Deals with the crucial problem of population substructure, which is a significant potential confounder in genetic research. Principal Component Analysis (PCA) is used after Linkage Disequilibrium (LD) trimming to obtain quantitative measures of ancestry that can be utilized as covariates to adjust for stratification.
3. Genome-Wide Association Study(GWAS): The QC'd dataset is subjected to a GWAS for two reasons. Initially, it serves as an exploratory approach to find genetic variations in our population that are strongly linked to the AD diagnostic phenotype (CN, MCI, AD). A high-confidence sample of SNPs is chosen for the machine learning models using the GWAS p-values, which also offer a biologically informed ranking of all 1.4 million variations. This rating is crucial for this study since it serves as a supervised feature selection filter.

4. Multi-Modal Machine Learning Framework: This is the core prective modelling stage. For the three-way classification challenge, an advanced stacked ensemble (super learner) architecture is created. Because of its modular nature, this framework may be used to train models on three different types of data: MRI-only, genetics-only, and an integrated Joint Genetics+MRI model. To guarantee that every stage-from feature engineering and preprocessing to hyperparameter tweaking and final evaluation-is carried out in a way that avoids data leakage and produces objective performance estimates, the entire framework is constructed within a nested cross-validation design.

With the goal of creating accurate and broadly applicable prediction models, this end-to-end pipeline offers a solid and moral strategy for addressing the difficulties of high-dimensional, multi-modal biomedical data processing.

4.2 Genetic Data Preprocessing and Quality Control (QC)

Any useful genetic analysis must have a strong quality control pipeline because it reduces the influence of genotyping errors and technical artifacts, which could otherwise produce erroneous results [66]. In order to ensure a repeatable and thoroughly documented procedure, our pipeline was constructed as an automated BASH script that methodically invoked PLINK v1.9 commands.

4.2.1 Initial Data Checks and Sex Verification

I carried out a crucial validation step prior to any data filtering, comparing each participant's recorded sex to their genetic sex. This process, which is carried out using PLINK's --check-sex command, is crucial for spotting possible mistakes in sample labelling or DNA mix-ups [67]. The technique depends on figuring out the inbreeding coefficient (F-statistic) using X chromosomal variations. Females have heterozygosity at X-linked SNPs due to having two X chromosomes, which causes F-statistic values to be close to 0. On the other hand, because males only have one X chromosome, they are genetically homozygous for all X-linked variations, which results in F-statistic values close to 1.

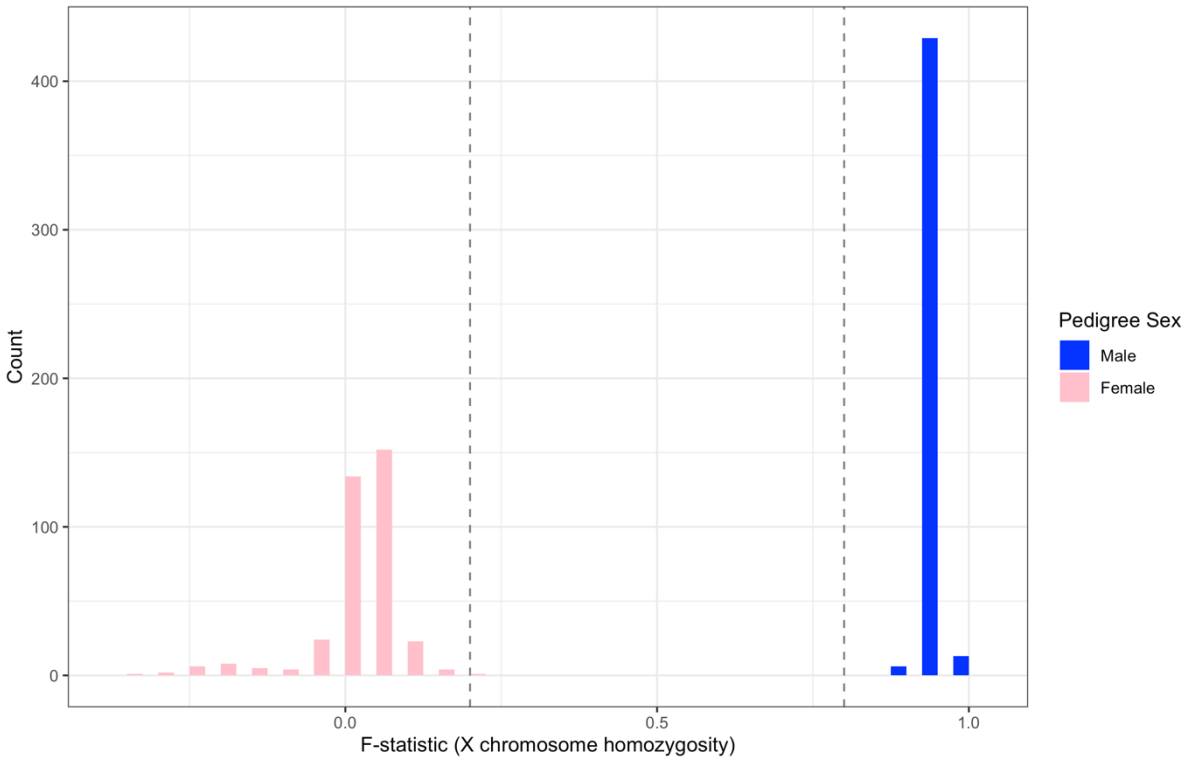

Figure 6. Sex Check based on X Chromosome Homozygosity

Figure 6 displayed the findings of this investigation. Two distinct distributions are evidently visible in the histogram: one for pedigree men (blue) centered around F=1, and another for pedigree females (pink) centered around F=0. In the script's visualization code, the dashed vertical lines at F=0.2 and F=0.8 stand for typical criteria for identifying disparities.

```
PLINK v1.9.0-b.7.8 64-bit (15 Jun 2025)
Options in effect:
  --bfile WGS_Omni25_BIN_wo_ConsentsIssues
  --check-sex
  --out initial_sex_check

Hostname: Shabbirs-MacBook-Air.local
Working directory: /Volumes/WD /WGS Analysis
Start time: Sat Sep  6 01:21:31 2025

Random number seed: 1757118091
16384 MB RAM detected; reserving 8192 MB for main workspace.
2379855 variants loaded from .bim file.
812 people (448 males, 364 females) loaded from .fam.
812 phenotype values loaded from .fam.
Using 1 thread (no multithreaded calculations invoked).
Before main variant filters, 812 founders and 0 nonfounders present.
Calculating allele frequencies... done.
Warning: 417855 het. haploid genotypes present (see initial_sex_check.hh );
many commands treat these as missing.
Warning: Nonmissing nonmale Y chromosome genotype(s) present; many commands
treat these as missing.
Total genotyping rate is 0.99744.
2379855 variants and 812 people pass filters and QC.
Phenotype data is quantitative.
--check-sex: 49975 Xchr and 0 Ychr variant(s) scanned, no problems detected.
Report written to initial_sex_check.sexcheck .
```

Figure 7. Log file for Sex Check

Perfect concordance between the recorded and genetic sex for all samples was confirmed by the PLINK log file (Figure 7), which showed that no issues were found among the 812

people. High confidence in the dataset's sample identification integrity was given by this successful check.

4.2.2 Variant and Sample Filtering Cascade

The genomic data was subjected to a series of screening procedures. Because sample-level QC can affect variant-level metrics and vice versa, the sequence in which these processes are completed is crucial.

- Autosomal Variant Selection: The --autosome flag was used to limit the study to the 22 autosomal chromosomes. Excluding the mitochondrial chromosome, which is haploid and inherited from the mother, and sex chromosomes, which differ in ploidy and inheritance patterns between males and females, is a usual practice in GWAS to streamline the study [68]. As a result, there were 2,314,174 variants instead of 2,379,855. Figure 8 displays the log file for this phase.

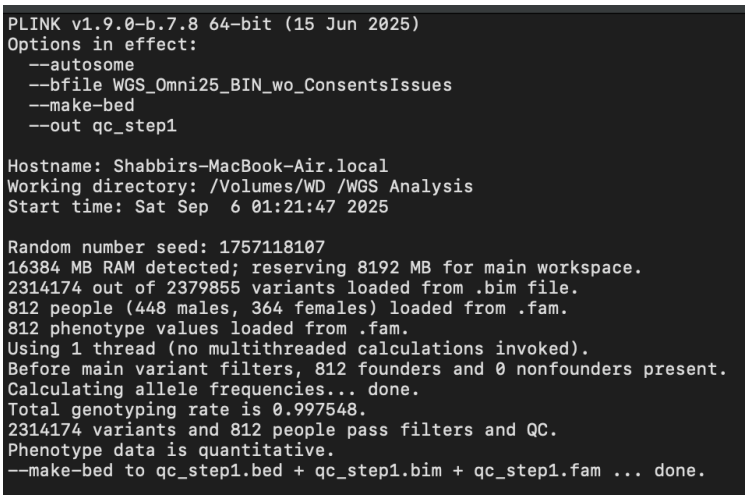

Figure 8. Log file for Autosomal Variant Selection

- Missingness Filtering: This two-step procedure eliminates data that isn't dependable because of genotyping errors.
    - Missingness of Sample (--mind 0.02): I started by eliminating anyone for whom over 2% of genotype calls were absent. A high missingness rate for a sample frequently denotes a systematic failure in the genotyping process or low-quality input DNA[69]. Zero samples were eliminated when this filter was applied to our 812 participants, demonstrating the excellent calibre of the

ADNI sample collecting and processing procedures. Figure 9 displays the log file for this phase.

```
PLINK v1.9.0-b.7.8 64-bit (15 Jun 2025)
Options in effect:
  --bfile qc_step1
  --make-bed
  --mind 0.02
  --out qc_step2

Hostname: Shabbirs-MacBook-Air.local
Working directory: /Volumes/WD /WGS Analysis
Start time: Sat Sep  6 01:22:02 2025

Random number seed: 1757118122
16384 MB RAM detected; reserving 8192 MB for main workspace.
2314174 variants loaded from .bim file.
812 people (448 males, 364 females) loaded from .fam.
812 phenotype values loaded from .fam.
0 people removed due to missing genotype data (--mind).
Using 1 thread (no multithreaded calculations invoked).
Before main variant filters, 812 founders and 0 nonfounders present.
Calculating allele frequencies... done.
Total genotyping rate is 0.997548.
2314174 variants and 812 people pass filters and QC.
Phenotype data is quantitative.
--make-bed to qc_step2.bed + qc_step2.bim + qc_step2.fam ... done.
```

Figure 9. Log file for Sample Missingness

- Missingness of Variants (--geno 0.01): I then eliminated any variants that were successfully genotyped in less than 99 percent of the samples that remained. A high variation missingness rate indicates that the SNP is hard to reliably genotype, maybe because it is located in a repetitive area of the genome or because the assay was poorly designed. In order to eliminate technically noisy characteristics, this filter is essential [70]. 109,620 variations were eliminated from the dataset in this step. This step's log file is seen in Figure 10.

```
PLINK v1.9.0-b.7.8 64-bit (15 Jun 2025)
Options in effect:
  --bfile qc_step2
  --geno 0.01
  --make-bed
  --out qc_step3

Hostname: Shabbirs-MacBook-Air.local
Working directory: /Volumes/WD /WGS Analysis
Start time: Sat Sep  6 01:22:29 2025

Random number seed: 1757118149
16384 MB RAM detected; reserving 8192 MB for main workspace.
2314174 variants loaded from .bim file.
812 people (448 males, 364 females) loaded from .fam.
812 phenotype values loaded from .fam.
Using 1 thread (no multithreaded calculations invoked).
Before main variant filters, 812 founders and 0 nonfounders present.
Calculating allele frequencies... done.
Total genotyping rate is 0.997548.
109620 variants removed due to missing genotype data (--geno).
2204554 variants and 812 people pass filters and QC.
Phenotype data is quantitative.
--make-bed to qc_step3.bed + qc_step3.bim + qc_step3.fam ... done.
```

Figure 10. Log file for Variant Missingness

- Allele Frequency Filtering (--mac 20): I used a minor allele count (MAC) filter to eliminate extremely rare variants that have little statistical power and are more likely to be genotyping errors. All SNPs for which the less common allele was detected less than 20 times across all chromosomes in the cohort were eliminated using the --mac 20 command. MAC filtering is frequently better than Minor Allele Frequency (MAF) filtering for WGS data[71]. In our sample of 794 people (1588 chromosomes), for example, a MAC of 20 is equivalent to a MAF of roughly 1.26%. A more reliable and sample-size-independent method of managing rare variations is to use MAC, which guarantees that a variant is supported by a minimum absolute number of observations [8]. With 1,433,980 variants remaining in the sample after 770,574 variants were eliminated, this filter had the most influence. This step's log file is displayed in Figure 11.

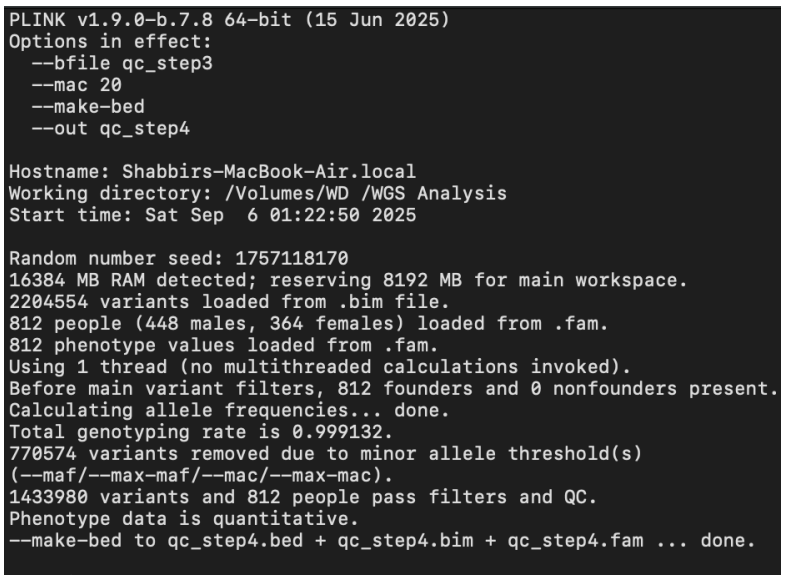

Figure 11. Log file for Allele Frequency Filtering

- Hardy-Weinberg Equilibrium(HWE) Filtering: SNPs exhibiting a considerable departure from HWE in the control population (or the full cohort, as used here) were the focus of the last variant-level filter. The predicted distribution of genotypes under random mating is described by the HWE principle. Genotyping errors, especially the systematic miscalling of heterozygous genotypes, might be indicated by a substantial deviation [72]. I used a p-value threshold of 1e−10 (--hwe 1e-10), which is rather strict. To take into consideration the enormous multiple testing burden in a WGS dataset, this stringent cutoff is required. Tens of thousands of genuine SNPs would be mistakenly eliminated with a standard p-value of 0.05. 729 variations that were

probably impacted by genotyping artifacts were eliminated in this stage. This step's log file is displayed in Figure 12.

```
PLINK v1.9.0-b.7.8 64-bit (15 Jun 2025)
Options in effect:
  --bfile qc_step4
  --hwe 1e-10
  --make-bed
  --out qc_step5

Hostname: Shabbirs-MacBook-Air.local
Working directory: /Volumes/WD /WGS Analysis
Start time: Sat Sep  6 01:23:11 2025

Random number seed: 1757118191
16384 MB RAM detected; reserving 8192 MB for main workspace.
1433980 variants loaded from .bim file.
812 people (448 males, 364 females) loaded from .fam.
812 phenotype values loaded from .fam.
Using 1 thread (no multithreaded calculations invoked).
Before main variant filters, 812 founders and 0 nonfounders present.
Calculating allele frequencies... done.
Total genotyping rate is 0.999037.
--hwe: 729 variants removed due to Hardy-Weinberg exact test.
1433251 variants and 812 people pass filters and QC.
Phenotype data is quantitative.
--make-bed to qc_step5.bed + qc_step5.bim + qc_step5.fam ... done.
```

Figure 12. Log file for Hardy-Weinberg Equilibrium (HWE) Filtering

4.2.3 Sample-Level Heterozygosity Check

I carried out a last sample-level QC step after cleaning the variation set in order to find people with unusual heterozygosity rates. The percentage of an individual's genotypes that are heterozygous is known as the heterozygosity rate, and it is determined using PLINK's --het command. The F-statistic, often known as the inbreeding coefficient, summarizes this rate. Outliers may be a sign of poor data quality:

- Excessive heterozygosity (negative F-statistic): This frequently indicates that two people's DNA may have been mingled together in the sample.
- Excessive homozygosity (positive F-statistic): This could be a sign of inbreeding in the person's lineage or, in certain situations, low-quality DNA that causes an overabundance of homozygous genotypes.

