Epigenetic Prediction Models for Alzheimer's Disease Progression

Pavithra Nagarajan
Thesis Advisor: Dr. Heming Wang

May 2022

1 Introduction

1.1 Background

Alzheimer's disease (AD), the leading cause of dementia, is an elusive and debilitating neurodegenerative disorder[37,38,69,89]. To this date there is no definitive cure nor a clearly identified upstream causative factor[35,37,63,89]. Early identification of elders who are at high risk for AD, before external symptomatic presentation, is imperative[52,61,109]. Predicting future AD progression may allow treatment to be administered in an optimal time period – providing tangible hope in reversing damage, stopping further progression, or modifying the course of the disease[38,61]. Simply the ability to delay further cognitive impairment will transform the lives of patients, caretakers, and their families[59,94].

Current theory supports this endeavor. Clinical presentation of AD is understood to be preceded by insidious pathological mechanisms (not limited to the brain) up to even two decades[8,11]. Along this AD 'continuum', individuals with mild cognitive impairment (MCI) are considered to be at a key standpoint between healthy aging and potential AD dementia[51,59]. However amongst all MCI patients, not all may transition to AD dementia[59]. While some may remain stable in this state and others may indeed transition, MCI patients may also revert back to being completely cognitively healthy[116]. Annually, the rate of MCI to AD transition may be 10% - 15%[59,116].

It is in this gray-area space where baseline biomarkers that identify MCI individuals who are at high risk for future AD transition fit right in. Currently, cerebrospinal fluid (CSF) biomarkers, amyloid PET, and even tau PET, exist as validated modes of information[51,107]. New modalities such as structural MRI and FDG-PET have shown incredible promise as well[107,109]. However, lumbar punctures needed for measuring CSF biomarkers are invasive[2,38,51]. Neuroimaging technology is plagued by high cost and the constraint of specialized facilities[38,116]. Thus, although such data sources are incredibly valuable for assuring proper clinical diagnosis, they may not be practical for widespread use for patient risk stratification, enrichment, or screening for clinical trials.

With this in mind, it is vital to note that late-onset AD is incredibly heterogeneous. Competing hypotheses are abundant for seeking to explain pathological pathways, including the amyloid cascade hypothesis, the inflammation hypothesis, the infection hypothesis, and the metabolic reprogramming hypothesis[2,26,37,111,118]. Amidst this theoretical controversy, toxic build-up of amyloid-beta peptide aggregates as well as hyperphosphorylated tau tangles are established as key hallmarks of the disease[53,101]. In terms of genetic factors, there is consensus in that the APOE4 e4 allele is a strong risk factor, where a single copy of the allele may confer a 4-fold elevated risk, and two copies, 12-fold[69]. However, the APOE e4 allele along with GWAS loci are able to explain only approximately 29% of an estimated total heritability of 60% - 80%, and it may also be that this allele's effect is not consistent across different ethnicities and even sex[3,31,53,107]. Contradiction further presents itself in instances where elderly with high amyloid-beta deposition do not present any signs of cognitive impairment[26].

Ultimately, today the understanding of AD has matured with strong connections to the disease found ranging from HSV-1 infection, sleep-deprivation, insulin resistance, mitochondrial dysregulation, to even menopause[8,11,26,37,69,70,89,101,118]. A glimpse of the fascinating elusivity of this disease can be perhaps showcased by the new understanding that amyloid-beta, so far regarded as an impetus for a neurotoxic environment, may in fact be an antimicrobial peptide protecting neuron cells[37,118]. This two-faced nature of amyloid is a striking revelation and reflects how much complexity AD inherently contains.

1.2 Motivation of Blood-Based Biomarkers

It is essential to understand that AD is complex in its pathology, as this necessitates a panel of biomarkers that captures multiple dimensions, and opens the search horizon to not be limited to just amyloid or tau burden. Plasma-based blood biomarkers enter this space as being prime candidates along with the necessary qualities of non-invasiveness and cost-effectiveness[2,116]. They may assist AD therapeutic research in various avenues: identifying cognitively healthy or MCI individuals that will transition to AD, granular risk profiling clinical trial participants across the AD continuum, and repeated sampling for longitudinal monitoring of treatment response[1,8,38,116].

However, peripheral blood-based biomarkers do not come without their own set of challenges. While proteins from the brain are exchanged in the cerebrospinal fluid space, very few such molecules find their way into the blood. Proteins found in the blood may be subjected to protease degradation, liver metabolism, or kidney clearance, placing into question their stability[12]. However, current evidence justifies their exploration. Plasma neutrophil levels may be associated with cognitive decline[95]. Peripheral blood neurofilament light may be a marker of neurodegeneration[95,109]. Plasma noradrenaline has shown to be correlated with MMSE score[38]. A protein kinase, GSK-3B, was found to be elevated in the peripheral blood of MCI and AD patients relative to healthy controls[2]. Bacterial infections not localized in the brain, may disrupt the blood-brain-barrier and promote neuroinflammation[37]. Diets involving glucose, ketones, or free fatty acids may differentially impact cognitive decline in AD patients[26,101].

1.3 The Promise of Epigenetics

Amidst this shifting diversity, epigenetics enters as an interesting candidate due to its ability to connect environmental factors to gene regulation, that in turn impact disease presentation[35,112]. Epigenetics carries the desirable traits of being reversible and stable, as the patterns are maintained across cell division cycles, but do not necessitate direct base pair changes[35,76,112]. DNA methylation (DNAm) is the most widely studied epigenetic mechanism partly due to the ease and availability of high-throughput assays, and is the data modality utilized for this thesis project[58,112]. DNAm involves the covalent addition of a methyl group to the 5' position of a cytosine ring, thus creating a 5-methylcytosine (5mC), and resulting in context-specific consequences[112]. For example, methylation of the enhancer regions is thought to decrease gene expression, while methylation in gene bodies may promote the opposite[112].

1.4 Peripheral Blood - Current Findings

DNAm patterns exist for AD, and were first identified through epigenome wide-association studies (EWAS) using post-mortem brain tissue[81,112]. Robust brain methylation dysregulations in AD patients have been noted, so much that this has warranted an exciting recent effort termed EWASplus in extending EWAS coverage[45]. DNAm patterns in the brain have been shown to be sensitive to age and be linked with memory[81]. Currently, there is active interest in a parallel shift in exploring DNAm patterns from peripheral blood, opposed to post-mortem brain tissue[112].

The challenge exists in that DNAm is incredibly specific, even at the cell level, and thus adjusting for key factors such as cell type proportions, lifestyle factors, and traditional batch effects are essential[14,112]. It has also been confirmed that DNAm patterns found in brain regions do not tightly overlap with DNAm patterns taken from peripheral blood tissue[112]. Despite these factors, exciting promise presents itself as new findings emerge. For example, the APP gene has been found to be consistently hypermethylated in peripheral blood and the brain[112]. A recent twin study investigating peripheral blood of twins discordant for AD, demonstrated differing DNAm signatures[53]. The COASY gene has been identified to have higher blood methylation levels in MCI and AD individuals, and to be correlated with MMSE score[51]. A recent longitudinal study showed DNAm signatures are already established in peripheral blood prior to symptomatic dementia[35].

Thus current literature involving peripheral blood-based DNAm signatures are new, evolving, and as relevant as ever. In this spirit, I sought to build prediction models for

MCI to AD progression using a peripheral blood-derived DNAm panel. To the best of my knowledge, this is new territory, as although recent work with DNAm has provided exciting EWAS findings, prediction was not explored mainly due to the fact that individual CpG sites did not pass the stringent Bonferroni threshold nor have large effect sizes[60]. I hypothesize that although individual CpGs may not provide high AUC scores, or high p-values that warrant easily seen motivation for feature selection for a biomarker panel, a collection of CpG sites in tandem may have predictive power when taken collectively. This reasoning follows recent findings that differentially methylated regions for AD, opposed to singular positions, have been found to be replicated across studies, and pass significance thresholds[80]. This thesis project serves thus as an avenue of exciting exploration, where I seek to analyze whether a peripheral blood DNAm signature indeed has potential in predicting future Alzheimer's disease progression.

1.5 Hypothesis

Specifically, my hypothesis is the following: whole blood DNA (CpG) methylation patterns exist that differentiate individuals that progress from MCI to Alzheimer's disease in 1 year, from those with MCI that do not. Identifying such sites and creating a methylation panel, carries clinical utility as it is non-invasive, cost-effective, and allows to begin treatment for Alzheimer's earlier when it may be more effective. Of course, identifying such a panel for the more general outcome of 'cognitive decline' may revolutionize monitoring disease progression. However, such a panel that generalizes temporally is difficult, and thus the first task to be addressed may be simplified to a binary classification model for the time frame of 1 year. If such a panel exists, it may justify future conscious work spent in identifying a multi-omics blood-based biomarker panel for monitoring Alzheimer's disease progression, or a collection of panels for different stages of the AD continuum.

2 Methods

2.1 Samples

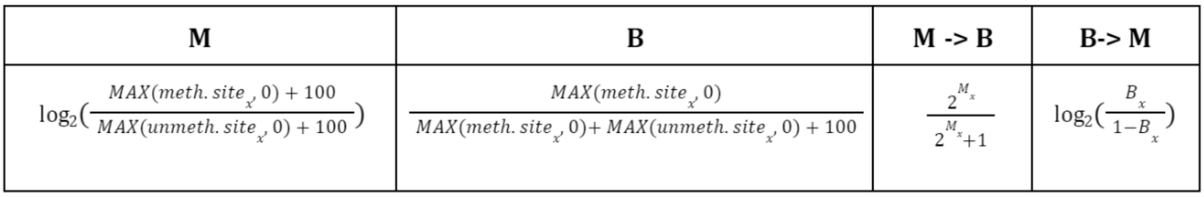
Whole blood Illumina Infinium Human Methylation 450K BeadChip array (450k array) data was made publically accessible by Roubroeks et al[86]. For 284 individuals in the AddNeuroMed cohort, 485,512 genome-wide whole blood CpG methylation measures, age, and sex were initially retrieved from GSE144858[86]. This dataset is a subset of the AddNeuroMed study, which was designed as a longitudinal observational cohort study for Alzheimer's disease biomarker identification, involving individuals of European ancestry. More details are found in Lovestone et al.'s paper, which shares excitement in the discovery of biomarkers for AD[64].

The 284 individuals were ≥65 years old and were classified into 1 of 3 health states at baseline, when blood was drawn. This resulted in 36 Alzheimer's disease (AD) blood samples, 109 amnestic mild cognitive impairment (MCI) blood samples, and 89 cognitively healthy blood samples[86]. AD diagnosis followed DSM-IV and NINCDS-ADRDA criteria[86]. MCI diagnosis was reserved for individuals who did not report significant daily life impediments according to Peterson's criteria, and received either a CDR score of 0.5, or a CDR memory score of 0.5 or 1[86]. Among these 109 MCI individuals, 38 MCI individuals transitioned to AD within 1 year post-baseline (MCI-AD), 67 individuals remained MCI stable (MCI-MCI), and 4 individuals transitioned to AD at an unclear, unknown time[86]. The time-of-transition ambiguity of the latter 4 individuals were advised to necessitate exclusion[86]. Batch effects were mitigated through thoughtful sample collection – blood samples were reportedly randomized by sex, center, and cognitive health diagnosis[86]. Psychiatric or neurological comorbidities were not present in these 284 individuals[86].

2.2 DNA Methylation Data

2.2.1 M Values vs B Values

Utilizing the GEOQuery R package for raw data download, and R minfi package's readGEORawFile function for parsing methylation data into accessible dataframes, both Beta values and M-values were obtained from GSE144858[4,24,65]. Illumina technology involves two probes working in tandem – one probe measures fluorescence intensity of methylated sites, while the other is for unmethylated sites[65]. The two measurements result in the final methylation measure. Beta values are popular, ranging from 0 to 1 due to their easy interpretability: % methylation[30]. M values have been recommended recently (although not as easy to directly interpret) because they help achieve homoscedasticity, often assumed when performing differential methylation analysis[30]. M values are essentially log-transformed beta values, with their exact formula shown in Table 1[30]. Offset values for M and B are by default 100, and are often in most contexts safely negligible[30]. Due to deriving differentially methylated positions in my protocol, I thus decided to use M values for any direct differential methylation analysis, and to transform M values back to Beta values for prediction modeling[13,27,42,48,83,110,120,121].



M	B	M -> B	B-> M
$\log_2(\frac{MAX(meth.site_x, 0) + 100}{MAX(unmeth.site_x, 0) + 100})$	$\frac{MAX(meth.site_x, 0)}{MAX(meth.site_x, 0)+ MAX(unmeth.site_x, 0) + 100}$	$\frac{2^{M_x}}{2^{M_x}+1}$	$\log_2(\frac{B_x}{1-B_x})$

Table 1: B and M value interconversions and equations

2.2.2 Data Quality Control

A key step when processing 450k array data is identification of cross-reactive probes: probes that ambiguously point to more than one location in the genome[65]. It is common practice to exclude such ambiguous markers, and thus using the xreactive_probes R function, such sites reported in literature by Chen et al. and Benton et al. were discarded[9,20,22]. Another key step is removing probes whose methylation value may be SNP-driven, thus leading to potentially false positive claims of CpG site association with disease status[112].

minfi's dropLociWithSnps function removed CpG loci directly overlapping SNPs, or being present 1 bp away from SNPs, with no mean allele frequency filter restriction[4,65]. Next, a quick check for data quality was conducted regarding detection p-values: denoting whether a CpG site is simply noise, or real signal[43,65]. The intuition is that as the observation gets more extreme than what would be expected simply due to random background signal, its p-value is smaller, denoting this signal is indeed real. No samples had mean detection p-value $\geq$0.05[88,113]. 1 probe was found to have median detection p-value $\geq$0.05 across samples, and thus discarded. Normalization was performed using minfi to prioritize a streamlined nature for data processing[4]. Stratified quantile normalization (taking into account Illumina probe types and CpG regional variation) was performed using minfi's preProcessQuantile function which also valuably notifies if there were any sex inconsistencies (there were none detected)[4]. While there is no consensus on the single best normalization method, preprocessQuantile is well-regarded in performance, and is recommended for single tissue-based data, which matches the characteristics of this dataset (whole blood tissue)[65,4]. Probes on sex chromosomes were discarded since this dataset consists of both males and females, as cross-reactive probes mapping to sex chromosomes have been shown to cause spurious associations, and gender may bias methylation patterns[44,68,79,88,113]. For the goal of predicting Alzheimer's disease prediction from mild cognitive impairment, samples were restricted to MCI individuals that either transitioned from AD over the course of 1 year post-baseline, or remained MCI stable.

After the above QC and data processing steps, this resulted in a 105 $\times$ 421290 matrix of M value methylation values. Beta values were retained for later interpretation and for downstream polygenic methylation score creation.

2.2.3 Cell-type Heterogeneity Adjustment

Whole blood samples in reality may be composed of several different cell types. Cell-type composition may be associated with health status, and DNA methylation patterns are cell-type specific, leading this to be a potential confounder in analysis[15,67]. We do not wish cell-type composition to drive the identification of CpG sites that are associated with

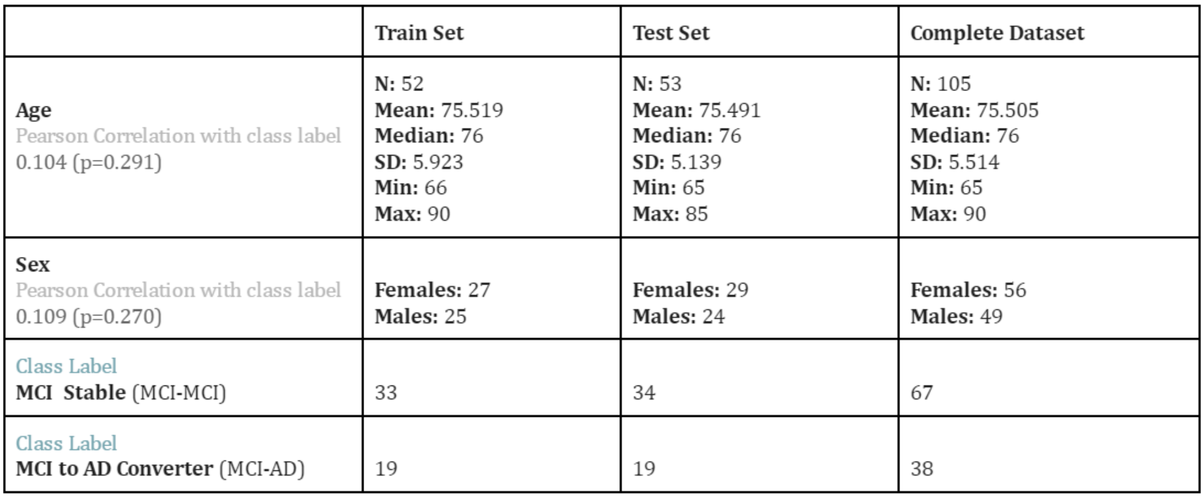
MCI to AD progression. Thus, it is essential to properly adjust for cell-type heterogeneity to prevent false positives. Adjustment can be either reference based, or reference-free[15,67]. Houseman et al.'s reference-based method was a popular method in early EWAS, where adjustment involved using the reference data to infer local data's cell type proportions[15]. But if the reference does not contain all the necessary cell types, or the current data has measurement errors, this will impede proper adjustment[15]. In fact, a lymphocyte subtype that is not present in this reference dataset may have led to false positive findings in past EWAS[15]. Thus, due to the limitations of reference-based cell type adjustment, reference-free was the chosen avenue for adjustment. Of the reference free-methods, surrogate variable analysis (SVA) is well-regarded and recommended[15,55,67]. The high-level intuition is that SVA assigns every CpG site a weight, which corresponds to how likely the site is influenced by latent variables[19]. This is useful to capture latent, unmodeled variation, and compartmentalize them into separate 'surrogate variables', that we can then adjust for[19]. SVA fails when strong cell-type composition confounding is present, as it does not converge[15,19]. SmartSVA is an improvement of SVA that solves this problem[15,19]. SmartSVA has excellent control of false positives, alongside high power, and the added bonus of being computationally faster compared to SVA[15,19].

2.3 Prediction Modeling using Internal Feature Selection

2.3.1 Train-Validation Split and K-Fold Cross Validation

A sample size of 105 is quite small when dealing with such high dimensional data (421290 features). Train-test splits can vary from 90:10 to 50:50, and different studies may employ different splits[47,92,96,98]. Due to the small sample size, to reserve enough data to comment on test set accuracy, a 50% - 50% train-validation split was employed, with stratification on sex and MCI conversion status[47]. A simple random split in cases of small sample sizes may lead to unstable estimates, and thus a methodology such as nested cross validation (nested CV) is recommended[96]. However, such a methodology results in different features that may be selected in each internal fold, and extremely small sample sizes for EWAS analyses may not result in proper identification of true positives[66]. Thus, a balance of

50:50 train-validation split was used to separate feature selection from model creation, and then 3-fold cross validation was utilized on the validation set to enable more stable estimates of model performance. Table 2 shows the train-validation split statistics, and that age nor sex have strong Pearson correlation with MCI conversion status. Supplementary Figure 1 provides a visual appraisal of age and sex distribution.