By determining the mean and standard deviation of the F-statistic for each of the 812 individuals and highlighting those that deviated more than three standard deviations from the mean, I was able to statistically identify outliers. Without depending on preconceived notions about the population, this common statistical procedure effectively detects severe outliers.

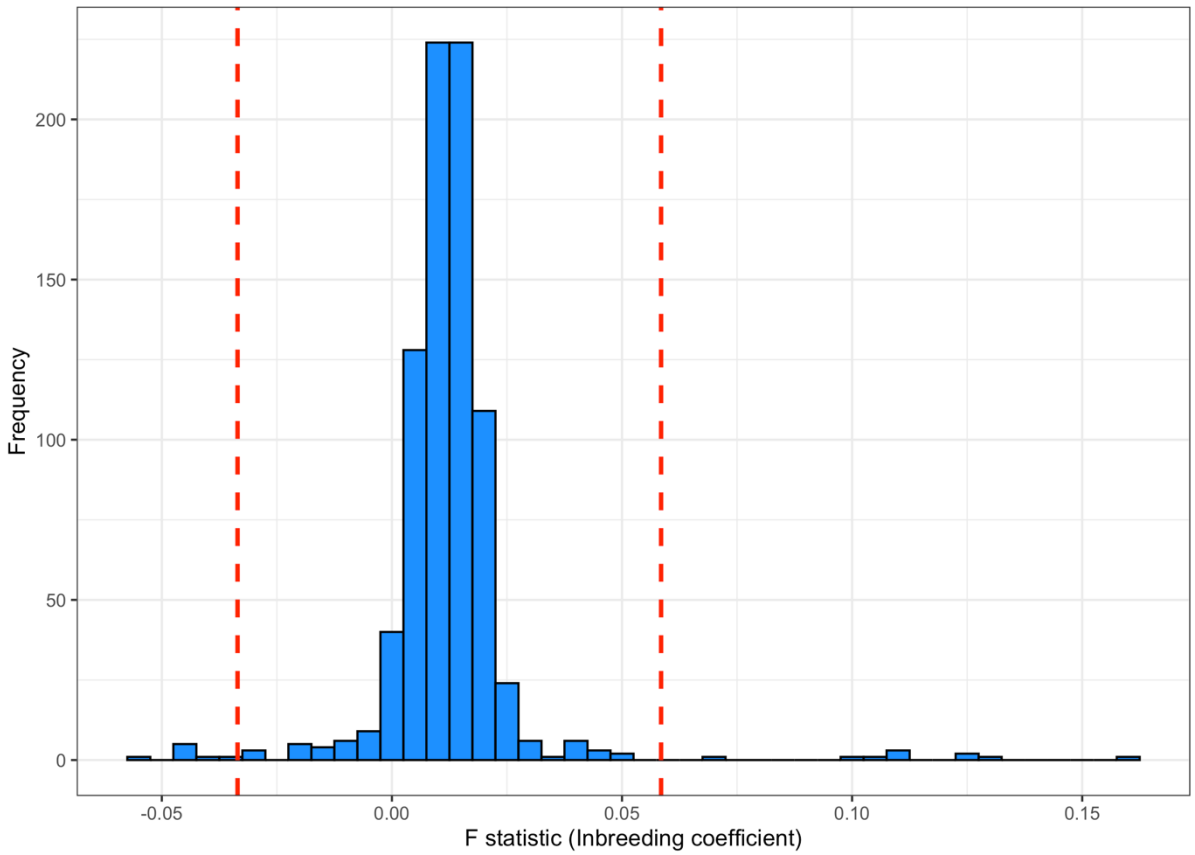

Figure 13. Sample Heterozygosity Distribution

The thresholds are indicated by red dashed lines in Figure 13, which displays an F-statistic distribution that is essentially normal. The --remove het_outliers_to_remove.txt command was used to eliminate 18 people from the dataset after they were determined to be outliers based on this criterion. As a result, the qc_final fileset contains the final, analysis-ready cohort of 794 people. This step's log file is displayed in Figure 14.

```
PLINK v1.9.0-b.7.8 64-bit (15 Jun 2025)
Options in effect:
  --bfile qc_step5
  --make-bed
  --out qc_final
  --remove het_outliers_to_remove.txt

Hostname: Shabbirs-MacBook-Air.local
Working directory: /Volumes/WD /WGS Analysis
Start time: Sat Sep  6 01:23:35 2025

Random number seed: 1757118215
16384 MB RAM detected; reserving 8192 MB for main workspace.
1433251 variants loaded from .bim file.
812 people (448 males, 364 females) loaded from .fam.
812 phenotype values loaded from .fam.
--remove: 794 people remaining.
Using 1 thread (no multithreaded calculations invoked).
Before main variant filters, 794 founders and 0 nonfounders present.
Calculating allele frequencies... done.
Total genotyping rate in remaining samples is 0.999041.
1433251 variants and 794 people pass filters and QC.
Phenotype data is quantitative.
--make-bed to qc_final.bed + qc_final.bim + qc_final.fam ... done.
```

Figure 14. Log file for Heterozygosity Outlier Removal

4.2.4 Summary of Quality Control

The raw data was refined in large part because to the QC pipeline. There were 2,379,855 variations and 812 individuals at the beginning of the process. The final dataset (qc_final) for downstream analysis included 794 people and 1,433,251 high-quality autosomal variants after being methodically filtered based on data quality and integrity parameters. This resulted in an exceptional final genotyping rate of 99.90%. This meticulous procedure guarantees that a strong basis of trustworthy genetic data will serve as the basis for any further studies.

4.3 Population Stratification Analysis

Because allele frequency differences are connected with both disease state and ancestry, genetic association studies are particularly vulnerable to confounding by population stratification, which can result in erroneous results [73]. I used PCA in a regular and efficient two-step procedure to lessen this.

4.3.1 Linkage Disequilibrium (LD) Pruning

Like many other multivariate statistical techniques, PCA makes the assumption that the input variables-in this case, SNPs-are independent. However, a phenomenon called Linkage Disequilibrium (LD) occurs when neighboring SNPs are frequently co-inherited in blocks as a result of the physical linkage of genes on chromosomes. The over-representation of certain genomic regions would emerge from doing PCA on the entire set of LD-rich SNPs, and the principle components that would be produced would represent the LD structure rather than broad, genome-wide ancestry patterns[74].

I started by using LD pruning to create a selection of roughly independent SNPs in order to address this. The PLINK command --indep-pairwise 200 50 0.1 was what I used. For dense WGS data, this command's parameters were carefully selected:

- 200: Indicates a 200 SNP window size.
- 50: Establishes the size of the steps. Each cycle shifts the window by 50 SNPs.
- 0.1: Establishes the upper limit of the greatest squared correlation (R2). Every pair of SNPs is evaluated within each window, and one SNP is eliminated if the pair's R2 is greater than 0.1. Perfect correlation is shown by an R2 of 1, whilst total independence is indicated by an R2 of 0. The SNP set that is produced is appropriate for PCA because of the strict 0.1 criterion, which successfully eliminates all but the weakest relationships.

1,324,139 variations that were in high LD with other markers were eliminated by this very successful trimming procedure. In the final set utilized for PCA, there were roughly 109,112 independent SNPs. This step's log file is displayed in Figure 15.

```
PLINK v1.9.0-b.7.8 64-bit (15 Jun 2025)
Options in effect:
  --bfile qc_final
  --indep-pairwise 200 50 0.1
  --out pca_ld_pruning

Hostname: Shabbirs-MacBook-Air.local
Working directory: /Volumes/WD /WGS Analysis
Start time: Sat Sep  6 01:23:44 2025

Random number seed: 1757118224
16384 MB RAM detected; reserving 8192 MB for main workspace.
1433251 variants loaded from .bim file.
794 people (440 males, 354 females) loaded from .fam.
794 phenotype values loaded from .fam.
Using 1 thread (no multithreaded calculations invoked).
Before main variant filters, 794 founders and 0 nonfounders present.
Calculating allele frequencies... done.
Total genotyping rate is 0.999041.
1433251 variants and 794 people pass filters and QC.
Phenotype data is quantitative.
Pruned 102176 variants from chromosome 1, leaving 8712.
Pruned 110567 variants from chromosome 2, leaving 8408.
Pruned 93741 variants from chromosome 3, leaving 7282.
Pruned 88043 variants from chromosome 4, leaving 6781.
Pruned 82935 variants from chromosome 5, leaving 6515.
Pruned 90778 variants from chromosome 6, leaving 6330.
Pruned 73794 variants from chromosome 7, leaving 5945.
Pruned 72771 variants from chromosome 8, leaving 5384.
Pruned 60235 variants from chromosome 9, leaving 5016.
Pruned 68298 variants from chromosome 10, leaving 5581.
Pruned 65144 variants from chromosome 11, leaving 5296.
Pruned 64668 variants from chromosome 12, leaving 5286.
Pruned 48720 variants from chromosome 13, leaving 4005.
Pruned 44133 variants from chromosome 14, leaving 3637.
Pruned 41076 variants from chromosome 15, leaving 3652.
Pruned 43735 variants from chromosome 16, leaving 4009.
Pruned 36517 variants from chromosome 17, leaving 3764.
Pruned 40396 variants from chromosome 18, leaving 3670.
Pruned 25626 variants from chromosome 19, leaving 3003.
Pruned 32260 variants from chromosome 20, leaving 3142.
Pruned 18762 variants from chromosome 21, leaving 1802.
Pruned 19764 variants from chromosome 22, leaving 1892.
Pruning complete.  1324139 of 1433251 variants removed.
Marker lists written to pca_ld_pruning.prune.in and pca_ld_pruning.prune.out .
```

Figure 15. Log file for Linkage Disequilibrium Pruning

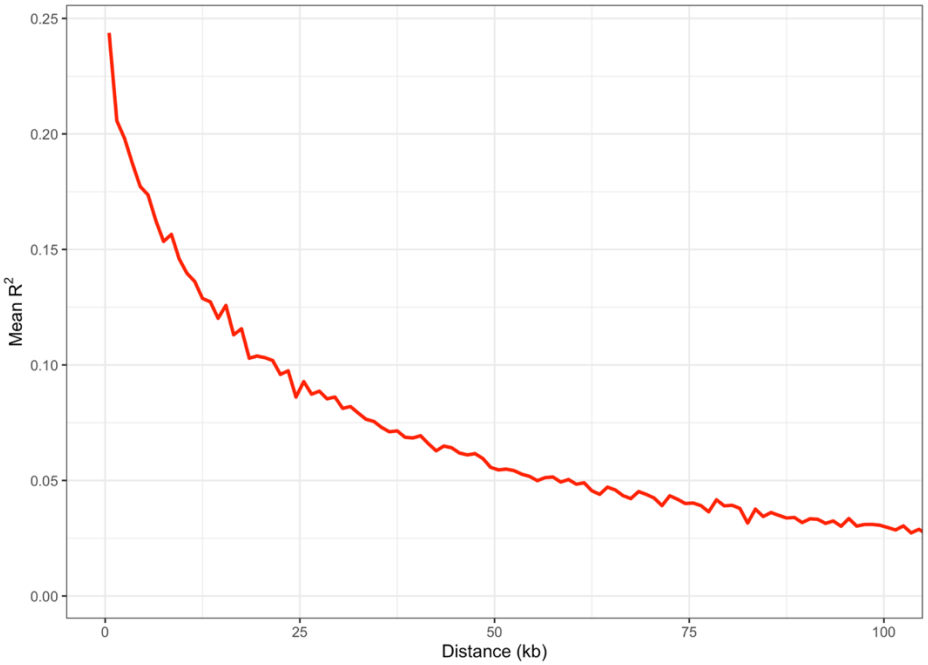

Figure 16. Linkage Disequilibrium Decay

Figure 16 showed the typical decline of LD as the physical distance between SNPs increased. Plotting the impacts of historical recombination in the population, as anticipated, reveals a sharp drop in mean R2 within the first 50–100 kilobases (kb), leveling out at a baseline level for more distant SNPs.

4.3.2 Principal Component Analysis (PCA)

Using the pruned set of 109,112 SNPs, I performed PCA to distil the major axes of genetic variation in our cohort of 794 individuals. PCA is a dimensionality reduction technique that transforms the high-dimensional genotype data into a set of orthogonal variables called principal components (PCs). In human genetics, the top PCs have been shown to correspond closely with geographical ancestry [9]. I calculated the top **10 PCs** using the --pca 10 command. The log file for this step is shown in Figure 17.

```
PLINK v1.9.0-b.7.8 64-bit (15 Jun 2025)
Options in effect:
  --bfile qc_final
  --extract pca_ld_pruning.prune.in
  --out ancestry_pca
  --pca 10

Hostname: Shabbirs-MacBook-Air.local
Working directory: /Volumes/WD /WGS Analysis
Start time: Sat Sep  6 01:23:48 2025

Random number seed: 1757118228
16384 MB RAM detected; reserving 8192 MB for main workspace.
1433251 variants loaded from .bim file.
794 people (440 males, 354 females) loaded from .fam.
794 phenotype values loaded from .fam.
--extract: 109112 variants remaining.
Using up to 9 threads (change this with --threads).
Before main variant filters, 794 founders and 0 nonfounders present.
Calculating allele frequencies... done.
Total genotyping rate is 0.99897.
109112 variants and 794 people pass filters and QC.
Phenotype data is quantitative.
Relationship matrix calculation complete.
--pca: Results saved to ancestry_pca.eigenval and ancestry_pca.eigenvec .
```

Figure 17. Log file for Principal Component Analysis

Two plots were used to evaluate the PCA results:

- PCA Scree Plot (Figure 18): The percentage of the overall genetic variance that each PC accounts for is shown in this bar chart. According to our plot, PC1 (the first PC) explains more than 50% of the variance, while PC2 explains almost 10%. The "elbow"-shaped variation explained by succeeding PCs quickly decreases. This suggests that the first few components contain the most important population substructure.