	Train Set	**Test Set**	**Complete Dataset**
Age Pearson Correlation with class label 0.104 (p=0.291)	**N:** 52 **Mean:** 75.519 **Median:** 76 **SD:** 5.923 **Min:** 66 **Max:** 90	**N:** 53 **Mean:** 75.491 **Median:** 76 **SD:** 5.139 **Min:** 65 **Max:** 85	**N:** 105 **Mean:** 75.505 **Median:** 76 **SD:** 5.514 **Min:** 65 **Max:** 90
Sex Pearson Correlation with class label 0.109 (p=0.270)	**Females:** 27 **Males:** 25	**Females:** 29 **Males:** 24	**Females:** 56 **Males:** 49
Class Label **MCI Stable** (MCI-MCI)	33	34	67
Class Label **MCI to AD Converter** (MCI-AD)	19	19	38

Table 2: Train-Validation split was conducted using a simple 50% - 50% split, stratified on the class label and sex.

2.3.2 Differential Methylation Analysis

Identifying CpG sites that have differing methylation levels corresponding to specific disease states, may have potential as actionable biomarkers. With this spirit, the first step of this analysis was to find differentially methylation positions (DMPs) and regions (DMRs), where the comparisons are between MCI-MCI individuals and MCI-AD individuals[65]. A popular R package for differential microarray analysis, limma, was utilized to identify DMPs[65,85]. Linear models (with the methylation M value as the outcome variable) were adjusted for sex, age, and surrogate variables identified by SmartSVA to address cell-proportion confounding and any hidden latent variables that may identify false positive CpG sites associated with MCI to AD conversion. To prevent overfitting, a small set of features were desired. Thus, a p-value cutoff of 5×10^{-4} was used, resulting in 15 DMPs. DMRcate identifies DMRs and carries the advantage of being data-agnostic and not constrained by the availability of annotations beyond the spatial level[76]. DMRcate is streamlined to depend on limma, and thus was utilized in this workflow to identify DMRs, with the same 5e-4 cutoff[76]. This resulted in 6 DMRs, which mapped to 35 CpG sites. No markers passed the FDR cutoff of 0.05, as observed in previous EWAS[60,80].

2.3.3 Prediction Models

5 prediction algorithms were identified as they each vary in factors of interpretability, simplicity, and linearity. These 5 models were applied both with and without SMOTE-Tomek adjustment. This meant a total of 10 models, separately applied to both DMP and DMR-derived features.

Decision Tree

First, the decision tree was chosen for its incredible advantage of transparency in decision making and ability to address nonlinearity. In brief, a decision tree involves recursive binary partitioning, and is in essence a hierarchical structure[21,50]. From a root node, binary decisions are decided for each internal node, until a stopping criterion is met[50]. At each node, since a binary decision is chosen, continuous variables will be reformulated to be dichotomous using thresholds, opening up an immense search space of possible cutoffs[50]. A good decision allows for homogeneity, such that each child node consists of samples that are solely MCI-stable, or MCI-AD[50]. This is reflected by minimizing an impurity score such as Gini impurity when an algorithm identifies what decision to use to create children nodes[50,21]. In this analysis, Scikit-learn's implementation was used[73]. A max depth of 5 chosen to prevent overfitting, with all other parameters kept as the default.

Random Forest

Next, the random forest was chosen as it carries the immense advantage of being robust to overfitting, relative to decision trees[28,90]. Random forests are a very popular ensemble method, and in short use base learners of uncorrelated decision trees[21]. To achieve this, bootstrapping is used to draw samples with replacement for each tree in the forest[21]. Within each individual tree, for every node, a random subset of variables is chosen to identify a binary decision function[21,90]. This allows the trees in the forest to be independent from each other, thus reducing variance which is desirable[21]. In the end the collection of trees form a consensus decision, and in the case of classification, is a simple majority vote[21]. Scikit-learn's implementation was once again used, with a max depth of 5 to form shallow trees to prevent overfitting[21,73,90].

Adaptive Boosting

AdaBoost is the most popular and widely-used boosting algorithm, which follows the intuition of improving on numerous sequential weak decision trees[102]. In essence an error function is employed that upweights in every iteration on erroneous classifications, thus allowing for gradual improvement[102]. One controls the number of iterations, and in scikit-learn's implementation 50 iterations were chosen (default) due to the small sample size, to prevent overfitting[73,108]. AdaBoost was also included as it has shown promising performance in imbalanced data situations[102].

Logistic Regression

Logistic regression was chosen for its power if a linear decision boundary is possible that separates MCI-stable and MCI-to-AD converters, and also for its transparency. Logistic regression enables the possibility of developing a polygenic methylation score, simply by taking the summation of the methylation values of biomarkers, multiplied by their respective coefficients converged upon by the algorithm[84]. Along this note, a second version of logistic regression was utilized, in separating individuals by the median methylation score. Individuals with scores above the median were predicted as MCI-to-AD converters, while those with below the median (and equal to) were predicted as MCI stable. This was inspired by past literature employing the use of the median value as a cutoff of individuals into "high-risk" and "low-risk" for a specific outcome variable[29,34,39]. Scikit-learn's implementation was used, ensuring regularization was not utilized, maintaining default parameters for all else[73].

SMOTE-Tomek to Address Class Imbalance

It is important to note that this dataset has a level of class imbalance – approximately 2:1 (67 MCI-stable vs 38 MCI-converters). In a small sample size this may result in prioritization of the majority class and thus unsatisfactory performance for the minority class: the individuals that transition from MCI to AD[7]. Therefore, SMOTE-Tomek was utilized as an oversampling method combined with data cleaning that prevents overfitting[7]. Specifically, SMOTE is a promising oversampling method that allows the minority class to also be fairly represented in the decision space[7]. To remove noisy examples lying on the wrong side of the decision boundary, Tomek links identify such problematic datapoints, to filter them out, allowing simpler models with better generalizability[7]. SMOTE-Tomek was applied using the imblearn Python package[56].

2.3.4 Model Performance Evaluation

Accuracy was not chosen as the optimizing criterion due to class imbalance. If accuracy were to be used, a model may simply decide on predicting all samples as MCI-stable, which may indeed result in high accuracy overall – yet this would result in abysmal performance for the minority class[7,8]. Therefore, as commonly done in practice, area under the Receiver Operating Characteristics Curve (referred to as auROC or simply AUC) was the performance metric optimized for all models, except for the Logistic Regression-derived median score based classification (does not utilize concept of probability threshold), which was optimized for balanced accuracy[7,8]. Balanced accuracy protects against bias for the majority class in imbalanced data scenarios, allows to evaluate whether a model is any better than random guessing (similar to the intuition that an AUC above 0.5 results in a more informed model than a random coin toss) and is the proper choice when choosing the best model across different data (in this case, defined by different feature sets)[23].

In the end, a total of 10 models were applied individually to DMP and DMRs, adjusting for age and sex.

2.4 Prediction Modeling using External Feature Selection

As is described in the results below, this initial iteration of modeling resulted in clear overfitting plausibly due to a small sample size used to extract features in such a high-dimensional dataset. To understand whether past literature findings may be utilized to overcome this, external feature selection was performed. Specifically two data sources were identified of promise – one derived from methylation data and another derived from gene expression data.

2.4.1 External Biomarker Prioritization

Methylation-Derived CpG sites

In 2021 baseline epigenetic signatures associated with MCI to AD conversion status were published by Li et al., utilizing ADNI 450k microarray data[60]. Biomarkers in the form of DMPs associated with MCI to AD conversion, DMPs overlapping between ADNI and the AddNeuroMed cohort, and finally DMRs for MCI to AD conversion respectively, were publically provided by the authors[60]. This collection of CpG sites were adapted for this analysis by ensuring sites were present in the local dataset and were mappable to genes for downstream interpretation as noted by the IlluminaHumanMethylation450kanno.ilmn12.hg19 R package[41]. This resulted in a total of 141 CpG sites. A table of these 141 CpG sites (a union of DMPs and DMRs) is hosted on Github here. Constraining to just DMP-specific features, this resulted in 31 CpG sites, hosted here. Constraining to just DMR-specific features, this resulted in 110 CpG sites, hosted here.

Gene-Expression Derived CpG sites

In 2021 Lee et al. identified AD-related genes in an extensively robust manner: integrating information from SNPs, transcripts, animal disease models of AD, and text mining from the ADNI and AddNeuroMed cohorts[54]. Integration of such wide-varied data is possible by convergent functional genomics (CFG)[54]. CFG scoring allows the ranking of biomarkers using varied sources of evidence[54]. A numeric point is assigned if a gene is identified as being supported by a specific line of evidence[54]. The authors highlighted genes with a score 3 and above to be informative for AD (max score being 7)[54]. This resulted in 439 differentially expressed genes that were well-supported by other lines of evidence, and mapped to corresponding 9535 CpG sites found in the local dataset used in this thesis. To narrow down these robust gene-derived features to a smaller pool, limma was run on the local dataset (105 samples), and CpG sites with $p<0.01$ were noted[85]. The intersection of these CpG sites were identified as biomarkers supported by multi-dimension external information, resulting in 64 CpG sites. A table of these 64 CpG sites is hosted on Github here.

2.4.2 Nested Cross Validation

In the case of small sample sizes, k-fold cross validation poses a challenge as increasing k increases variance[91]. Moreover, it can produce strongly biased estimates of test performance in small sample sizes[105]. Leave-one-out cross validation can still lead to overfitting in small samples[105]. In this scenario, nested cross validation enters as a strong alternative, as it prevent data leakage, allows for robust feature selection, and a robust estimate of model performance[91,105]. As the dataset consists of 105 samples, 5x3 nested cross validation was chosen. This allows an outer train fold size of 84, an outer test fold size of 21, an inner train fold size of 56, and an inner test fold size of 28. The inner loop is reserved for feature selection, and the outer loop is reserved for model performance evaluation.

As discussed earlier, 10 models were observed under this nested CV design: Decision Tree, Decision Tree+SMOTE-Tomek, Random Forest, Random Forest+SMOTE-Tomek,AdaBoost, AdaBoost+SMOTE-Tomek, Logistic Regression, Logistic Regression+SMOTE-Tomek, Logistic Regression Median Score, Logistic Regression Median Score + SMOTE-Tomek.

2.4.3 Feature Selection Algorithms

Within the inner folds, 3 feature selection (FS) algorithms (in the spirit of triangulation) were utilized, to identify 10 CpG features each from the pre-prioritized biomarker sets. These algorithms are briefly detailed below. The reasons for using 3 algorithms is that 'consensus' features may be extracted from the entire nested CV process, to identify high-priority methylation sites for future direction. This idea is inspired by Parvandeh et al.'s novel ideology of consensus nested CV, as well as Lima et al.'s comprehensive work encouraging covariate stability to be directly incorporated in protocols by not depending on a single FS algorithm[62,72]. To understand whether indeed triangulation offers the best feature set possible, or whether individual algorithms may offer optimal performance, the individual top 10 features identified by each of the 3 FS algorithms were also retained as candidate feature sets.

Minimum Redundancy Maximum Relevance

Maximum relevance is the most popular paradigm towards feature selection – simply it is the intuition of selecting features that matter the most with respect towards the outcome of interest[74]. In this perspective, there is to be some sort of metric quantifying importance – and in the case of MRMR's implementation, it can be mutual information or F-statistics[25,32,74,82]. However, this single layer of prioritization may result in a pool of features that have redundancy (interdependency). Thus, the next layer of filtering involves minimizing this repetitiveness[25,32,74,82]. The MRMR algorithm was introduced by Peng et al., and implemented with the mrmr_selection package written in Python, using the F-statistic for identifying relevant variables, and Pearson correlation to identify redundant variables[74?] .

Fisher Score

Fisher score follows the intuition of identifying features that allow discordant samples to be separated by a large distance, but distances within samples in of the same class to be small[78,106,117]. In short, this summarizes to minimizing within-class distance and maximizing between-class distances[78,106,117]. This is a simple filter approach and in practice, it is one of the most popular FS algorithms, and was implemented in this workflow using the scikit-feature repository[106,117].

L2-Regularization

L2-regularized logistic regression was utilized as it has been shown to be superior to L1-

regularized logistic regression when multicollinearity is present among variables[49]. Methylation is context-dependent, and may have inherent correlation structure across genomic regions (inspiring DMR identification)[40,57]. Thus this regularization method was chosen to prioritize top features, and implemented through Python's Scikit-learn package[73].

2.4.4 Model Performance Evaluation

In the end, the union set of the top 10 features found by the 3 algorithms across the inner folds were noted. Along side this, separate feature sets corresponding to each of the 3 FS algorithms were retained.

Ultimately, the nested CV design dedicates the inner folds for choosing the optimal feature set. In the case of gene-derived features, this resulted in 4 feature set options described above. In the case of methylation-derived features, the original DMP and DMR feature sets were individually included as well as potential feature sets of their own accord, resulting in a total of 6 feature set options.

As before, auROC was the performance metric optimized for all models, except for Logistic Regression-derived median score based classification, which was optimized for balanced accuracy instead. To select the final model, a holistic evaluation was made using sensitivity, specificity, and visual appraisal of confusion matrices. A highly sensitive model allows screening of individuals, by prioritizing individuals to undergo secondary neuropsychological assessments, neuroimaging, and specialized clinical assessment for AD. To the contrary, a highly specific model is useful in not being overzealous in overtreating MCI patients when it is unnecessary. The ideal scenario would be a model with both high sensitivity and specificity. If faced with the difficult dilemma of a tradeoff, the false negative carries more penalty in the AD context. Thus high sensitivity would be prioritized.

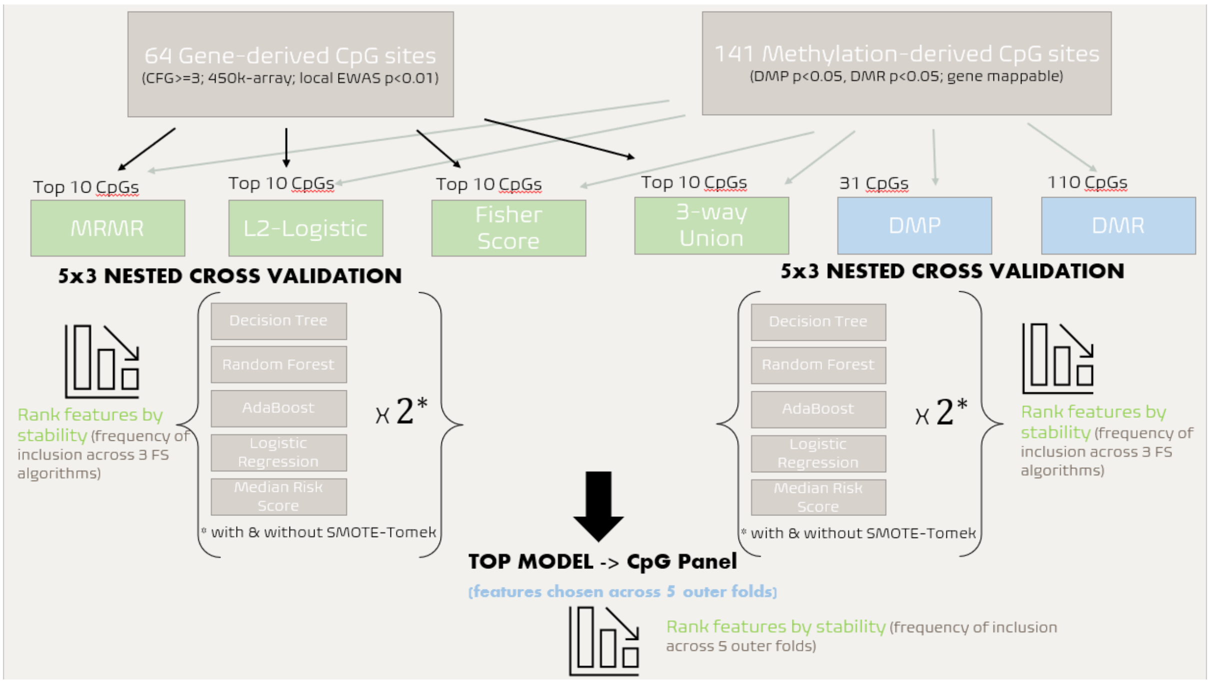

64 Gene-derived CpG sites
(CFG>=3; 450k-array; local EWAS p<0.01)
141 Methylation-derived CpG sites
(DMP p<0.05, DMR p<0.05; gene mappable)
Top 10 CpGs
MRMR
Top 10 CpGs
L2-Logistic
Top 10 CpGs
Fisher Score
Top 10 CpGs
3-way Union
31 CpGs
DMP
110 CpGs
DMR
5x3 NESTED CROSS VALIDATION
Decision Tree
Random Forest
AdaBoost
Logistic Regression
Median Risk Score
x 2*
* with & without SMOTE-Tomek
Rank features by stability (frequency of inclusion across 3 FS algorithms)
5x3 NESTED CROSS VALIDATION
Decision Tree
Random Forest
AdaBoost
Logistic Regression
Median Risk Score
x 2*
* with & without SMOTE-Tomek
Rank features by stability (frequency of inclusion across 3 FS algorithms)
TOP MODEL -> CpG Panel
(features chosen across 5 outer folds)
Rank features by stability (frequency of inclusion across 5 outer folds)

Figure 1: Overall Design of External Feature Prioritized Prediction Models

2.4.5 Biointeractions Analyses

Pathway Enrichment Analysis

The final biomarker panel was envisioned as the features chosen across the five outer folds in the nested CV process, ranked by their frequency of inclusion, to denote their stability and thus their importance[97]. Using the R package missMethyl, this collection of features was analyzed for the top enriched KEGG pathways to holistically interpret what information the panel reflects.[77].

Biomarker Prioritization

High priority methylation biomarkers were identified as the features that appear consistently across all 5 outer folds. Any such CpG sites were mapped to their genes, and queried at the Drug-Gene Interaction DataBase (DGIdb) to identify relationships with established therapeutic mechanisms, at AlzGene's CFG Rank resource to identify whether evidence suggests the genes are altered early in the AD continuum, and analyzed using CavityPlus, to assess druggability of the genes structurally[36,54,114,119,115]. This enables appraisal of top biomarkers for potential in drug design for therapeutic interventions.

3 Results

3.1 Prediction Modeling using Internal Feature Selection

For prediction modeling using internal feature selection with an initial 50-50 train-validation split, Figures 2 and 3 illustrate the mean AUC scores across the 3-fold CV procedure. The top train set performance for DMP-derived features was the Random Forest model, with mean AUC 0.951. The top test set performance for DMP-derived features was the AdaBoost model with mean AUC 0.633. The top train set performance for DMR-derived features was the Random Forest model, with mean AUC 0.796. The top test set performance for DMR-derived features was the Decision Tree model, with mean AUC 0.546. Clear overfitting is present as can be noted visually, and thus this was not chosen as the top model paradigm.