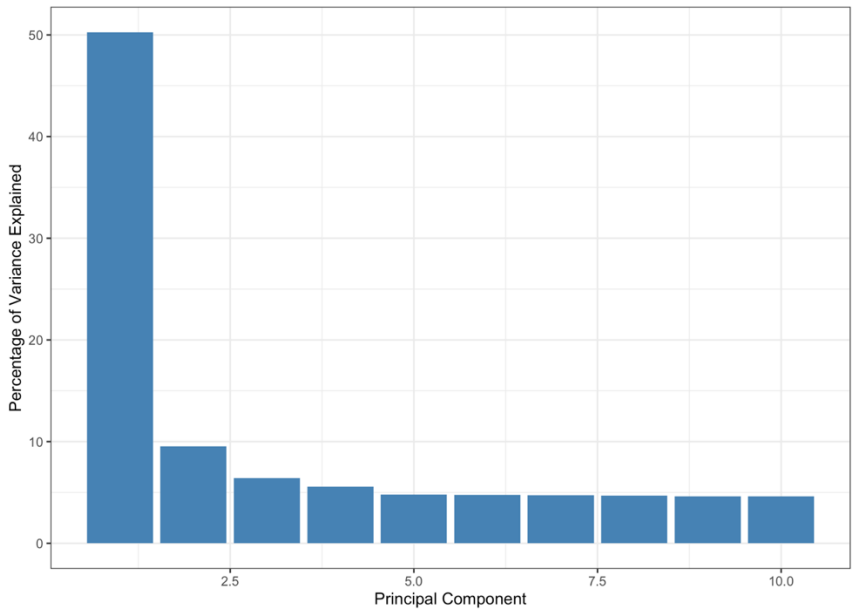

Figure 18. PCA Scree Plot

- Genetic Ancestry PCA Plot (Figure 19): The genetic ties between individuals are graphically shown by this scatter plot of PC1 vs PC2. Along with smaller clusters and scattered individuals that might represent different ancestries or admixed individuals, the map displays a big, compact cluster that most likely represents people of European ancestry, the main group in ADNI. This graphic demonstrates the need to account for genetic variation in the cohort and validates its existence. To account for these ancestral differences, the resulting GWAS included the computed top 10 PCs, stored in ancestry_pca.eigenvec, as covariates.

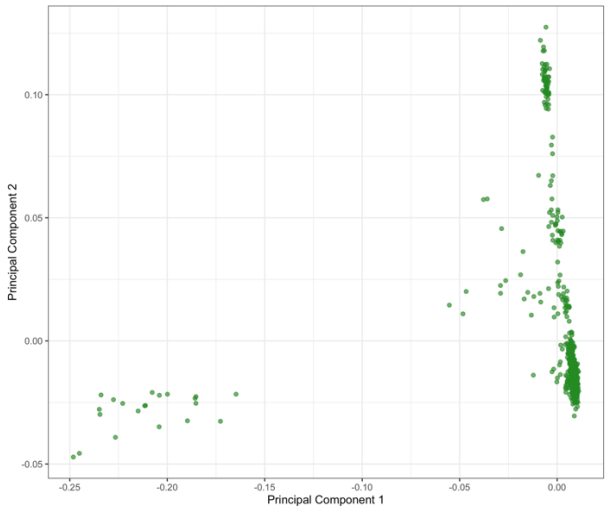

Figure 19. Genetic Ancestry PCA

4.4 Genome-Wide Association Study (GWAS)

While the project's ultimate goal is to use machine learning to classify diseases, performing a GWAS was an essential first step. In order to find SNPs associated with the AD phenotype, we conducted a large-scale association screen using the GWAS. More significantly, we were able to produce a data-driven ranking of all variations based on their statistical significance. Our main feature selection method for the machine learning models is based on this ranking. I used PLINK's --linear flag to do a quantitative association analysis on the final qc_final dataset. This command tests the relationship between each of the 1,433,251 SNPs and the phenotype by fitting a linear regression model for each of them. In order to construct a dose-response relationship in which risk alleles are projected to shift individuals towards a higher (more severe) phenotypic score, GWAS frequently treats the numerical coding of our phenotype (CN=1, MCI=2, AD=3) as quantitative, even though it is strictly an ordinal scale. Importantly, I added the top 10 PCs from our ancestry analysis as covariates in the regression model (--covar ancestry_pca.eigenvec --covar-number 1-10), to avoid population stratification from skewing the results.
After adjusting for ancestry, the significance of the link for that SNP is shown by the p-value for the beta_SNP coefficient. This step's log file is displayed in Figure 20.

```
PLINK v1.9.0-b.7.8 64-bit (15 Jun 2025)
Options in effect:
  --bfile qc_final
  --covar ancestry_pca.eigenvec
  --covar-number 1-10
  --linear
  --out gwas_results

Hostname: Shabbirs-MacBook-Air.local
Working directory: /Volumes/WD /WGS Analysis
Start time: Sat Sep  6 01:24:14 2025

Random number seed: 1757118254
16384 MB RAM detected; reserving 8192 MB for main workspace.
1433251 variants loaded from .bim file.
794 people (440 males, 354 females) loaded from .fam.
794 phenotype values loaded from .fam.
Using 1 thread (no multithreaded calculations invoked).
--covar: 10 covariates loaded.
Before main variant filters, 794 founders and 0 nonfounders present.
Calculating allele frequencies... done.
Total genotyping rate is 0.999041.
1433251 variants and 794 people pass filters and QC.
Phenotype data is quantitative.
Writing linear model association results to gwas_results.assoc.linear ... done.
```

Figure 20. Log file for Genome Wide Association Studies

3.4.1 GWAS Result Visualization and Interpretation

To evaluate the overall quality of the GWAS results and find important loci, they were visualized after being written to gwas_results.assoc.linear.

- Manhattan Plot (Figure 21): An overview of the association findings throughout the genome is given by this figure. Stronger statistical significance is indicated by greater peaks on the y-axis, which reflects the −log10. Two typical significance thresholds are

shown in the plot: a "genome-wide significance" line at $p=5\times10^{-8}$ (red) and a "suggestive" line at $p=1\times10^{-5}$ (blue). Because it accounts for the roughly one million independent tests conducted throughout the genome, the genome-wide threshold is frequently used in research on populations with European ancestry [75]. On chromosome 19, our Manhattan plot displays a distinct and remarkable signal, with several SNPs surpassing the genome-wide significance threshold. Given that the Apolipoprotein E (APOE) gene, which is without a doubt the most important genetic risk factor for late-onset AD, is located at this locus, our analysis is strongly validated [76]. This strong signal demonstrates that the most significant known genetic connection with the disease was successfully captured by our GWAS.

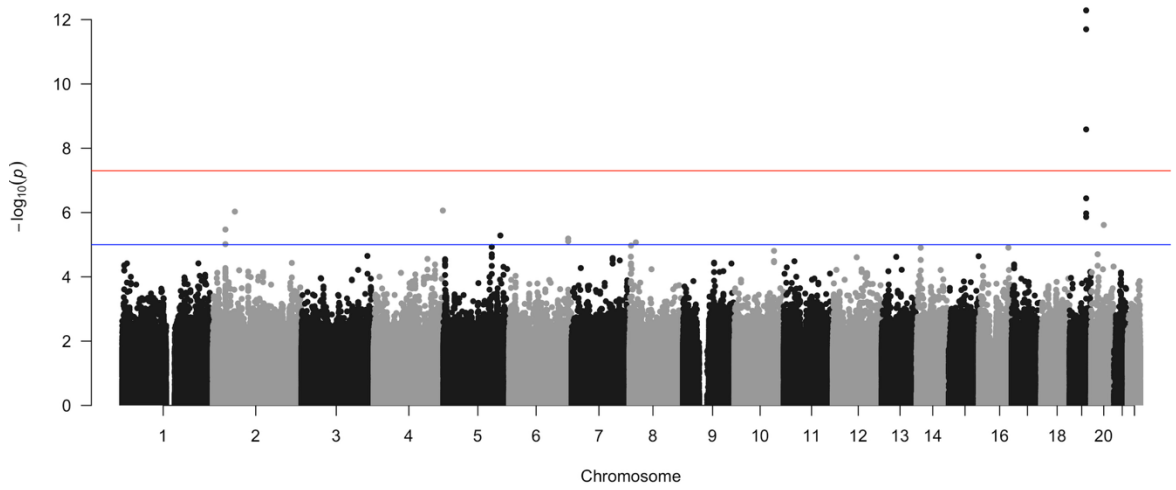

Figure 21. Manhattan Plot

- Q-Q Plot (Figure 22): A diagnostic tool for comparing the distribution of the observed p-values to the expected uniform distribution under the null hypothesis of no correlation is the quantile-quantile (Q-Q) plot. Since most SNPs are not anticipated to be linked to the trait in a well-controlled GWAS, their p-values ought to be in line with the null distribution, or the diagonal line. The great majority of the observed p-values, as demonstrated by our Q-Q plot, perfectly match the expected diagonal. This proves that our use of PCA variables effectively adjusted for population stratification by showing that test statistics do not exhibit systematic inflation. As would be expected for a study with real genetic signals, the collection of actually related SNPs (mostly the APOE locus) is represented by the sharp divergence from the diagonal near the tail of the distribution. A well-conducted analysis would be indicated by the genomic inflation factor (lambda_GC), which measures this divergence, being extremely close to 1.0.

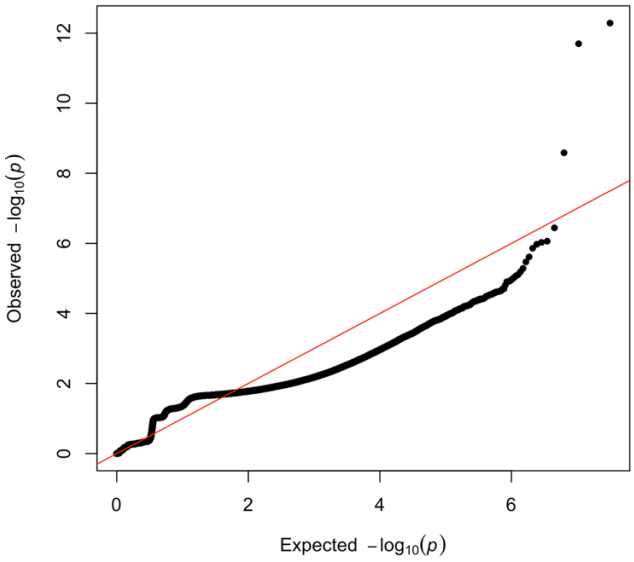

Figure 22. Q-Q Plot of GWAS p-values

Exploratory Data Analysis was done on the samples present after complete QC before using the data for Machine Learning. The diagrams generated are in the Appendix.

4.5 Machine Learning Methodology

I created a complex super-ensemble machine learning system to carry out the three-class (CN, MCI, and AD) diagnostic classification, building on the meticulously processed data and GWAS insights. The complete Python-based framework was created with three separate modeling experiments in mind: genetics-only, MRI-only, and an integrated Joint Genetics+MRI model. These studies focused on robustness, limiting data leakage, and utilizing the complimentary knowledge from genetics and neuroimaging.

4.5.1 Overview of Super-Ensemble (Stacked) Framework

A stacked ensemble, or "super learner," architecture is the foundation of our predictive modeling [77]. In order to reduce model variance and capture more intricate patterns than any one model could on its own, this two-level technique combines the outputs of many base learning algorithms to improve prediction performance.

1. Base Learners (Level 0): A wide variety of distinct machine learning models (such as Random Forest, Gradient Boosting, and Neural Networks) make up this level. The core feature data-such as SNPs or MRI measurements-is used to train these algorithms. Through a cross-validation process, their forecasts are produced in an out-of-fold (OOF) fashion rather than using their direct predictions for the final output.

2. Meta-Learner (Level 1): There is just one distinct model at this level. The "meta-feature" matrix is a new feature set created by concatenating the OOF forecasts from each Level 0 model. The best combination of the base learners' predictions is then taught to the meta-learner using this matrix.

A five-tiered cross-validation loop encircles the entire system. Accordingly, the dataset is divided into five folds, and the basic learners are trained and OOF predictions are produced five times, using each fold as the validation set once. By doing this, the meta-learner is guaranteed to be trained on predictions for the complete dataset, each of which was produced by models that were not exposed to that specific data point during training. This methodical procedure is essential for getting objective performance evaluations of the group.

4.5.2 Data Preparation for Machine Learning

The QC and GWAS algorithms' outputs were converted into machine learning-ready forms.

- Genetic Data: The enormous dimensionality of genetic data is its main drawback. I addressed this by applying a supervised feature selection filter to the GWAS results. After sorting the gwas_results.assoc.linear file by p-value, the top 50,000 distinct SNPs were chosen. This figure strikes a practical compromise between managing the computing task and preserving a significant amount of the possible polygenic signal. A more concentrated, smaller genetic dataset (ml_top50k_data) was extracted using this list of the most important SNPs. The PLINK.raw format, a text file with individuals in rows and SNPs in columns, was then used to export this dataset. Genotypes were classified as the number of minor alleles (0, 1, or 2). For example, if G is the minor allele, then AA is the genotype that has no copies of the minor allele. 1 for the genotype (e.g., AG) that has one copy of the minor allele. 2 for the genotype (e.g., GG) that has two copies of the minor allele.
- MRI Data: mri.csv was used to load the tabular MRI feature data. The Python script makes sure that the collection of participants is in line with those in the final genetic dataset and that just the most recent scan for each participant is used.
- Joint Data: Participants with both genetic and MRI data were found for the joint modality. The models were thus able to learn from both forms of data at the same time by concatenating their respective feature vectors horizontally to produce a single, broad feature matrix.

4.5.3 Leak-Proof Feature Engineering and Dimensionality Reduction

"Data leakage," in which information from the test or validation set unintentionally affects the training process and results in artificially inflated performance metrics, is a typical machine learning pitfalls. Our whole pretreatment pipeline was created to be used individually in each fold of the cross-validation loop in order to avoid this. Both the training data and the held-out validation data were transformed using a pipeline of transformations that was fitted only to the training data of a particular fold.
This per-fold pipeline was made up of a series of actions intended to produce several "views" of the data:

1. Imputation: Any remaining missing values were handled using scikit-learn's SimpleImputer. Depending on what came next, different approaches were used, such as mean imputation for techniques like ANOVA and SVD and ‘most frequent' (mode) imputation for techniques like Chi-squared that perform better with discrete-like data.
2. Scaling: For many algorithms, feature scaling is crucial. For the majority of views, StandardScaler-which standardizes features by eliminating the mean and scaling to unit variance-was employed. As required by the Chi-squared statistic, features were transformed to a non-negative [0, 1] range for the Chi-squared view using MinMaxScaler.
3. Feature Selection (FS): Four distinct supervised feature selection methods were used in parallel to produce a variety of inputs for the ensemble, each of which produced a distinct "view" of the data. One adjustable hyperparameter was the number of features (K_BEST_SNP, K_BEST_MRI, etc.) to retain. The four techniques were:
    - ANOVA F-test: A univariate filter that chooses features with the most separated means among the three diagnostic groups by using the highest F-statistic from an Analysis of Variance test.
    - Mutual Information (MI): A non-parametric technique that gauges how each attribute and the target variable are related. It complements the linear ANOVA test by capturing intricate, non-linear interactions.
    - L1-Regularization (Lasso): Involves training a logistic regression model with an L1 penalty. Feature selection is carried out automatically by the L1 penalty, which makes the coefficients of less significant characteristics exactly zero. The characteristics with the biggest non-zero coefficients were the ones I chose.

- Chi-squared Test: An additional univariate filter appropriate for categorical-like, non-negative characteristics. It checks to see if a feature and the class labels are dependent.