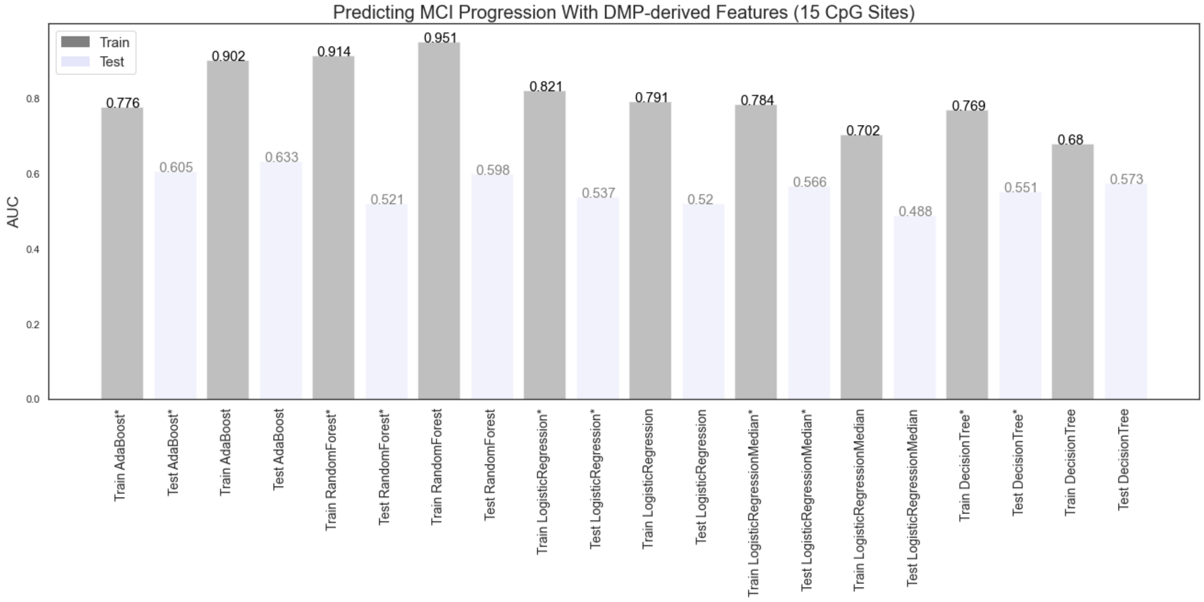


Figure 2: Internal Feature Selection (DMPs) Model Performance

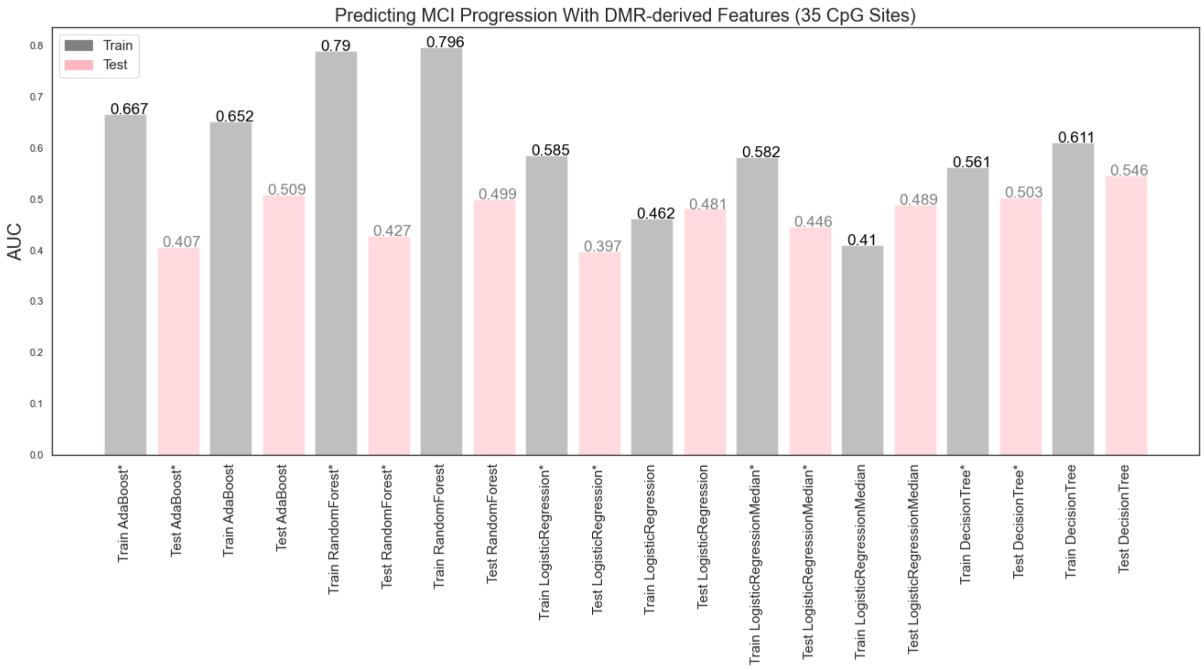
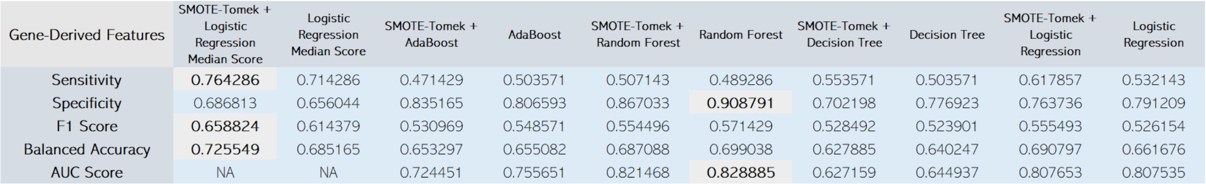


Figure 3: Internal Feature Selection (DMRs) Model Performance

3.2 Prediction Modeling Using External Feature Prioritization

For prediction modeling using external feature prioritization, and 5x3 nested CV for further feature selection, gene-derived features resulted in mean sensitivity score being highest for SMOTE-Tomek + Logistic Regression Median Score (0.764), mean specificity score being highest for Random Forest (0.909), mean F1 score being highest for SMOTE-Tomek + Logistic Regression Median Score (0.659), mean balanced accuracy being highest for SMOTE-Tomek + Logistic Regression Median Score (0.726), and mean AUC score being highest for Random Forest (0.829) (Figure 4). For methylation-derived features, mean sensitivity score was highest for SMOTE-Tomek + Logistic Regression Median Score (0.657), mean specificity was highest for Random Forest (0.942), mean F1 score was highest for SMOTE-Tomek + Random Forest (0.623), mean balanced accuracy was highest for Random Forest (0.719), and mean AUC score was highest for SMOTE-Tomek + Random Forest (0.827) (Supplementary Table 3).

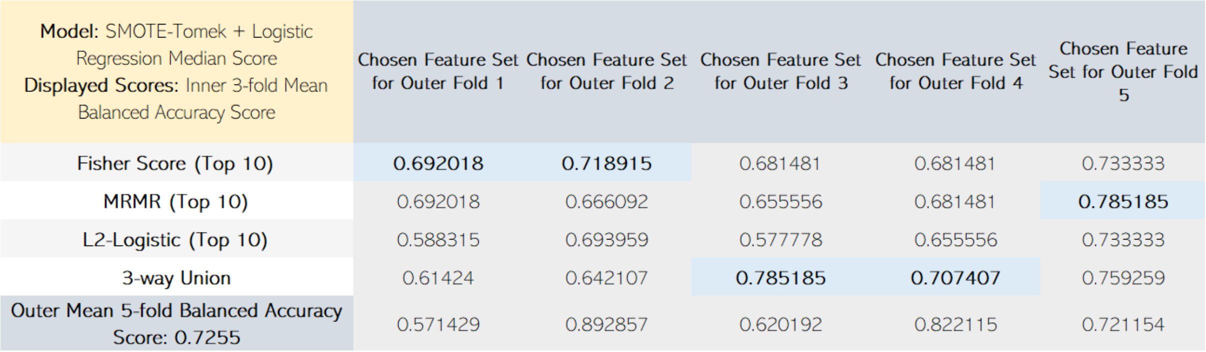
No model simultaneously was able to achieve above 0.7 sensitivity and specificity, so sensitivity was chosen as the metric to prioritize the final model, given AD's clinical context. Starting treatment intervention too late may result in no plausible benefit in modifying disease progression. Thus, gene-derived features and the SMOTE-Tomek + Logistic Regression Median Score model were retained to report the final methylation panel as well as top stable biomarkers. The confusion matrix for this final chosen model is provided in Supplementary Figure 5.



Gene-Derived Features	SMOTE-Tomek + Logistic Regression Median Score	Logistic Regression Median Score	SMOTE-Tomek + AdaBoost	AdaBoost	SMOTE-Tomek + Random Forest	Random Forest	SMOTE-Tomek + Decision Tree	Decision Tree	SMOTE-Tomek + Logistic Regression	Logistic Regression
Sensitivity	0.764286	0.714286	0.471429	0.503571	0.507143	0.489286	0.553571	0.503571	0.617857	0.532143
Specificity	0.686813	0.656044	0.835165	0.806593	0.867033	0.908791	0.702198	0.776923	0.763736	0.791209
F1 Score	0.658824	0.614379	0.530969	0.548571	0.554496	0.571429	0.528492	0.523901	0.555493	0.526154
Balanced Accuracy	0.725549	0.685165	0.653297	0.655082	0.687088	0.699038	0.627885	0.640247	0.690797	0.661676
AUC Score	NA	NA	0.724451	0.755651	0.821468	0.828885	0.627159	0.644937	0.807653	0.807535

Figure 4: Gene-derived Features 5x3 Nested CV Model Performance Comparison

Across the 5x3 nested CV scheme for this top chosen model, the following feature sets were chosen for each of the 5 outer folds based on inner fold optimization of Balanced Accuracy score: Fisher Score, Fisher Score, Union (3-way between Fisher, L2-Logistic, and MRMR), Union, and MRMR respectively shown in Figure 5. Consensus features across the 5x3 nested cross validation scheme ranked by frequency of inclusion (covariate stability) are shown in Figures 6 - 7.