4. Dimensionality Reduction with SVD: Truncated Singular Value Decomposition (SVD) was used to project the resultant (still high-dimensional) feature set onto a lower-dimensional space following feature selection. One effective matrix factorization method for reducing dimensionality is SVD. I produced dense, information-rich "meta-features" that represent the primary axes of variation in the chosen feature space by retaining the top n components (SVD_A = 300, for example). This stage aids in regularizing the models and de-noising the data. I produced a rich and varied set of inputs for the base learners by developing four different views (ANOVA+SVD, MI+SVD, L1+SVD, and Chi2+SVD) for each modality. This is a crucial tactic for effective ensembling.

4.5.4 Base Learners and Hyperparameter Optimization

Four strong and unique classification algorithms made up our ensemble's Level 0 and were selected to offer a variety of learning strategies:

- Random Forest (RF): Based on bagging (bootstrap aggregating), this ensemble model is its own. It averages the predictions of numerous decision trees that have been trained on various random subsets of the data and features. RF is renowned for its low variance, robustness, and proficiency with high-dimensional data.
- XGBoost (XGB): A very effective and optimized gradient boosting algorithm. Models are constructed in a sequential fashion, with each new model fixing the mistakes of the one before it. It is often a top performer in machine learning competitions and has built-in regularization to avoid overfitting.
- LightGBM (LGBM): A relatively recent gradient boosting framework that is comparable to XGBoost but frequently much faster. It achieves this speed without compromising accuracy by using gradient-based one-side sampling (GOSS) and a revolutionary leaf-wise tree development technique.

  • Multi-Layer Perceptron (MLP): A traditional feedforward artificial neural network. Through a technique known as backpropagation, its many layers of nodes (neurons) enable it to learn extremely intricate, non-linear correlations between inputs and outputs.

I used RandomizedSearchCV for hyperparameter tweaking in order to optimize each base learner's performance. This approach takes a given number of parameter combinations (TUNING_N_ITER = 20) and samples them from predefined distributions (e.g., integer ranges for tree depths, log-uniform for learning rates). In many cases, it is more effective than a thorough Grid Search, particularly when dealing with vast parameter spaces. In order to strictly prevent the leaking of validation data knowledge into the tuning process, this search was conducted on the training data of each major fold within an inner 3-fold cross-validation loop. A reliable statistic for multi-class classification, the one-vs-rest macro-averaged ROC-AUC score, was used to optimize the search.

4.5.5 Stacked Generalization and Final Prediction

To generate the final projections, the stacking procedure was meticulously planned.

1. OOF Prediction Generation: The tuned base learners were trained on the training section and utilized to forecast probabilities for the samples in the validation portion of each of the five primary cross-validation folds. These OOF forecasts were gathered.
2. Meta-Feature Matrix Construction: The OOF predictions were compiled into a meta-feature matrix, Z_oof, following the completion of all five folds. Concatenation of the predictions from the base models was done for each of the four feature selection "views," For instance, the meta-matrix for a single view of the Genetics-only modality would have (794 samples) x (4 models * 3 classes) = (794 x 12) dimensions if I utilized four base learners. Additionally, each base model's output's prediction entropy-a measure of the model's uncertainty-is computed by the script and appended as an extra meta-feature.
3. Meta-Learner Training: Because LightGBM classifiers are fast and perform well on tabular data, they were selected as the Level 1 meta-learner. Using a 5-fold CV and a randomized search across the full OOF meta-matrix, its own hyperparameters were adjusted. The final model was then produced by training the optimal meta-learner on the entire Z_oof matrix.

The meta-learner is trained on objective inputs thanks to this methodical procedure, which teaches it how to effectively balance and integrate the many "opinions" of the base learners to provide a better final classification.

4.5.6 Model Evaluation

The OOF predictions produced by the final tuned meta-learner were used to assess the overall super-ensemble's final performance. This offers an objective assessment of the model's performance using fresh, untested data. I employed a set of metrics suitable for classifying many classes, some of which may be unbalanced:

- Macro-averaged ROC-AUC (One-vs-Rest): This measure evaluates how well the model distinguishes between each class and the others. Each class is given equal weight regardless of its frequency when the "macro" average computes the metric separately for each class and then takes the average.
- Balanced Accuracy : The average recall (sensitivity) for each class. Because it penalizes models that only outperform minority classes on the majority class, it provides a robust metric for imbalanced datasets.
- Macro F1-Score: The F1-score is calculated by taking the harmonic mean of recall and precision. Once more offering a fair assessment of performance across all classes, macro-averaging calculates the F1-score for each class and then averages them.

Along with these quantitative metrics, I also produced qualitative diagnostics, such as per-class ROC and Precision-Recall curves to give a more detailed understanding of the model's performance for each diagnostic category and confusion matrices to show the specific error patterns (e.g., misclassifying MCI as CN vs. AD).

5. Results and Evaluation

5.1 Overview

The full grid of experiments spanned three modalities-genetic only, MRI only, and joint genetic+MRI-crossed with four feature-selection families (ANOVA, Chi-square, L1-penalised logistic, and mutual information) and increasing levels of model ensembling, from single learners to a four-model stack. Across this landscape, performance was consistently strongest when information from both modalities was fused and when probability outputs from multiple, diverse base learners were blended by the meta-learner. Aggregating results across runs highlights this pattern. Joint models achieved the highest average macro-AUC and macro-F1, with genetic-only models close behind and MRI-only models markedly weaker. On average, the joint configuration yielded macro-AUC =0.945 and macro-F1 =0.838; genetic-only averaged macro-AUC =0.936 and macro-F1 =0.832; MRI-only averaged macro-AUC =0.695 and macro-F1 =0.496. These averages set the backdrop for the

best overall configuration, which combined both views and a maximal set of base learners. The Experiment Results for Genetics, MRI and Genetics+MRI is shown in Figure 23, 24 and 25 respectively.

Feature Selection	Base Models	ROC AUC Macro Overall	Balanced Accuracy	F1 Macro	AP Macro Overall
ANOVA	LGBM	97.28%	90.45%	90.64%	94.80%
	MLP	88.74%	75.17%	75.64%	79.90%
	RF	95.56%	87.35%	87.51%	91.07%
	XGB	96.94%	89.49%	89.52%	93.85%
	RF+XGB	97.18%	91.25%	91.36%	94.28%
	RF+XGB+LGBM	97.39%	91.28%	91.37%	95.02%
	RF+XGB+LGBM+MLP	97.60%	90.03%	90.12%	95.59%
CHI2	LGBM	95.96%	88.03%	88.25%	92.59%
	MLP	86.80%	70.60%	71.02%	76.15%
	RF	94.08%	84.26%	84.45%	88.02%
	XGB	96.84%	87.38%	87.75%	94.36%
	RF+XGB	96.28%	89.25%	89.49%	92.74%
	RF+XGB+LGBM	96.26%	88.47%	88.77%	92.57%
	RF+XGB+LGBM+MLP	96.92%	89.19%	89.39%	93.66%
L1	LGBM	94.33%	84.65%	84.76%	88.78%
	MLP	85.08%	67.04%	67.63%	73.39%
	RF	93.74%	81.88%	81.80%	88.43%
	XGB	95.40%	85.60%	85.61%	91.39%
	RF+XGB	95.25%	83.05%	83.12%	91.47%
	RF+XGB+LGBM	95.00%	83.72%	83.94%	91.10%
	RF+XGB+LGBM+MLP	95.40%	84.47%	84.46%	91.45%
MI	LGBM	97.28%	90.45%	90.64%	94.80%
	MLP	88.74%	75.17%	75.64%	79.90%
	RF	95.56%	87.35%	87.51%	91.07%
	XGB	96.94%	89.49%	89.52%	93.85%
	RF+XGB	97.18%	91.25%	91.36%	94.28%
	RF+XGB+LGBM	97.39%	91.28%	91.37%	95.02%
	RF+XGB+LGBM+MLP	97.60%	90.03%	90.12%	95.59%

Table 1. Genetic Experiment Results

Feature Selection	Base Models	ROC AUC Macro Overall	Balanced Accuracy	F1 Macro	AP Macro Overall
ANOVA	LGBM	65.76%	45.65%	45.34%	48.27%
	MLP	68.30%	50.50%	50.41%	49.94%
	RF	68.86%	48.57%	48.43%	51.37%
	XGB	68.82%	48.34%	48.48%	51.08%
	RF+XGB	69.54%	50.65%	50.67%	51.35%
	RF+XGB+LGBM	68.81%	49.26%	49.10%	51.32%
	RF+XGB+LGBM+MLP	68.52%	45.99%	45.36%	51.64%
CHI2	LGBM	70.36%	53.03%	52.96%	50.37%
	MLP	69.33%	50.86%	50.70%	52.19%
	RF	69.51%	49.03%	49.12%	52.85%
	XGB	71.27%	51.89%	51.79%	54.08%
	RF+XGB	71.05%	49.62%	49.74%	54.68%
	RF+XGB+LGBM	70.99%	49.71%	49.89%	54.51%
	RF+XGB+LGBM+MLP	71.86%	52.14%	52.40%	54.78%
L1	LGBM	67.06%	47.93%	48.03%	49.06%
	MLP	70.90%	50.59%	50.47%	52.45%
	RF	67.29%	47.08%	46.69%	50.46%
	XGB	71.67%	53.05%	53.12%	54.23%
	RF+XGB	70.41%	50.58%	50.44%	53.24%
	RF+XGB+LGBM	68.64%	49.02%	48.51%	51.04%
	RF+XGB+LGBM+MLP	70.91%	50.31%	50.12%	53.34%
MI	LGBM	68.91%	48.23%	48.36%	50.65%
	MLP	69.85%	50.71%	51.21%	52.60%
	RF	64.98%	46.18%	46.15%	45.88%
	XGB	71.78%	51.63%	52.04%	55.69%
	RF+XGB	70.24%	50.98%	51.17%	53.56%
	RF+XGB+LGBM	69.64%	48.79%	48.85%	52.55%
	RF+XGB+LGBM+MLP	71.01%	50.11%	50.45%	54.17%

Table 2. MRI Experiment Results

Feature Selection	Base Models	ROC AUC Macro Overall	Balanced Accuracy	F1 Macro	AP Macro Overall
ANOVA	LGBM	95.14%	85.74%	85.93%	90.44%
	MLP	91.27%	76.15%	75.95%	84.21%
	RF	95.76%	86.88%	86.87%	92.53%
	XGB	95.77%	86.11%	86.21%	92.19%
	RF+XGB	96.46%	87.88%	87.95%	93.55%
	RF+XGB+LGBM	96.49%	86.95%	87.03%	93.74%
	RF+XGB+LGBM+MLP	97.64%	89.86%	89.95%	95.66%
CHI2	LGBM	96.35%	89.04%	89.30%	92.37%
	MLP	91.17%	77.57%	77.76%	84.32%
	RF	95.17%	85.93%	86.12%	90.55%
	XGB	97.18%	88.09%	88.34%	94.78%
	RF+XGB	96.50%	88.23%	88.35%	93.32%
	RF+XGB+LGBM	96.53%	88.97%	89.10%	92.97%
	RF+XGB+LGBM+MLP	97.85%	90.43%	90.50%	96.19%
L1	LGBM	94.88%	84.34%	84.64%	89.93%
	MLP	82.26%	66.43%	65.77%	68.30%
	RF	93.24%	78.72%	78.93%	86.79%
	XGB	96.08%	86.53%	86.62%	92.37%
	RF+XGB	95.09%	85.04%	85.10%	90.01%
	RF+XGB+LGBM	94.73%	82.98%	83.08%	88.96%
	RF+XGB+LGBM+MLP	95.21%	85.72%	85.71%	90.39%
MI	LGBM	94.40%	80.28%	81.17%	89.46%
	MLP	83.25%	64.99%	64.89%	72.05%
	RF	94.24%	80.95%	81.28%	89.79%
	XGB	95.62%	84.37%	84.44%	91.71%
	RF+XGB	96.02%	84.88%	84.85%	92.77%
	RF+XGB+LGBM	95.49%	83.83%	84.10%	91.86%
	RF+XGB+LGBM+MLP	96.28%	86.43%	86.50%	93.21%

Table 3. Genetics+MRI Experiment Results

5.2 Impact of Feature Selection Methods

Feature selection shaped performance differently by modality. For genetic inputs, ANOVA and Chi-square were the most effective families on average, with ANOVA slightly ahead on macro-AUC and macro-F1, followed by Chi-square; L1 and mutual information trailed. For joint inputs, Chi-square tended to deliver the strongest averages, followed closely by ANOVA, with L1 and mutual information again behind. For MRI-only inputs, even the best selector (Chi-square) produced only modest gains; scores remained far below those of genetic or joint systems, suggesting that the single-visit MRI features available here carried limited class-separating signal compared with the wide genetic panel or the combination of both.

5.3 Impact of Base Learners Diversity

The number and diversity of base learners also mattered. Moving from a single model to two or more generally improved macro-AUC and macro-F1 for genetic and joint modalities. The four-model stack-Random Forest, XGBoost, LightGBM and MLP-was consistently among the strongest variants, especially for the joint view where the gains from ensembling were largest on average. The added diversity appeared to reduce variance and broaden the operating characteristics of the meta-learner, yielding more robust probability estimates across folds. In contrast, with MRI-only inputs, ensembling produced only marginal benefits, consistent with the view that the limiting factor in that setting was signal content rather than model capacity.

5.4. Best Performing Model

The top-ranked experiment overall was the **joint** configuration with Chi-square feature selection and the RF+XGB+LGBM+MLP base-learner set. Out-of-fold (OOF) evaluation for this model produced a macro-AUC (OvR) of 0.978**,** macro-F1 of 0.905, and balanced accuracy of 0.904, with an overall accuracy =0.903 across 793 subjects. The class wise precision, recall and F1 score from the OOF predictions as shown in Figure 26.

Diagosis Label	Precision	Recall	F1 Score	Support
Congitively Normal (CN)	90.60%	91.40%	91.00%	232
Mild Cognitive Impairment (MCI)	86.40%	88.40%	87.40%	302
Alzheimers Dieases (AD)	94.80%	91.50%	93.10%	259

Table 4. Class Wise Precision, Recall and F1 Scores of the best performing model.

True Label	**CN**	212	20	0
	MCI	22	267	13
	AD	0	22	237
		CN	**MCI**	**AD**
		Predicted label		

Figure 23. Confusion Matrix of best performing model

These summary statistics align with the confusion matrix pattern (Figure 26): of the 232 CN cases, 212 were correctly identified, with 20 mislabelled as MCI and none as AD; of the 302 MCI cases, 267 were correctly identified, with 22 called CN and 13 called AD; of the 259 AD cases, 237 were correctly identified, with 22 called MCI and none called CN. The OOF estimates suggest high discrimination and balanced performance across classes, with the expected difficulty concentrated around MCI as an intermediate phenotype.

The error profile of the best model is clinically intuitive. There was no direct confusion between CN and AD: CN→AD and AD→CN errors were both zero. CN errors mostly flowed to MCI (8.62%), and AD errors mostly flowed to MCI (8.49%). Within MCI,

misclassifications were split toward CN (7.28%) and AD (4.30%). The model rarely leaps across the diagnostic spectrum; it errs by moving one step toward the neighbouring class. This pattern is expected in a progressive disease context and underpins the strong macro-F1 despite class overlap in the feature space. These figures are derived from the OOF confusion matrix and associated report.