Model: SMOTE-Tomek + Logistic Regression Median Score **Displayed Scores:** Inner 3-fold Mean Balanced Accuracy Score	Chosen Feature Set for Outer Fold 1	Chosen Feature Set for Outer Fold 2	Chosen Feature Set for Outer Fold 3	Chosen Feature Set for Outer Fold 4	Chosen Feature Set for Outer Fold 5
Fisher Score (Top 10)	0.692018	0.718915	0.681481	0.681481	0.733333
MRMR (Top 10)	0.692018	0.666092	0.655556	0.681481	0.785185
L2-Logistic (Top 10)	0.588315	0.693959	0.577778	0.655556	0.733333
3-way Union	0.61424	0.642107	0.785185	0.707407	0.759259
Outer Mean 5-fold Balanced Accuracy Score: 0.7255	0.571429	0.892857	0.620192	0.822115	0.721154

Figure 5: Gene-derived Feature Sets Chosen Across 5x3 Nested CV Derived From Top Model

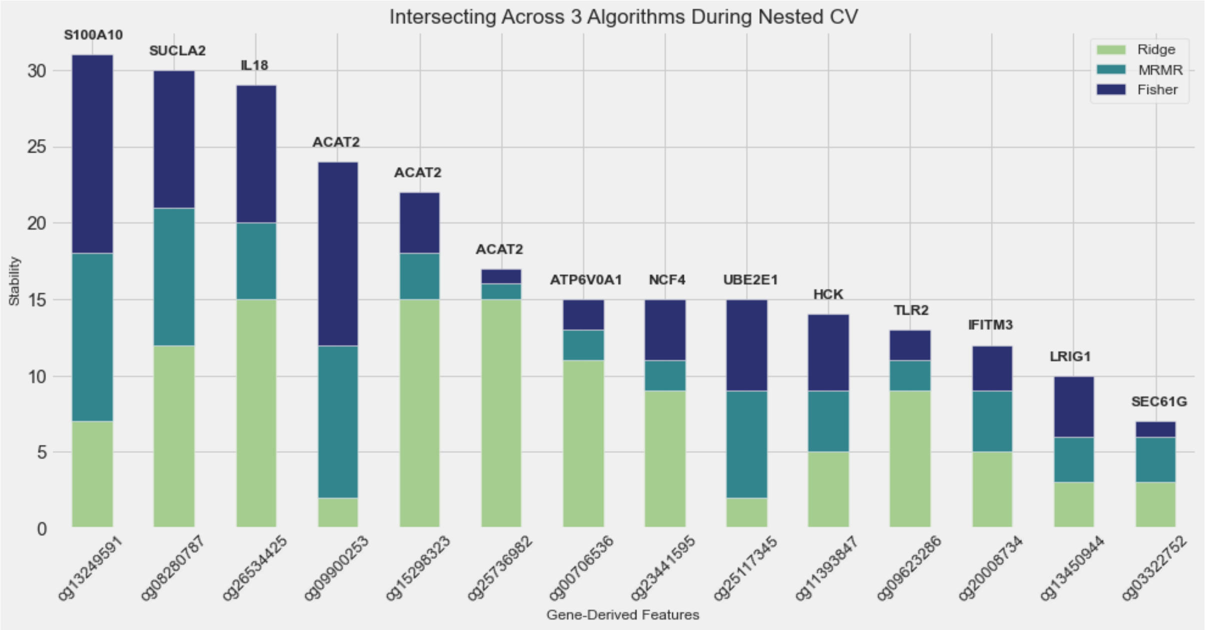


Figure 6: Covariate Stability for Gene-derived Features across 3 FS algorithms and 5x3 Nested CV

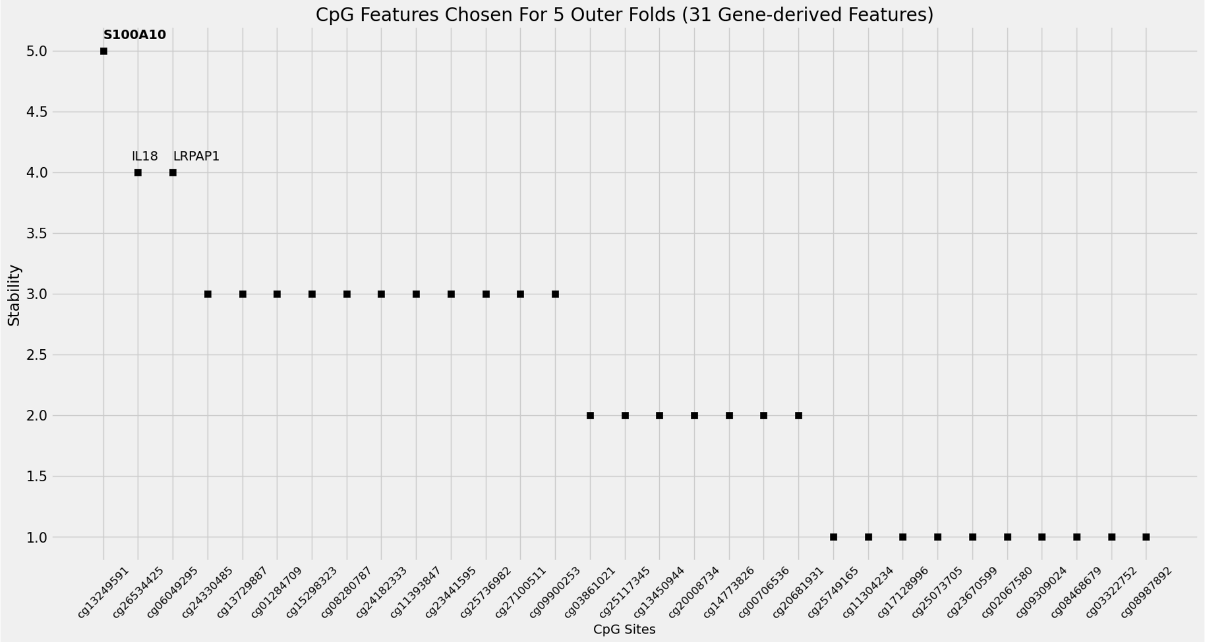
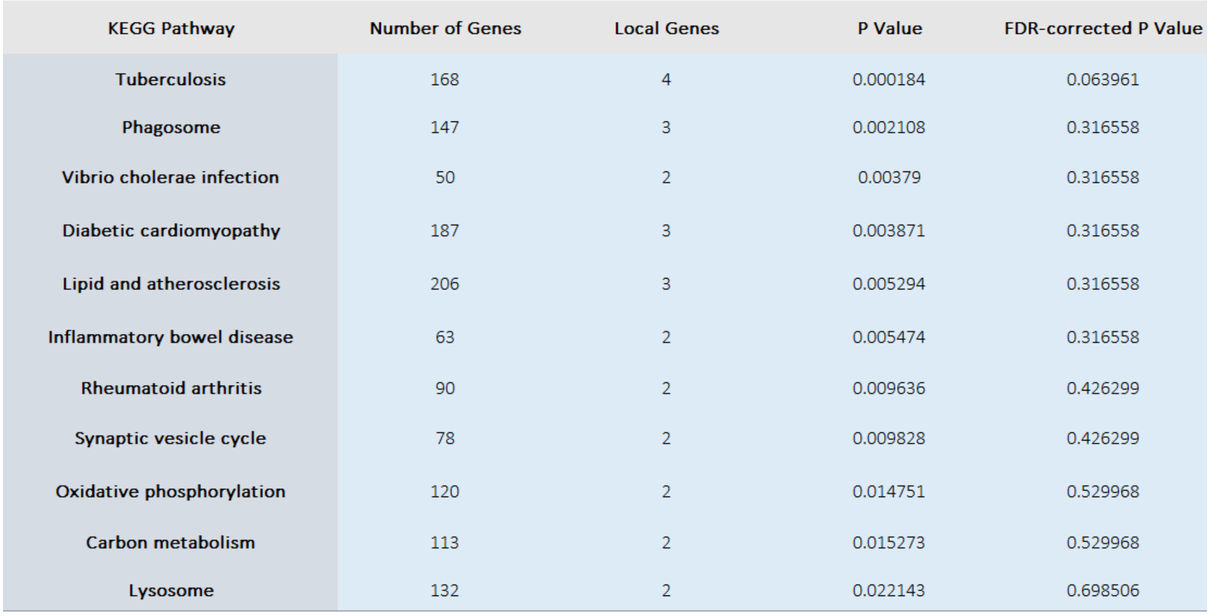


Figure 7: Gene-derived 31 CpG Panel Ranked by Covariate Stability across 5 Outer Folds Derived From Top Model

The final collection of CpGs resulted from these 5 feature sets was 31 CpG sites (Final Panel). The top biomarker resulted as a single methylation marker: cg13249591, mapping to the S100A10 gene. This is the highest ranked feature according to covariate stability without considering model performance, and also the most stable covariate consistently chosen across all 5 outer folds, which incorporates model performance.

3.3 Bioinformatic Analyses

KEGG pathway enrichment analysis did not result in pathway hits that passed the FDR threshold of 0.05. The top 10 pathways are shown in Figure 8 ranked by unadjusted p-value. From the local dataset, differential methylation analysis showed hypomethylation at the top marker associated with MCI to AD conversion, cg13249591 (Supplementary Table 4). Summary statistics from Lee et. al show increased gene expression of S100A10 with respect to AD[54] (Supplementary Table 4). This relationship of hypomethylation matched with increased gene expression is incredibly exciting, compounded by the fact that cg13249591 is found in the transcription start site (TSS200) and is annotated to be promoter associated (Final Panel). AlzGene's CFG rank resource identified S100A10 as an early differentially expressed gene, and correlation of its gene expression with tau line mouse models was reported to be statistically significant ($p<0.05$) with a coefficient of 0.599 (Supplementary Table 5). Querying S100A10 at the Drug Gene Interaction database highlighted an interaction with the clinical drug, dexamethasone. CavityPlus, run using the PDBe ID of 1a4p, labeled the top predicted cavity site with a strong druggability score and a predicted pKd of 6.99 (Supplementary Figure 8).