The precision–recall(Figure 25) and ROC(Figure 24) analyses of the best model support the strong separation with a small performance penalty for MCI compared to CN and AD. The one-vs-rest ROC curves show class-wise AUCs of 0.987 for CN, 0.961 for MCI, and 0.987 for AD, with a micro-AUC of 0.980. The corresponding precision–recall curves yield average precision (AP) of 0.974 for CN, 0.934 for MCI, and 0.978 for AD, with a micro-AP of 0.963. The lift over the diagonal is most pronounced for CN and AD, and remains substantial for MCI, reflecting the genuine separability of the two classes and the residual ambiguity of borderline cases. These values align with the tabulated summary in the results file.

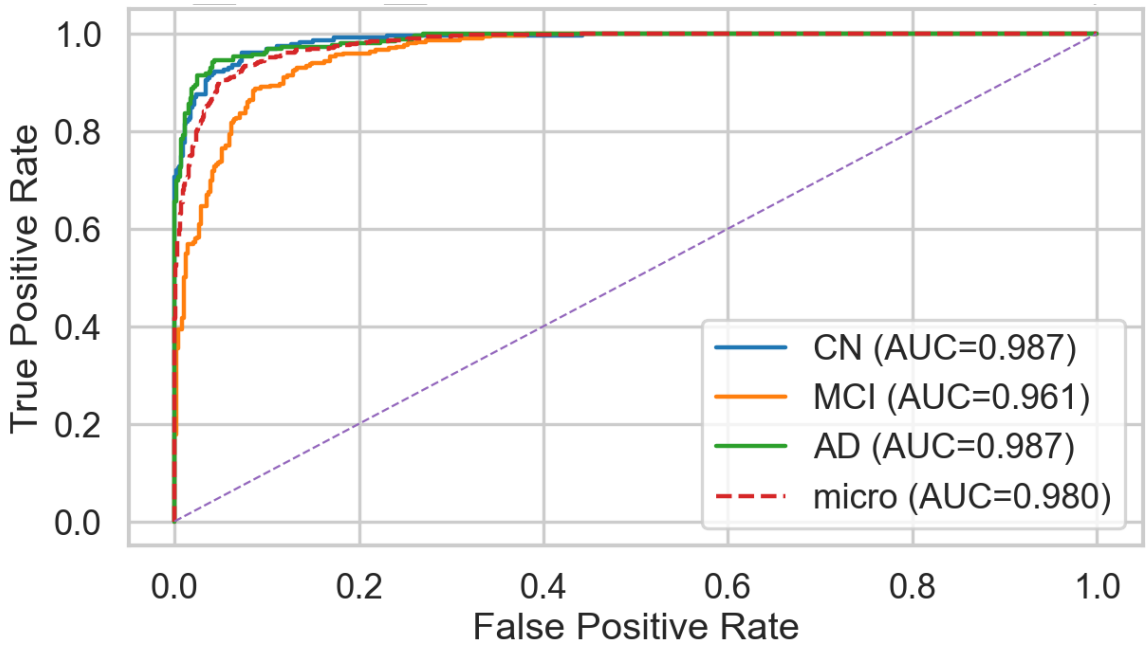

Figure 24. ROC Curve of best performing model

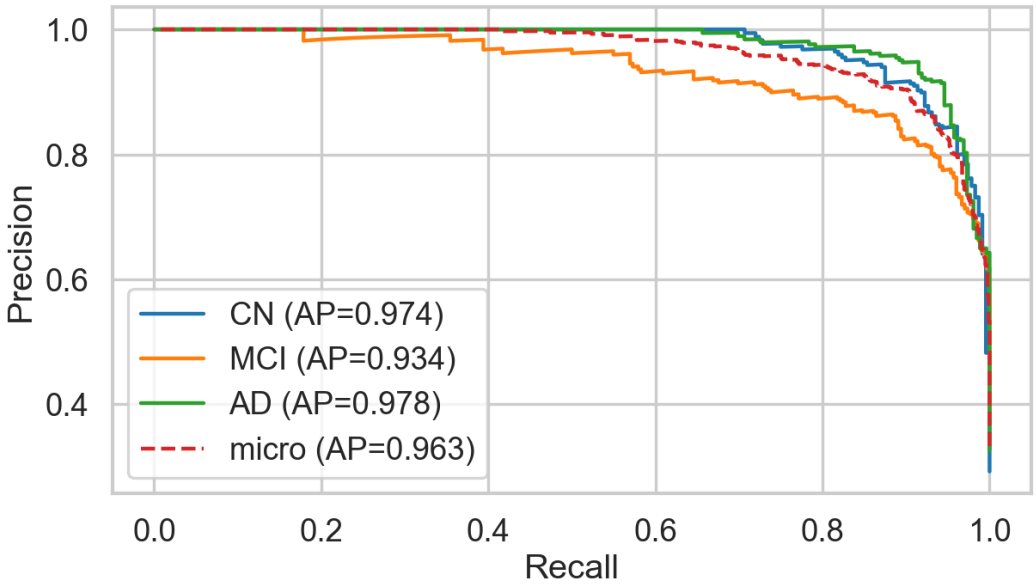

Figure 25. Precision-Recall curve for best performing model

Dimensionality-reduction diagnostics for the joint Chi-square view show that 300 SVD components account for roughly two-thirds of the variance on average across folds (shown in Figure 26). The cumulative explained-variance curve rises smoothly, with no sharp elbow within the retained dimensionality. This shape suggests that predictive signal is distributed across many low-variance directions rather than concentrated in a handful of dominant axes. In practice, the chosen dimensionality offers a good bias–variance compromise: it is sufficiently expressive to preserve discriminative structure for the meta-learner while still regularising the high-dimensional genotype space and curbing overfitting. The explained-variance behaviour is captured in Figure and is consistent with the out-of-fold generalisation metrics reported above.

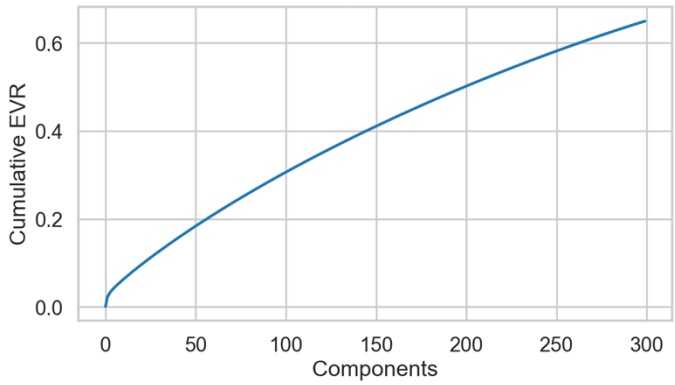

Figure 26. SVD Cumulate Explained Variance (mean across folds) for the best performing model.

The tuned hyperparameters of the four base learners in the best joint model offer additional context. The Random Forest selected 358 trees with a max depth of 16, sqrt feature sampling, and conservative split settings, indicating a preference for breadth with controlled depth to avoid overfitting. LightGBM converged to 62 leaves and a learning rate of 0.062, with moderate subsample and colsample_bytree rates and a small L2 regularisation term, a typical configuration for stable multiclass performance. XGBoost selected a max depth of 8, a learning rate of 0.03, subsample of 0.75, colsample_bytree of 0.72, and a light lambda regulariser, consistent with a balanced bias–variance trade-off. The MLP used a two-hidden-layer architecture with a batch size of 128, a learning rate of 1.1×10^{-3}, and an alpha of 9.7×10^{-4}, which act as a modest capacity non-linear learner that contributes complementary decision boundaries to the ensemble without dominating it. These settings describe well-regularised and diverse base learners, which is the regime where stacking provides the largest benefit.

The joint Chi-square four-model stack outperformed the strongest unimodal baselines. The best genetic-only configuration, which also used the four-model stack with ANOVA selection, achieved macro-AUC =0.976, macro-F1 =0.901, and balanced accuracy =0.900, slightly lower than the joint counterpart. This suggests the genetic view carried most discriminative information. The best MRI-only run-again using the four-model stack and Chi-square selection achieved macro-AUC =0.719, macro-F1 =0.524, and balanced accuracy =0.521, indicating the low stand-alone utility of the MRI feature set. These results align with the modality averages cited earlier and reflect the relative signal-to-noise ratios of the inputs, not limitations of the learning algorithms.

Trends within each modality mirror general expectations for selector-model interactions. ANOVA and Chi-square were the most reliable selectors for genetic features, as they suit both F-test rankings and Chi-square scoring. L1-penalised logistic selection worked but trailed, likely due to sparsity pressure discarding weak but complementary SNPs. Mutual information averaged lowest, possibly due to its sensitivity to estimation noise in high dimensions. Chi-square slightly favoured joint features, likely because it harmonised the combined feature space and class-conditional shifts manifested as robust frequency differences. For MRI-only, all selectors converged to a similar and lower plateau, with observed differences small compared to genetic and joint models. OOF probability calibration appeared reasonable based on ROC and PR curves, and the meta-learner's use of

entropy features down-weighted overly confident base predictions. Micro-averaged curves tracked closely to macro summaries, consistent with balanced class supports. The absence of CN↔AD confusions suggests robust decision surface separation, with ambiguity confined to MCI boundaries. The balanced accuracy of 0.904 and macro-F1 of 0.905 for the best model reflect genuine class-wise balance, not dominance by a single class.

In summary, the experimental results show a coherent story. Combining genetic and MRI information, selecting features with Chi-square, compressing to a moderate number of SVD components, and stacking diverse base learners yields the best and most stable out-of-fold performance. Genetic-only systems are close, suggesting genetic information is highly informative. MRI-only systems are comparatively weak. The error structure is clinically plausible and concentrated at CN–MCI and MCI–AD boundaries. The tuned hyperparameters of the models are in a regularised regime that aligns with strong generalisation. The variance-explained diagnostics suggest predictive information is distributed across many components. The remainder of this chapter interprets these findings and explores their implications for Alzheimer's disease stratification.

6. Discussion

Three key findings emerge from the data: the effectiveness of model diversity in conjunction with careful dimensionality reduction for high-dimensional biomedical prediction; the predominance of genetic signal for cross-sectional CN/MCI/AD discrimination in this cohort; and the added value-but not dominance-of MRI when fused with genetics.

6.1 Complementary nature of MRI data for Genetic Data

While MRI's standalone performance is limited, the joint view enhances the genetic-only picture, suggesting supplementary information beyond SNP data. Though the best joint model's macro-AUC and macro-F1 only marginally outperform the best genetic model, the observed benefits of fusion were incremental. This equilibrium makes sense. Genomics provides upstream risk profiles spread over many loci with minor impacts, while MRI volumetrics and features indicate downstream anatomical manifestations of disease processes. When combined, the MRI perspective helps define decision boundaries by identifying people whose imaging phenotype deviates from age-matched norms but whose genetic risk may be moderate. Genetic characteristics can smooth out measurement noise and

inter-scanner variability to stabilise forecasts for those whose MRI readings fall close to class boundaries. Thus, the slight improvement from fusion is consistent with MRI providing case-specific refinement and genetics mostly explaining the between-class signal.

6.2 Error Topology

The error topology of the optimal model provides further evidence of clinically coherent bounds. The absence of CN↔AD misclassifications indicates high confidence in differentiating between the illness extremes. Most mistakes occur through MCI, which varies in biology and diagnosis. MCI describes people heading to AD, those who remain stable, and those with non-AD cognitive problems. Unsurprisingly, CN–MCI and MCI–AD boundaries are permeable to an algorithm trained on aggregated labels. Since the model's errors are conservative and rarely jump across the spectrum, CN and AD errors drift exclusively toward MCI and MCI, respectively. This makes the error profile safer for triage and early warning. For instance, biasing toward precision for AD when prioritising diagnostic resources or sensitivity for the AD vs. rest decision when screening could be exploited by threshold-tuning.

6.3 Feature Selection trends

The selector-specific trends show how the data structure interacts with statistical filters. Chi-square selection performed well for the joint view because class-conditional shifts are strong frequency differences that the Chi-square score captures, and min–max scaling and non-negativity equalise genetic doses and MRI features. Under balanced classes and high sample counts per allele, class-conditional changes in per-locus means that the F statistic represents well are likely to be reflected in ANOVA for genetic-only inputs. L1's poorer performance warns of removing weak but complementary predictors that, when combined, have significant discriminative power in complex disease genetics where the signal is distributed across numerous features with tiny effects. Mutual information is strong, but estimation noise can make it fragile in high dimensions. Overall, the signal structure is better suited to selectors that promote broad, moderate feature inclusion and then use SVD to compress correlated directions than to those that impose great sparsity.

6.4 SVD explained-variance profile

The SVD explained-variance profile supports this interpretation. Without a severe elbow, the cumulative curve climbs gently and crosses around two-thirds of the variance at 300

components. This morphology is typical of ‘long-tail’ structures, where each axis carries a small amount of predictive information. Maintaining a moderate number of components balances expressivity and regularisation. Too few remove weak signals, while too many raise variance without benefits. This trade-off appears successful based on robust out-of-fold metrics with 300 components. A significant increase might yield diminishing returns, but the lack of an elbow suggests a slight additional component increase could yield marginal profits if computational budget permits.

6.5 Ensemble Behaviour

The collective behaviour is worth noting. The best solo performers were single models based on gradient-boosting (XGBoost and LightGBM), which handled varied feature scales and captured non-linear interactions. The MLP, adjusted to a limited capacity, produced non-linear decision boundaries. Random Forests provided baselines with reduced variance and stability. The meta-learner exploited variations in probability calibration and error diversity. The best run’s adjusted hyperparameters showed that all four base models were in a regularised regime, which is advantageous for stacking as it prevents overfitting. This regime included a limited-capacity MLP with explicit regularisation, a middle-sized forest, moderate depths and learning rates for the boosted trees. This balance likely explains why the MLP, when added to the tree-based learners, improved the joint model. The meta-learner also restrained overconfident base predictions for ambiguous circumstances by using an entropy feature.

6.6 Relative weakness of MRI-only

Given the limitations of MRI alone, a cautious interpretation is warranted. The specific MRI feature set used here—single-visit, summary measurements after simple preprocessing—is the focus, not the irrelevance of MRI to AD. Richer representations (e.g., regional cortical thickness, subfield segmentation, texture embeddings, or deep feature maps from raw volumes) and longitudinal change metrics enhance MRI’s discriminative power. Therefore, the modest MRI-only scores likely reflect feature representation ceiling effects, rather than a fundamental imaging constraint. MRI still provides value to the joint configuration: even a basic imaging summary can refine a genetically driven boundary when used sparingly.

6.7 Deployment Prospectives

A genetics-first screening strategy would already address much of the potential

discrimination in this dataset when genotyping is available and MRI acquisition is costly or logistically limited. Fusion produces a safer error profile, preventing CN↔AD flips, and improves performance in situations where MRI is easily accessible. A combination of varied base learners, SVD, and per-fold feature selection is a reliable formula for strong out-of-fold generalisation in both cases.

6.8 Limitations

Several restrictions discourage broad generalisations. The evaluation is within a single cohort; the gold standard is external validation on a geographically or chronologically separate dataset. OOF processes provide almost unbiased estimates when tuning is done within folds. The label MCI is time-dependent and heterogeneous, so ground-truth noise is unavoidable without longitudinal results, which lowers ceiling performance. Richer features may change the relative contributions of modalities; MRI features were limited to those in the CSV and may not represent state-of-the-art neuroimaging representations. A selected, post-QC subset of variants formed the basis for the genetic model; decisions on minor-allele thresholds, Hardy-Weinberg filtering, and LD pruning define the hypothesis space and may have cohort-specific interactions with feature selectors.