KEGG Pathway	Number of Genes	Local Genes	P Value	FDR-corrected P Value
Tuberculosis	168	4	0.000184	0.063961
Phagosome	147	3	0.002108	0.316558
Vibrio cholerae infection	50	2	0.00379	0.316558
Diabetic cardiomyopathy	187	3	0.003871	0.316558
Lipid and atherosclerosis	206	3	0.005294	0.316558
Inflammatory bowel disease	63	2	0.005474	0.316558
Rheumatoid arthritis	90	2	0.009636	0.426299
Synaptic vesicle cycle	78	2	0.009828	0.426299
Oxidative phosphorylation	120	2	0.014751	0.529968
Carbon metabolism	113	2	0.015273	0.529968
Lysosome	132	2	0.022143	0.698506

Figure 8: Top 10 KEGG pathways for 31 CpG Panel

4 Discussion

4.1 DNA Methylation Holds Promise

In essence, a highly-sensitive 31 CpG panel was identified to hold promise for predicting MCI individuals who will progress to Alzheimer's Disease within 1 year, through the design of a methylation risk score. Results show that there is justification of DNA methylation markers being incorporated in the prediction model paradigm for MCI to AD progression. As these prediction models only relied on age, sex, and DNA methylation markers, the potential of including APOE e4 allele dosage, transcriptomics, and metabolomics, (all-blood-based and non-invasive) sources of data, is promising. It is exciting that purely methylation data and 2 sources of demographic information is able to carry notable promise in prediction. However, it is imperative to validate these findings using external datasets.

4.2 Model Considerations

SMOTE-Tomek + Logistic Regression Median Score using gene-derived features was identified as the top model for highest sensitivity score. This suggests that identifying biomarkers of methylation can be candidate-gene driven opposed to solely epigenetic driven by past EWAS. Methylation is known to directly influence gene expression patterns, and thus this intuition holds plausibility, and in this analysis showed better performance (in sensitivity, balanced accuracy, AUC score, F1 score) compared to candidate markers identified by past EWAS.

One one hand, the finding that a methylation risk median score approach performs best for sensitivity, presents a challenge. The exact median score value may not generalize across datasets. Moreover, in the case of nested CV, multiple cutoffs will be used (1 for each outer fold model). For further direction, it would hold value to utilize multiple, larger scale datasets from different cohorts, thus making possible a clear train-test split for analyzing whether a generalizable methylation risk score cutoff can indeed be developed.

On the other hand, this finding illustrates that the problem of predicting individuals who transition to AD may be linearly separable, justifying the use of simplistic, more transparent models opposed to deep learning (while also ensuring underfitting does not result.) This problem of underfitting is illustrated as for example across the analyses, the Decision Tree showed low performance, while Random Forest, AdaBoost, and Logistic Regression based models had consistently higher learning capability.

What is interesting is that AUC score appears to prioritize the majority class, even in a class imbalance scenario that is not extreme (1:2 imbalance ratio). If one took the AUC score to reflect model performance, this would result in skewed optimism. This is directly illustrated how in this analysis the top performing model optimizing for AUC score was the Random Forest model with gene-driven features (0.8289), a nontrivial score suggesting superior performance if taken out of context. However, on closer look at sensitivity, and specificity, along with the visual appraisal of the confusion matrix, it is clear that AUC score has prioritized the majority class performance (Supplementary Figure 9). This may likely be due to the small sample size present in this analysis, but also may be because AUC may not be the best choice for optimization, which challenges wide-varied use in prediction schemes. For future direction, it would be incredibly interesting to compare whether area under the Precision-Recall curve or the new idea of a concordant partial area under the ROC curve may suffice to better prioritize the minority class without expense of the majority class, and vice versa[87,16].

Further direction also points to stacking ensemble methods that utilize base learners that can be a mix of models, with a higher level meta-learner[18,17,10]. This would allow for example utilizing AdaBoost, Random Forest, and Logistic Regression as base learners, incorporating each model's predictions. This methodology is becoming more popular, showing promise with their high performance, increase in generalizability, and reduced overfitting[18,10].

4.3 S100A10

One of the significant findings of this analysis was that S100A10 was consistently chosen across the the nested CV paradigm. It is vital to understand why this finding of S100A10 as a potentially high priority biomarker is exciting in the AD context – specifically in the space of hypotheses implicating systemic inflammation as the impetus or accelerator of AD neuropathology.

Plasminogen is a zymogen[99]. Once converted to its active form, a protease termed plasmin, this can encourage pathogen spread by the degradation of vascular barriers[99]. This connects with Alzheimer's as opportunistic reactivation of pathogens have been implicated in hypotheses surrounding infection-driven causal factors for AD progression[37]. C.neoformans is one such pathogen that also has the ability to penetrate the blood brain barrier[99]. Specific pathogens are increasingly being implicated in AD progression not limited to C. neoformans, including C. albicans and HSV-1[37,99,93,71,118]. Pathogens able to bind to plasminogen may manipulate processes involved with plasmin conversion.[99]. In short, pathogen invasion can be enhanced by plasminogen to plasmin conversion, and it is with high importance that this is emphasized as today there is increasing evidence of the role of infection in the Alzheimer's disease progression continuum.

Where does S100A10 fit in this pathogen story? Well, S100A10 is known to be found in a complex with the Annexin A2 peptide, and increases plasminogen production[5,104]. Annexin A2 has been reported to be utilized by viruses for endocytosis and S100A10 may help for pathogen transcytosis[104,33]. These connections are vital to note as peripheral inflammation due to pathogen infection can cause blood brain barrier dysfunction, which is thought to be an early sign of cognitive dysfunction and may be related with AD pathology[103]. Further, there is a mechanistic link between the periphery and the central nervous system as plasminogen can be transported to the brain from the peripheral blood circulation[103].

Beyond theory, can we effectively act on this supposed connection between systemic inflammation and neuroinflammation, that may indeed be an early player for AD progression? A first line of evidence stems from a recent publication that found that a combination of dexamethasone and acyclovir (well established clinical drugs) can effectively prevent amyloid oligomer-induced cognitive impairment and neuroinflammatton in AD mouse models[46]. Dexamethasone was promisingly also the singular interaction that the Drug-Gene Interaction Database found for S100A10[36]. In terms of safety, it is imperative to note that long term treatment of glucocorticoids (such as dexamethasone) can lead to neurotoxicity plausibly explained by increased quinolinic acid levels[46]. However, ayclovir counteracts against this, and thus may directly protect against this unwanted consequence[46]. The second line of evidence comes with the fact that S100A10 is known to be inducible by numerous transcription factors, providing hope for even methylation-based therapeutics, that can control gene silencing and reverse upregulation that may be harmful[100]. Lastly, the final exciting evidence of recent years is the discovery that blood-derived plasminogen regulates brain inflammation[6]. Astonishingly, in a mouse model, lower levels of blood plasminogen were seen to improve neuroinflammation, even decreasing microglial cell activation[6].

In terms of potential drug design, the CAVITY algorithm showed promise that S100A10 indeed may contain a cavity with strong possibility as a binding site for a small metabolite, drug molecule, or a macromolecule[119]. Although KEGG pathway enrichment analysis did not present results with FDR-corrected $p\text{-value}<0.05$, the top pathways identified tuberculosis (Figure 8). The fact that the top pathway implicates a virus is promising given the above results, compounded by the fact that Tuberculosis infection has been found to be associated with AD risk[75].

This evidence suggests that it is worthwhile to further elucidate the intersection of peripheral blood S100A10, viral and bacterial infection, and neuroinflammation, as it may possibly enable early intervention and prediction of Alzheimer's Disease for MCI individuals.

5 Conclusion

In conclusion, blood-based methylation data holds promise for Alzheimer's disease progression prediction, and recommends inclusion in non-invasive biomarker panels. With respect to modeling decision making, feature selection for methylation markers can be derived from candidate genes, and not be confined to solely epigenetic sources. The sensitivity-specificity trade-off is a real challenge for prediction modeling, and thus it would be fruitful to identify whether stacking ensembles that allow the integration of multiple varied base learners improve model performance. Class imbalance may present an issue in small datasets, and thus it is important to be holistic in the appraisal of model performance, and perhaps turn to alternative measures opposed to auROC to prevent skew towards the majority class. Lastly, specific to this thesis work, S100A10 was identified from methylation-based models to be a highly promising gene candidate that warrants further investigation in the role of neuroinflammation, infection, and Alzheimer's Disease.