6.9 Implications for Practice and Future Work

Future research and practice have several implications. First, the CN–MCI–AD boundary shape suggests that thresholded, cost-sensitive decision rules could be adjusted for specific objectives, such as prioritising specialist evaluation or screening for AD. Second, model calibration should be optimised to enhance decision-theoretic performance at clinically significant operating points. Third, adding richer volumetric characteristics and longitudinal deltas could reinforce the MRI pathway, improving the joint model even without significant gains in stand-alone MRI performance. Fourth, incorporating feature grouping or hierarchical selection may enhance interpretability and stability without compromising discrimination. Fifth, cross-view consistency features, which detect discrepancies between genetic and MRI probabilities, could be included in the meta-learner.

The genetic view has the most discriminative power for cross-sectional CN/MCI/AD classification in this cohort. Fusion with MRI produces a clinically reasonable error profile and a moderate but significant uplift. The optimal setup, including joint Chi-square selection, 300-component SVD, and a four-model stack, achieved macro-AUC =0.978, macro-F1 =0.905, and balanced accuracy =0.904 on out-of-fold evaluation. Combined with calibrated

thresholds and interpretability tools, this model is a strong contender for risk classification and triage. Future research should prioritise external validation, richer MRI representations, meticulous calibration, and fairness analysis to ensure reliable and equitable performance in real-world scenarios.

7. Conclusion

The aim of this work was to create and validate a principled, end-to-end pipeline for multi-class AD staging that starts with rigorous WGS and ends with a stacked ensemble combining structural MRI and genetics. The main argument was that algorithmic robustness, through leak-resistant preprocessing and diversity-aware ensembling, and biological credibility, through strict QC, stratification correction, and GWAS-guided feature selection, were necessary for clinically useful performance [47], [54], [58].

Empirical evidence supports this reasoning across experiments. The first set of fundamental genetics procedures performed as expected: Manhattan and Q-Q plots showed a dominant APOE signal with low inflation, suggesting that confounding was well controlled using LD-pruned PCA variables. Poorly managed population structure can inflate apparent accuracy, so this is crucial for genetic feature models. In our case, the GWAS validated the biological signal and provided an educated ranking for feature selection.

Second, the modelling architecture showed the anticipated advantages of stacking. Complementary inductive biases were supplied by the Level-0 ensemble (RF, XGB, LGBM, and MLP) across the SVD-compressed views generated by ANOVA, χ^2, L1, and mutual information. Under rigorous out-of-fold evaluation, the Level-1 LightGBM meta-learner produced calibrated, high-discrimination predictions by taking advantage of these complementary error profiles. This approach outperformed single learners and smaller ensembles, especially in the joint modality where the variety of inputs and models reduces variance and stabilises decision limits [47]–[50].

Third, multimodal fusion significantly improved performance. Combining Genetics+MRI outperformed genetics-only and MRI-only. While combining modalities combined stable genetic risk gradients with neuroanatomical markers of disease stage, improving both macro-AUC and macro-F1, the modest value of MRI-only likely reflects the limited separability of

single-timepoint cortical and subcortical measures compared to the breadth of genome-wide signal. Using the whole four-model stack with χ^2-selected features yielded the best results, with macro-AUC 0.978, macro-F1 0.905, balanced accuracy 0.904, and overall accuracy 0.903. There were no direct CN↔AD misclassifications, and the mistakes were concentrated in the MCI boundary, where clinical ambiguity is highest, according to the confusion matrix and class-wise metrics, which demonstrated a clinically reasonable error profile.

These results support three assertions. First, integrating QC, stratification control, and GWAS-guided feature selection into a single, leak-resistant approach provides a consistent, biologically credible path from raw WGS and harmonised MRI to robust, three-class diagnostic prediction. In complex, high-dimensional biomedical situations, stacking heterogeneous learners on numerous, per-fold designed perspectives aligns with proven ensemble learning theory and offers measurable advantages over single models. Third, multimodal fusion is synergistic. While MRI refines stage-specific patterns, genetics provides a stable, cohort-portable signal; when combined, they produce benefits that neither could separately.

There are two main implications. Methodologically, the paper stresses that choosing the right classifier is equally important as carefully handling data loss. To avoid overly optimistic estimates that have hindered clinical translation, all preprocessing, selection, and dimensionality reduction are done within CV folds, and hyperparameters are adjusted within nested inner folds [56]–[58].

In essence, the joint model suggests practical tools for risk enrichment and trial pre-screening, especially when genotyping is easily accessible and MRI may be added later.

However, there are several limitations. Performance on external datasets and across other ancestries hasn’t been established. The analyses are cross-sectional and limited to a single cohort. More longitudinal or subfield-level metrics should improve separability, especially around MCI, as MRI features are summaries of a single visit. The deployment contexts contribute additional drift (scanner variation, recruitment disparities) that need to be quantified prospectively, even though out-of-fold evaluation is an unbiased estimator of generalisation. Further investigations, such as embedded sparse models, stability selection, or biologically limited feature grouping, could refine the genetic and joint feature spaces, even though χ^2 and ANOVA selectors performed well in this case [22], [55], [58].

The dissertation's clear contributions are a transparent pipeline from WGS QC to GWAS to a stacked ensemble, thorough nested validation, and proof that combining genetics and structural MRI produces superior three-way classification of CN, MCI, and AD in ADNI. This integrated approach offers a competitive multi-modal decision support template, true to genetic epidemiology as sequencing and imaging pipelines standardise.

Algorithmically disciplined and physiologically based pipelines are advantageous for AD classification. This work illustrates converting WGS and MRI information into calibrated predictions with clinically interpretable error profiles by combining rigorous statistical genetics and carefully designed stacking on per-fold perspectives. Future enhancements, such as external validation, longitudinal progression modelling, multi-ancestry calibration, and adding additional omics, will transition the field from intriguing research curves to reliable patient-centred tools.

8. Future Work

The current study demonstrates that calibrated, high-discrimination three-way categorisation in ADNI can be achieved using a methodical, leak-resistant pipeline that starts with strict WGS-level quality control and ancestry correction and ends with layered ensembling across per-fold designed views. These achievements should be viewed as a foundation for more extensive generalisation, clinical credibility, and biological understanding. The contributions and limitations lead to methodological and translational threads that should be pursued to improve scientific interpretability and external validity. External validation under a real dataset shift is a top priority. Out-of-fold evaluation with nested tuning cannot replace testing on geographically or chronologically separate samples. Prospective evaluation on a held-out ADNI wave and independent resources that integrate clinical, imaging, and genetic data is required to evaluate the stability of the stacked architecture and uncover failure modes. External validation should also include explicit probability calibration at clinically relevant operating points to ensure that risk thresholds used for referral or trial pre-screening are decision-theoretically sound. The dissertation's discussion highlights that real-world drift may be caused by differences in recruitment and scanner variance, so it's crucial to prospectively assess these impacts before deployment.

The second extension axis is longitudinal modelling. Designations like MCI are diverse and time-dependent, and some people will transition after the snapshot used here. So, the current

analysis is cross-sectional. Instead of focusing on static diagnosis, time-to-event or trajectory models that encode within-person change in cortical thickness, hippocampus subfields, or composite cognitive scores would address progression risk and staging. ADNI was designed to identify longitudinal biomarkers, and repeated measures may help define the areas with the lowest cross-sectional signal, especially near MCI. Incorporating fold-wise feature creation into the CV loop would enhance temporal structure while maintaining the pipeline's leak-resistance by tracking deltas, slopes, and subject-specific baselines. The MRI pathway's representation learning is a third thread. The single-visit, summary morphometry's low discriminative capability may be due to feature ceiling effects. Future studies should assess richer volumetric representations, such as texture descriptors, subfield-resolved hippocampus metrics, and embeddings trained from raw volumes using regularised convolutional encoders. These improvements might complement the Genetics+MRI fusion and produce a safer error profile, even if MRI is only used for cross-sectional classification. The current ensemble's behaviour predicts this result. To prevent optimistic bias, deep feature maps must be learnt within tightly leak-proof folds.

Genetics offers two intriguing avenues. Firstly, addressing 'missing heritability' beyond common-variant dosages may involve rare-variant and structural-variant loads. Pathway-anchored aggregations and gene-level collapsing tests enhance interpretability and stability without compromising discrimination. Secondly, incorporating cohort-specific association evidence into learning, beyond p-value ranking, may produce biologically based and algorithmically well-conditioned feature sets. This thorough variant finding is becoming feasible due to declining sequencing costs and developing WGS processes.

Long-term focus is crucial for generalisability across ancestries and subgroups. Models trained on predominantly European cohorts may show calibration drift and performance loss when transferred to other populations due to polygenic signals and risk scores' sensitivity to ancestral composition. Therefore, systematic fairness analyses, reporting global ROC-AUC and F1, fold-wise stratification checks, and subgroup calibration curves are essential. Group-aware meta-learning or reweighting techniques, evaluated rigorously, may be compared against ancestry-invariant representations learnt under adversarial restrictions when sample sizes permit. Examining age, sex, and ancestry impacts before deployment is crucial for future clinical translation.

Transparency of the model and biological plausibility should progress simultaneously. Sparsity-aware selectors and stacked tree ensembles can be explained post-hoc, but genomic characteristics are highly collinear, and the model has multiple views. Stability selection and permutation tests should be used to triangulate per-fold SHAP-style explanations. Gene-set and pathway enrichment analysis should link priority loci and regions to external knowledge. To confirm discriminative regions match established AD pathways, attributions in the imaging view should be traced back to neuroanatomy. Using the APOE-centric signal and the low inflation previously shown as internal positive controls, fold-wise top features can be cross-referenced with cohort GWAS findings in the genetic view to evaluate concordance. By taking these actions, explanations will move from descriptive images to verifiable biological assertions.

Diversifying the Level-0 ensemble and enhancing the meta-learner's context awareness at the architectural level is possible. A leak-proof stack of gradient-boosted trees, a regularised MLP, and a random forest produces complementary error profiles and robust calibration. Cross-view consistency features, which measure divergence between the genetics-only and MRI-only posteriors, can further enhance ambiguous case detection. Confidence-aware loss functions that punish overconfidence near the CN–MCI boundary, where most errors cluster, can be added to the entropy-based features of the meta-learner. Further research should be done on other base learners, such as linear large-margin models with calibrated outputs, using layered tuning and ablations to measure genuine marginal value beyond increased computing footprint.

Strong, repeatable engineering is crucial for translational preparedness. Data versioning and immutable fold partitions prevent silent leakage, while containerised workflows encapsulating PLINK QC, LD-pruned PCA, GWAS, and fold-wise feature engineering facilitate pipeline auditing and extension. Decision-curve analysis and net-benefit calculation should be used in clinical settings with traditional measures to determine whether and how model-assisted triage improves outcomes compared to standard care. A joint pathway for centres where MRI can be added opportunistically to sharpen boundaries and reduce clinically implausible errors, and a genotyping-first screen that captures much of the achievable discrimination when imaging is constrained, are operational profiles that merit prototyping. Such a pragmatic understanding of use-cases is already argued for in the dissertation.

To strengthen the connection between biology and explanation, model longitudinal trajectories, upgrade MRI and genetics representations to prevent leaks, examine ancestry transfer and subgroup fairness, and harden the pipeline for future use, we can transform a high-performing research system into a clinically valid, generalisable, and interpretable decision-support tool. As sequencing becomes more commonplace and imaging pipelines continue to standardise, this tool will remain true to the rigour of statistical genetics and ensemble-learning principles.

9. References

[1]
C. Patterson, “World Alzheimer Report 2018,” 2018. Available: https://www.alzint.org/u/WorldAlzheimerReport2018.pdf

[2]
R. L. Frozza, M. V. Lourenco, and F. G. De, “Challenges for Alzheimer’s Disease Therapy: Insights from Novel Mechanisms Beyond Memory Defects,” *Frontiers in Neuroscience*, vol. 12, Feb. 2018, doi: https://doi.org/10.3389/fnins.2018.00037.

[3]
D. Brooker, J. L. Fontaine, S. Evans, J. Bray, and K. Saad, “Public health guidance to facilitate timely diagnosis of dementia: ALzheimer’s COoperative Valuation in Europe recommendations,” *International Journal of Geriatric Psychiatry*, vol. 29, no. 7, pp. 682–693, Jan. 2014, doi: https://doi.org/10.1002/gps.4066.

[4]
A. Milne, “Dementia screening and early diagnosis: The case for and against,” *Health, Risk & Society*, vol. 12, no. 1, pp. 65–76, Feb. 2010, doi: https://doi.org/10.1080/13698570903509497.

[5]
J. E. Gaugler, H. Ascher-Svanum, D. L. Roth, T. Fafowora, A. Siderowf, and T. G. Beach, “Characteristics of patients misdiagnosed with Alzheimer’s disease and their medication use: an analysis of the NACC-UDS database,” *BMC Geriatrics*, vol. 13, no. 1, Dec. 2013, doi: https://doi.org/10.1186/1471-2318-13-137.

[6]

A. Goate *et al.*, “Segregation of a missense mutation in the amyloid precursor protein gene with familial Alzheimer’s disease,” *Nature*, vol. 349, no. 6311, pp. 704–706, Feb. 1991, doi: https://doi.org/10.1038/349704a0.

[7]

R. Sherrington *et al.*, “Cloning of a gene bearing missense mutations in early-onset familial Alzheimer’s disease,” *Nature*, vol. 375, no. 6534, pp. 754–760, Jun. 1995, doi: https://doi.org/10.1038/375754a0.

[8]

E. Levy-Lahad *et al.*, “Candidate gene for the chromosome 1 familial Alzheimer’s disease locus,” *Science*, vol. 269, no. 5226, pp. 973–977, Aug. 1995, doi: https://doi.org/10.1126/science.7638622.

[9]

P. C. Ng and E. F. Kirkness, “Whole Genome Sequencing,” *Methods in molecular biology*, pp. 215–226, Jan. 2010, doi: https://doi.org/10.1007/978-1-60327-367-1_12.

[10]

G.-B. Chen, “Estimating heritability of complex traits from genome-wide association studies using IBS-based Hasemanâ€“Elston regression,” *Frontiers in Genetics*, vol. 5, Apr. 2014, doi: https://doi.org/10.3389/fgene.2014.00107.

[11]

B. MEI and Z. WANG, “An efficient method to handle the ‘large p, small n’ problem for genomewide association studies using Haseman–Elston regression,” *Journal of Genetics*, vol. 95, no. 4, pp. 847–852, Dec. 2016, doi: https://doi.org/10.1007/s12041-016-0705-3.

[12]

P. S. Kohli and S. Arora, “Application of Machine Learning in Disease Prediction,” *2018 4th International Conference on Computing Communication and Automation (ICCCA)*, Dec. 2018, doi: https://doi.org/10.1109/ccaa.2018.8777449.

[13]

S. Uddin, A. Khan, M. E. Hossain, and M. A. Moni, “Comparing different supervised machine learning algorithms for disease prediction,” *BMC Medical Informatics and Decision Making*, vol. 19, no. 1, Dec. 2019, doi: https://doi.org/10.1186/s12911-019-1004-8.

[14]

D. Jain and V. Singh, “Feature selection and classification systems for chronic disease prediction: A review,” *Egyptian Informatics Journal*, vol. 19, no. 3, pp. 179–189, Apr. 2018, doi: https://doi.org/10.1016/j.eij.2018.03.002.

[15]
E. H. Corder *et al.*, “Gene Dose of Apolipoprotein E Type 4 Allele and the Risk of Alzheimer’s Disease in Late Onset Families,” *Science*, vol. 261, no. 5123, pp. 921–923, Aug. 1993, doi: https://doi.org/10.1126/science.8346443.
[16]
R. Mahley, “Apolipoprotein E: cholesterol transport protein with expanding role in cell biology,” *Science*, vol. 240, no. 4852, pp. 622–630, Apr. 1988, doi: https://doi.org/10.1126/science.3283935.
[17]
C. T. Lin, Y. F. Xu, J. Y. Wu, and L. Chan, “Immunoreactive apolipoprotein E is a widely distributed cellular protein. Immunohistochemical localization of apolipoprotein E in baboon tissues.,” *Journal of Clinical Investigation*, vol. 78, no. 4, pp. 947–958, Oct. 1986, doi: https://doi.org/10.1172/jci112685.
[18]
Nabil Elshourbagy, Warren S.-L. Liao, R. W. Mahley, and J. B. Taylor, “Apolipoprotein E mRNA is abundant in the brain and adrenals, as well as in the liver, and is present in other peripheral tissues of rats and marmosets.,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 82, no. 1, pp. 203–207, Jan. 1985, doi: https://doi.org/10.1073/pnas.82.1.203.
[19]
“Risk of Alzheimer Disease with the ±4 Allele for... : Alzheimer Disease & Associated Disorders,” *LWW*, 2025. https://journals.lww.com/alzheimerjournal/abstract/1998/03000/risk_of_alzheimer_disease_with_the__4_allele_for.6.aspx (accessed Aug. 21, 2025).
[20]
M. Landen, A. Thorsell, A. Wallin, and K. Blennow, “The apolipoprotein E allele epsilon 4 does not correlate with the number of senile plaques or neurofibrillary tangles in patients with Alzheimer’s disease.,” *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 61, no. 4, pp. 352–356, Oct. 1996, doi: https://doi.org/10.1136/jnnp.61.4.352.
[21]
NIHAging, “Alzheimer’s Disease Genetics Fact Sheet,” *National Institute on Aging*, Mar. 2023. https://www.nia.nih.gov/health/alzheimers-causes-and-risk-factors/alzheimers-disease-genetics-fact-sheet (accessed Aug. 21, 2025).
[22]

J. Schwarzerova *et al.*, “A perspective on genetic and polygenic risk scores-advances and limitations and overview of associated tools,” *Briefings in Bioinformatics*, vol. 25, no. 3, Mar. 2024, doi: https://doi.org/10.1093/bib/bbae240.
[23]
R. P. Igo, T. G. Kinzy, and J. N. Cooke Bailey, “Genetic Risk Scores,” *Current Protocols in Human Genetics*, vol. 104, no. 1, Nov. 2019, doi: https://doi.org/10.1002/cphg.95.
[24]
A. Simona, W. Song, D. W. Bates, and C. F. Samer, “Polygenic risk scores in pharmacogenomics: opportunities and challenges-a mini review,” *Frontiers in Genetics*, vol. 14, Jun. 2023, doi: https://doi.org/10.3389/fgene.2023.1217049.
[25]
Y. Ding *et al.*, “Large uncertainty in individual polygenic risk score estimation impacts PRS-based risk stratification,” *Nature Genetics*, vol. 54, no. 1, pp. 30–39, Dec. 2021, doi: https://doi.org/10.1038/s41588-021-00961-5.
[26]
J. M. Fullerton and J. I. Nurnberger, “Polygenic risk scores in psychiatry: Will they be useful for clinicians?,” *F1000Research*, vol. 8, pp. 1293–1293, Jul. 2019, doi: https://doi.org/10.12688/f1000research.18491.1.
[27]
B. Fan, Z.-Q. Du, D. M. Gorbach, and M. F. Rothschild, “Development and Application of High-density SNP Arrays in Genomic Studies of Domestic Animals,” *Asian-Australasian Journal of Animal Sciences*, vol. 23, no. 7, pp. 833–847, Jun. 2010, doi: https://doi.org/2010.23.7.833.
[28]
F. C. Ceballos, S. Hazelhurst, and M. Ramsay, “Assessing runs of Homozygosity: a comparison of SNP Array and whole genome sequence low coverage data,” *BMC Genomics*, vol. 19, no. 1, Jan. 2018, doi: https://doi.org/10.1186/s12864-018-4489-0.
[29]
M. Khani, E. Gibbons, J. Bras, and R. Guerreiro, “Challenge accepted: uncovering the role of rare genetic variants in Alzheimer’s disease,” *Molecular Neurodegeneration*, vol. 17, no. 1, Jan. 2022, doi: https://doi.org/10.1186/s13024-021-00505-9.
[30]
“DNA Sequencing Costs: Data,” *Genome.gov*, 2019. https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data (accessed Aug. 22, 2025).

[31]
S. T. Park and J. Kim, “Trends in Next-Generation Sequencing and a New Era for Whole Genome Sequencing,” *International Neurourology Journal*, vol. 20, no. Suppl 2, pp. S76-83, Nov. 2016, doi: https://doi.org/10.5213/inj.1632742.371.
[32]
C. G. van El *et al.*, “Whole-genome sequencing in health care,” *European Journal of Human Genetics*, vol. 21, no. 6, pp. 580–584, May 2013, doi: https://doi.org/10.1038/ejhg.2013.46.
[33]
A. Khromykh and B. D. Solomon, “The Benefits of Whole-Genome Sequencing Now and in the Future,” *Molecular Syndromology*, vol. 6, no. 3, pp. 108–109, 2015, doi: https://doi.org/10.1159/000438732.
[34]
D. Smedley *et al.*, “100,000 Genomes Pilot on Rare-Disease Diagnosis in Health Care - Preliminary Report,” *New England Journal of Medicine*, vol. 385, no. 20, pp. 1868–1880, Nov. 2021, doi: https://doi.org/10.1056/nejmoa2035790.
[35]
“What is early-onset Alzheimer’s?| IU School of Medicine,” *Iu.edu*, 2025. https://medicine.iu.edu/expertise/alzheimers/research/translational/early-onset/what-is-early-onset-alzheimers (accessed Aug. 23, 2025).
[36]
IBM, “Support Vector Machine,” *Ibm.com*, Dec. 12, 2023. https://www.ibm.com/think/topics/support-vector-machine (accessed Aug. 23, 2025).
[37]
J. Daniel and J. Martin, “Speech and Language Processing,” Jan. 2023. Available: https://web.stanford.edu/~jurafsky/slp3/5.pdf
[38]
A. Gao, “Lecture 7 Decision Trees,” 2021. Available: https://www.cs.toronto.edu/~axgao/cs486686_f21/lecture_notes/Lecture_07_on_Decision_Trees.pdf
[39]
H. A. Salman, A. Kalakech, and Amani Steiti, “Random Forest Algorithm Overview,” *Deleted Journal*, vol. 2024, pp. 69–79, Jun. 2024, doi: https://doi.org/10.58496/bjml/2024/007.
[40]

"Alzheimer's Disease Neuroimaging Initiative," *ADNI*, 2022. https://adni.loni.usc.edu/about/ (accessed Aug. 23, 2025).
[41]
J. Ye *et al.*, "Sparse learning and stability selection for predicting MCI to AD conversion using baseline ADNI data," *BMC Neurology*, vol. 12, no. 1, Jun. 2012, doi: https://doi.org/10.1186/1471-2377-12-46.
[42]
K. T. N.P and D. Varghese, "A Novel Approach for Diagnosing Alzheimer's Disease Using SVM," *2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI)*, pp. 895–898, May 2018, doi: https://doi.org/10.1109/icoei.2018.8553789.
[43]
S. Dimitriadis, Dimitris Liparas, and None Alzheimer's DNI, "How random is the random forest? Random forest algorithm on the service of structural imaging biomarkers for Alzheimer's disease: from Alzheimer's disease neuroimaging initiative (ADNI) database," *Neural Regeneration Research*, vol. 13, no. 6, pp. 962–962, Jan. 2018, doi: https://doi.org/10.4103/1673-5374.233433.
[44]
S. Kim *et al.*, "Genome-wide association study of CSF biomarkers A 1-42, t-tau, and p-tau181p in the ADNI cohort," *Neurology*, vol. 76, no. 1, pp. 69–79, Dec. 2010, doi: https://doi.org/10.1212/wnl.0b013e318204a397.
[45]
L. Shen *et al.*, "Whole genome association study of brain-wide imaging phenotypes for identifying quantitative trait loci in MCI and AD: A study of the ADNI cohort," *NeuroImage*, vol. 53, no. 3, pp. 1051–1063, Jan. 2010, doi: https://doi.org/10.1016/j.neuroimage.2010.01.042.
[46]
F. Long, L. Wang, W. Cai, K. Lesnik, and H. Liu, "Predicting the performance of anaerobic digestion using machine learning algorithms and genomic data," *Water Research*, vol. 199, pp. 117182–117182, Apr. 2021, doi: https://doi.org/10.1016/j.watres.2021.117182.
[47]
Zhi-Hua Zhou, *Ensemble methods : foundations and algorithms*. Boca Raton ; London ; New York: Crc Press, Cop, 2012.
[48]

X. Dong, Z. Yu, W. Cao, Y. Shi, and Q. Ma, “A survey on ensemble learning,” *Frontiers of Computer Science*, vol. 14, no. 2, pp. 241–258, Aug. 2019, doi: https://doi.org/10.1007/s11704-019-8208-z.

[49]

Mbali Kalirane, “Bagging, Boosting and Stacking: Ensemble Learning in ML Models,” *Analytics Vidhya*, Jan. 20, 2023. https://www.analyticsvidhya.com/blog/2023/01/ensemble-learning-methods-bagging-boosting-and-stacking/ (accessed Aug. 25, 2025).

[50]

Yann LeCun, Yoshua Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015, doi: https://doi.org/10.1038/nature14539.

[51]

G. Paaß, “Deep Learning: How do deep neural networks work?,” *Lamarr Institute for Machine Learning and Artificial Intelligence*, Apr. 21, 2021. https://lamarr-institute.org/blog/deep-neural-networks/ (accessed Aug. 27, 2025).

[52]

K. R. Kruthika, Rajeswari, and H. D. Maheshappa, “Multistage classifier-based approach for Alzheimer’s disease prediction and retrieval,” *Informatics in Medicine Unlocked*, vol. 14, pp. 34–42, 2019, doi: https://doi.org/10.1016/j.imu.2018.12.003.

[53]

R. Ju, C. Hu, P. Zhou, and Q. Li, “Early Diagnosis of Alzheimer’s Disease Based on Resting-State Brain Networks and Deep Learning,” *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 16, no. 1, pp. 244–257, Nov. 2017, doi: https://doi.org/10.1109/tcbb.2017.2776910.

[54]

X. Ying, “An Overview of Overfitting and its Solutions,” *Journal of Physics: Conference Series*, vol. 1168, no. 2, p. 022022, Feb. 2019, doi: https://doi.org/10.1088/1742-6596/1168/2/022022.

[55]

I. Belcic and C. Stryker, “Feature Selection,” *Ibm.com*, Mar. 18, 2025. https://www.ibm.com/think/topics/feature-selection (accessed Sep. 03, 2025).

[56]

I. K. Nti, Owusu Nyarko-Boateng, and Justice Aning, “Performance of Machine Learning Algorithms with Different K Values in K-fold CrossValidation,” *International Journal of*

Information Technology and Computer Science, vol. 13, no. 6, pp. 61–71, Dec. 2021, doi: https://doi.org/10.5815/ijitcs.2021.06.05.
[57]
Y. Jung, "Multiple predictingK-fold cross-validation for model selection," *Journal of Nonparametric Statistics*, vol. 30, no. 1, pp. 197–215, Nov. 2017, doi: https://doi.org/10.1080/10485252.2017.1404598.
[58]
S. Parvandeh, H.-W. Yeh, M. P. Paulus, and B. A. McKinney, "Consensus features nested cross-validation," *Bioinformatics*, vol. 36, no. 10, pp. 3093–3098, Jan. 2020, doi: https://doi.org/10.1093/bioinformatics/btaa046.
[59]
M. W. Weiner *et al.*, "The Alzheimer's Disease Neuroimaging Initiative: A review of papers published since its inception," *Alzheimer's & Dementia*, vol. 8, no. 1S, Nov. 2011, doi: https://doi.org/10.1016/j.jalz.2011.09.172.
[60]
F. O. Bagger *et al.*, "Whole genome sequencing in clinical practice," *BMC Medical Genomics*, vol. 17, no. 1, Jan. 2024, doi: https://doi.org/10.1186/s12920-024-01795-w.
[61]
"FreeSurfer," *FreeSurfer*, 2025. https://surfer.nmr.mgh.harvard.edu/ (accessed Sep. 11, 2025).
[62]
B. Fischl *et al.*, "Automatically Parcellating the Human Cerebral Cortex," *Cerebral Cortex*, vol. 14, no. 1, pp. 11–22, Dec. 2003, doi: https://doi.org/10.1093/cercor/bhg087.
[63]
R. S. Desikan *et al.*, "An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest," *NeuroImage*, vol. 31, no. 3, pp. 968–980, Mar. 2006, doi: https://doi.org/10.1016/j.neuroimage.2006.01.021.
[64]
N. Spotorno, O. Strandberg, G. Vis, E. Stomrud, M. Nilsson, and O. Hansson, "Measures of cortical microstructure are linked to amyloid pathology in Alzheimer's disease," *Brain*, vol. 146, no. 4, pp. 1602–1614, Sep. 2022, doi: https://doi.org/10.1093/brain/awac343.
[65]

S. Purcell *et al.*, “PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses,” *The American Journal of Human Genetics*, vol. 81, no. 3, pp. 559–575, Aug. 2007, doi: https://doi.org/10.1086/519795.
[66]
C. A. Anderson, F. H. Pettersson, G. M. Clarke, L. R. Cardon, A. P. Morris, and K. T. Zondervan, “Data quality control in genetic case-control association studies,” *Nature Protocols*, vol. 5, no. 9, pp. 1564–1573, Aug. 2010, doi: https://doi.org/10.1038/nprot.2010.116.
[67]
S. R. Ellingson and D. W. Fardo, “Automated quality control for genome wide association studies,” *F1000Research*, vol. 5, pp. 1889–1889, Jul. 2016, doi: https://doi.org/10.12688/f1000research.9271.1.
[68]
R. P. Adelson *et al.*, “Empirical design of a variant quality control pipeline for whole genome sequencing data using replicate discordance,” *Scientific Reports*, vol. 9, no. 1, Nov. 2019, doi: https://doi.org/10.1038/s41598-019-52614-7.
[69]
G. Jun *et al.*, “Detecting and Estimating Contamination of Human DNA Samples in Sequencing and Array-Based Genotype Data,” *The American Journal of Human Genetics*, vol. 91, no. 5, pp. 839–848, Oct. 2012, doi: https://doi.org/10.1016/j.ajhg.2012.09.004.
[70]
R. Nielsen, J. S. Paul, A. Albrechtsen, and Y. S. Song, “Genotype and SNP calling from next-generation sequencing data,” *Nature Reviews Genetics*, vol. 12, no. 6, pp. 443–451, May 2011, doi: https://doi.org/10.1038/nrg2986.
[71]
W. S. Pearman, L. Urban, and A. Alexander, “Commonly used Hardy–Weinberg equilibrium filtering schemes impact population structure inferences using RADseq data,” *Molecular Ecology Resources*, vol. 22, no. 7, pp. 2599–2613, May 2022, doi: https://doi.org/10.1111/1755-0998.13646.
[72]
J. E. Wigginton, D. J. Cutler, and G. R. Abecasis, “A Note on Exact Tests of Hardy-Weinberg Equilibrium,” *The American Journal of Human Genetics*, vol. 76, no. 5, pp. 887–893, Apr. 2005, doi: https://doi.org/10.1086/429864.
[73]

A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich, "Principal components analysis corrects for stratification in genome-wide association studies," *Nature Genetics*, vol. 38, no. 8, pp. 904–909, Jul. 2006, doi: https://doi.org/10.1038/ng1847.

[74]

F. Zhang and D. Wagener, "An approach to incorporate linkage disequilibrium structure into genomic association analysis," *Journal of genetics and genomics/Journal of Genetics and Genomics*, vol. 35, no. 6, pp. 381–385, Jun. 2008, doi: https://doi.org/10.1016/s1673-8527(08)60055-7.

[75]

K. Roeder and L. Wasserman, "Genome-Wide Significance Levels and Weighted Hypothesis Testing," *Statistical Science*, vol. 24, no. 4, Nov. 2009, doi: https://doi.org/10.1214/09-sts289.

[76]

E. H. Corder *et al.*, "Gene Dose of Apolipoprotein E Type 4 Allele and the Risk of Alzheimer's Disease in Late Onset Families," *Science*, vol. 261, no. 5123, pp. 921–923, Aug. 1993, doi: https://doi.org/10.1126/science.8346443.

[77]

Mark, E. C. Polley, and A. E. Hubbard, "Super Learner," *Statistical Applications in Genetics and Molecular Biology*, vol. 6, no. 1, Jan. 2007, doi: https://doi.org/10.2202/1544-6115.1309.

Appendix

1. Ethics SOP Form : CS_REC_2 SOP2.2 Text Data.docx

2. Exploratory Data Analysis Diagrams:-

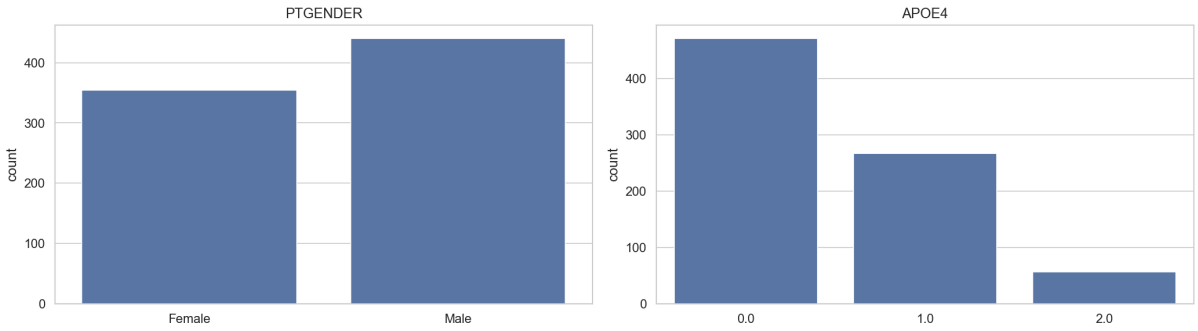

Figure 27. Age Distribution and APOE4 Distribution

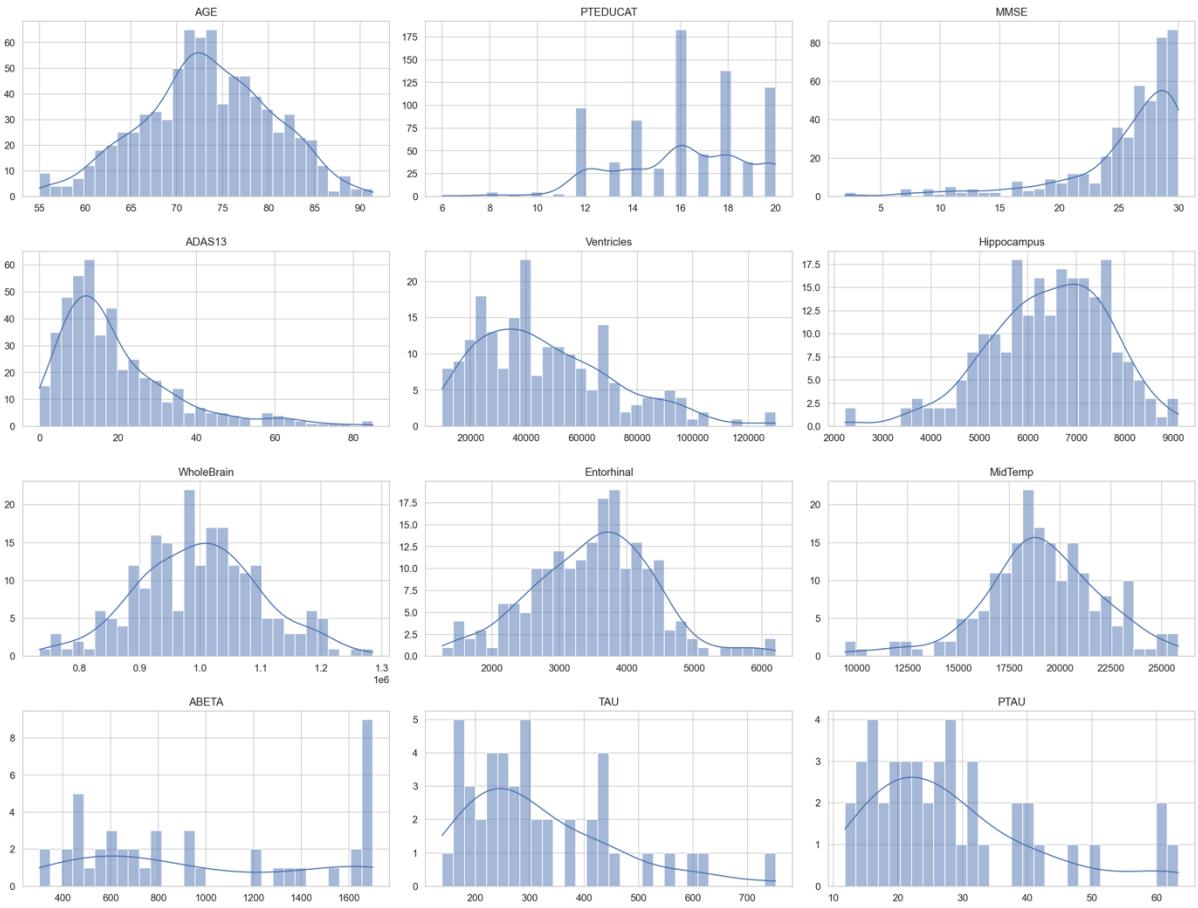

Figure 28. Distribution of Key Continuous Variables

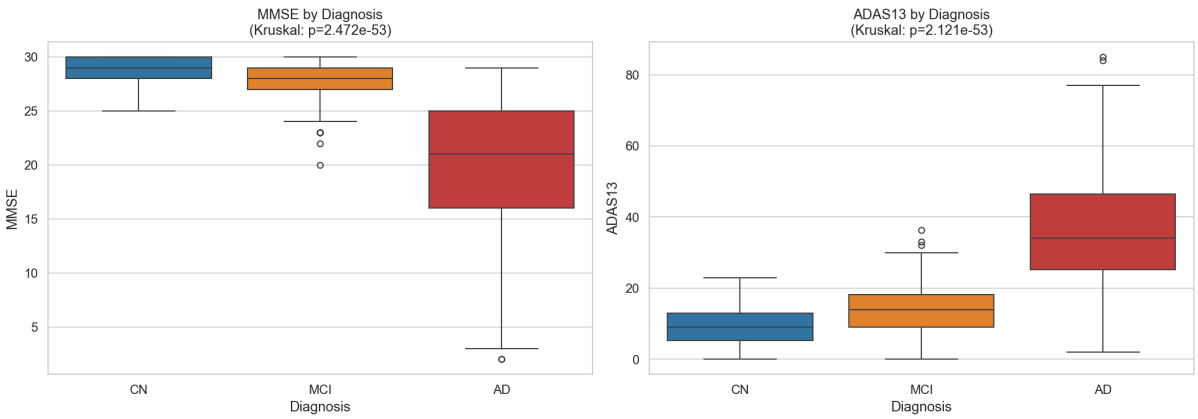

Figure 29. Cognition by Diagnosis

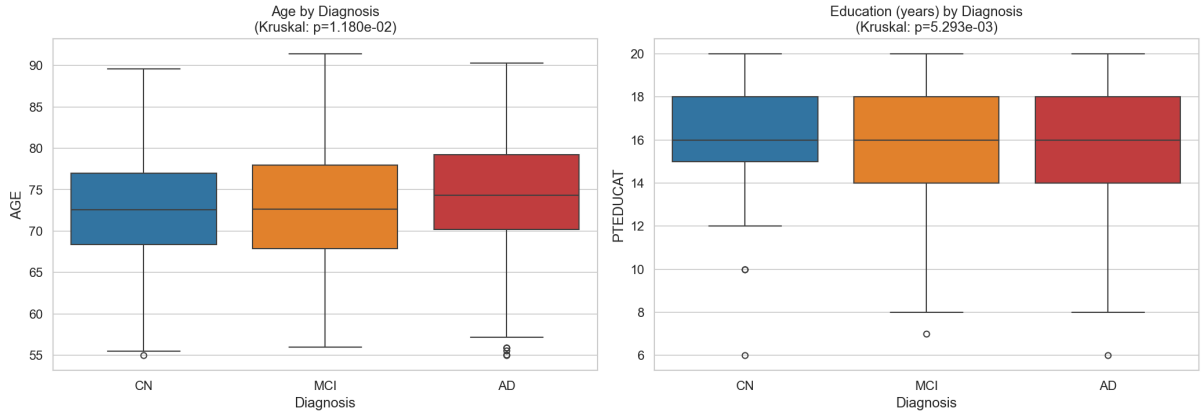

Figure 30. Demographics by Diagnosis

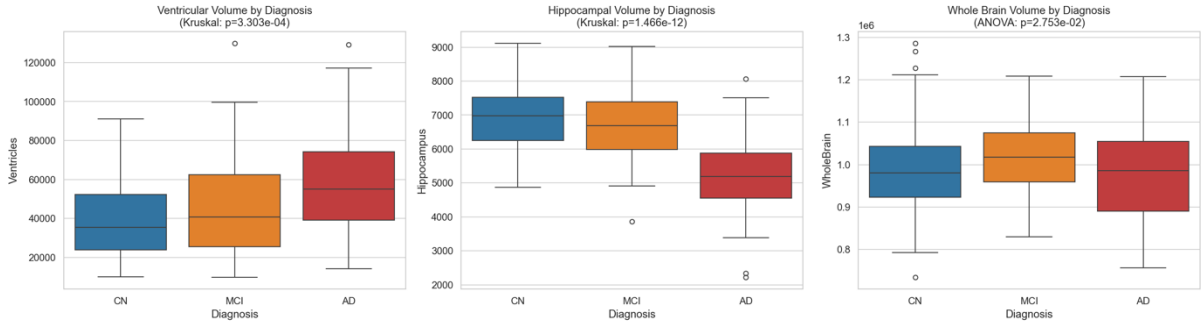

Figure 31. Key MRI Values by Diagnosis

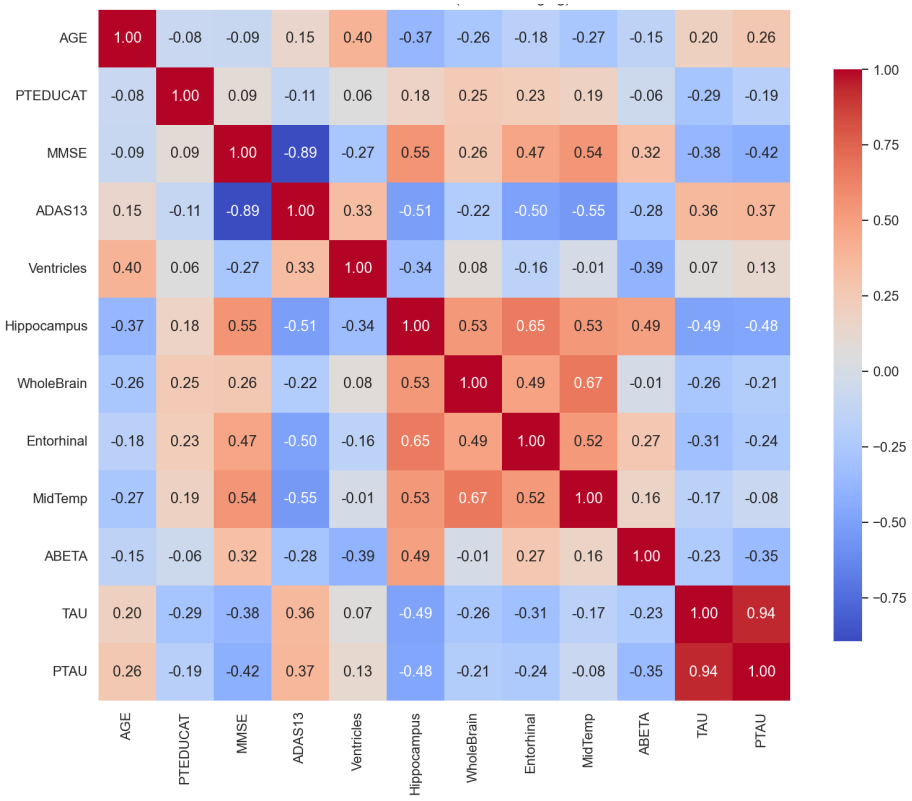

Figure 32. Pearson Correlation

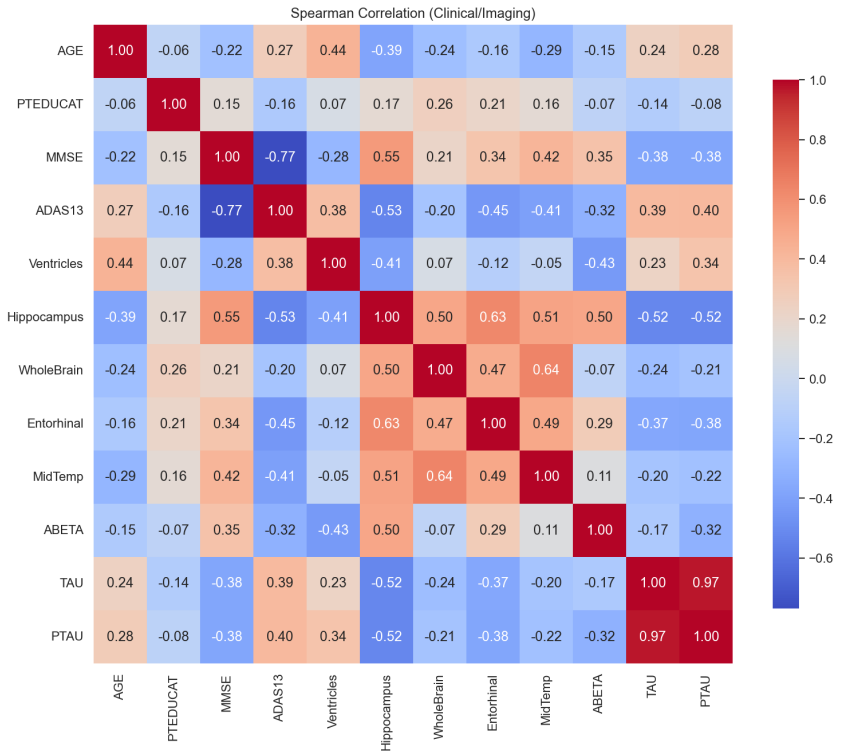

Figure 33. Spearman Correlation