6 Supplementary Materials

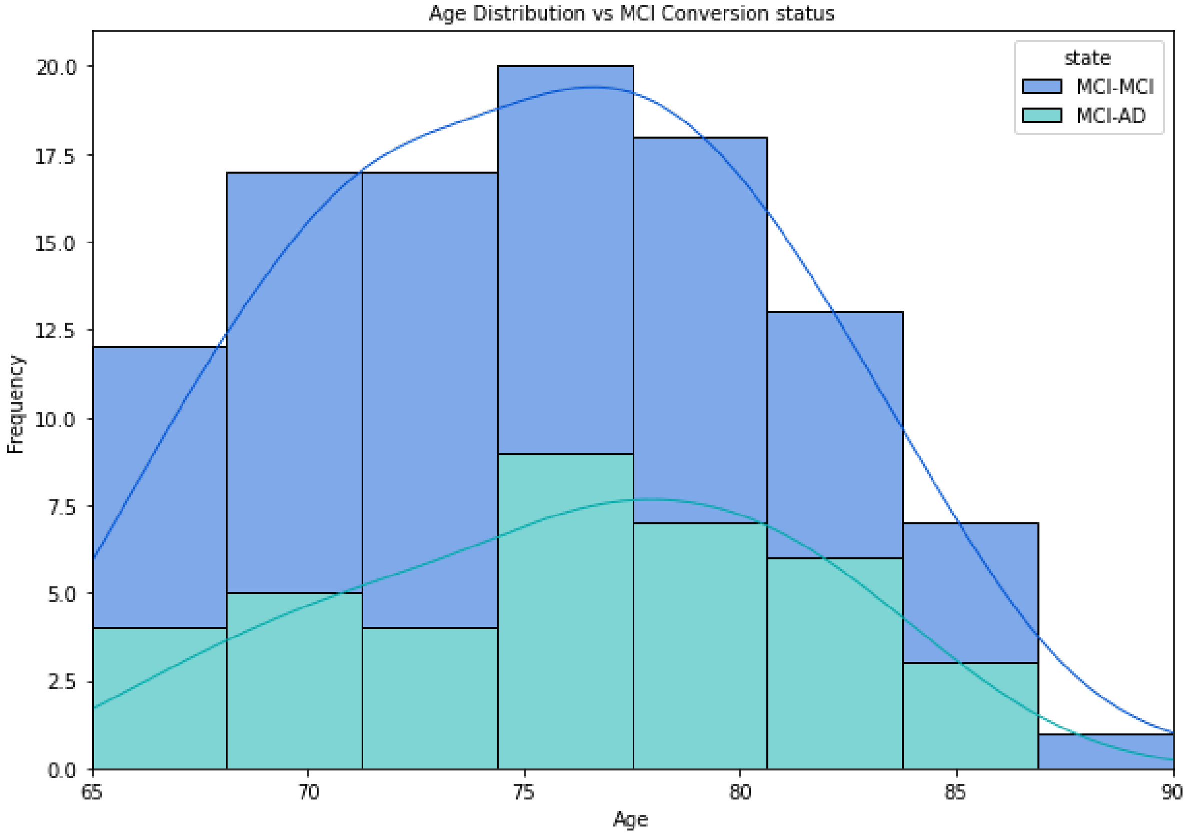


Figure 1: No clear pattern appears to exist for age between MCI-stable and MCI-to-AD converters.

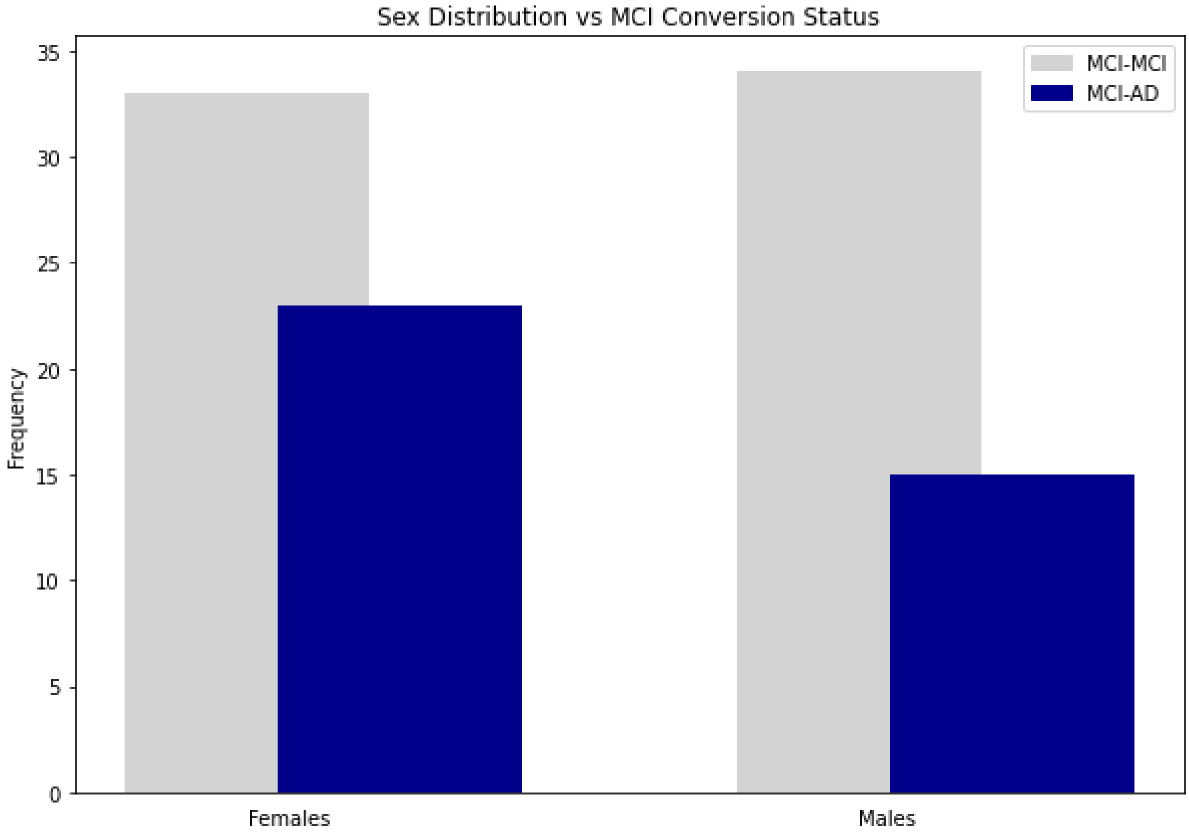
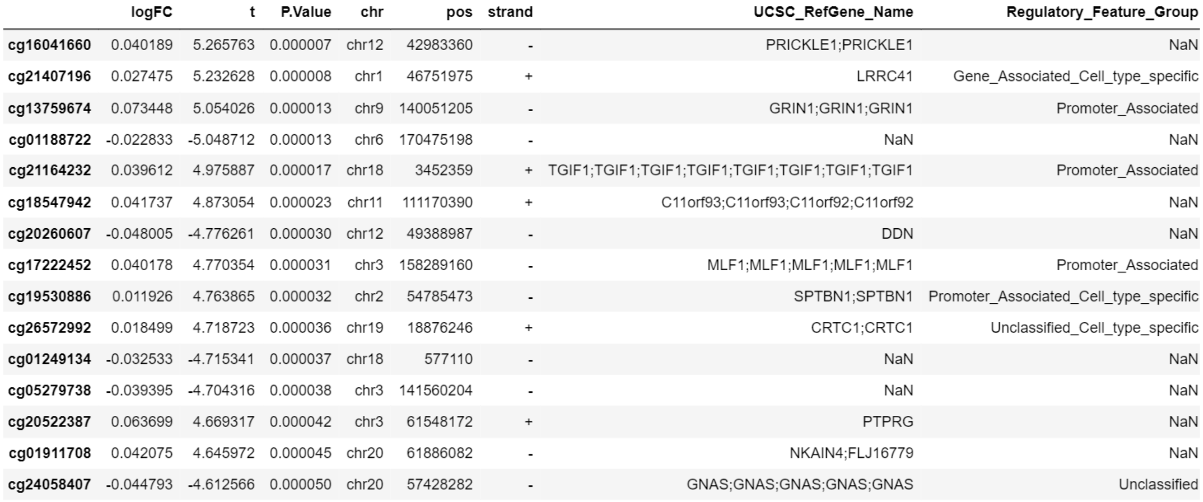
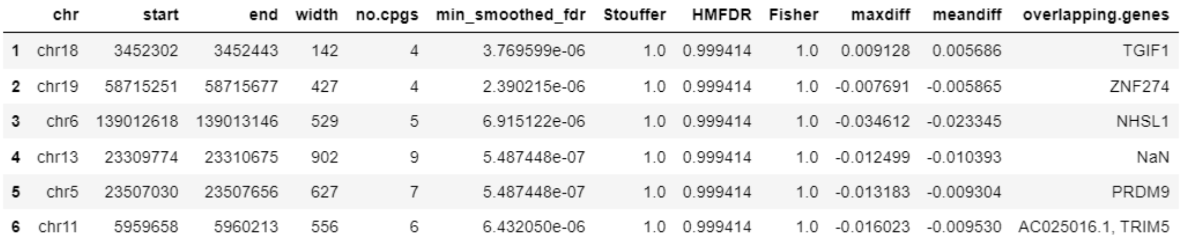


Figure 2: Both males and females are present across the class label distribution.



	logFC	t	P.Value	chr	pos	strand	UCSC_RefGene_Name	Regulatory_Feature_Group
cg16041660	0.040189	5.265763	0.000007	chr12	42983360	-	PRICKLE1;PRICKLE1	NaN
cg21407196	0.027475	5.232628	0.000008	chr1	46751975	+	LRRC41	Gene_Associated_Cell_type_specific
cg13759674	0.073448	5.054026	0.000013	chr9	140051205	-	GRIN1;GRIN1;GRIN1	Promoter_Associated
cg01188722	-0.022833	-5.048712	0.000013	chr6	170475198	-	NaN	NaN
cg21164232	0.039612	4.975887	0.000017	chr18	3452359	+	TGIF1;TGIF1;TGIF1;TGIF1;TGIF1;TGIF1;TGIF1;TGIF1	Promoter_Associated
cg18547942	0.041737	4.873054	0.000023	chr11	111170390	+	C11orf93;C11orf93;C11orf92;C11orf92	NaN
cg20260607	-0.048005	-4.776261	0.000030	chr12	49388987	-	DDN	NaN
cg17222452	0.040178	4.770354	0.000031	chr3	158289160	-	MLF1;MLF1;MLF1;MLF1;MLF1	Promoter_Associated
cg19530886	0.011926	4.763865	0.000032	chr2	54785473	-	SPTBN1;SPTBN1	Promoter_Associated_Cell_type_specific
cg26572992	0.018499	4.718723	0.000036	chr19	18876246	+	CRTC1;CRTC1	Unclassified_Cell_type_specific
cg01249134	-0.032533	-4.715341	0.000037	chr18	577110	-	NaN	NaN
cg05279738	-0.039395	-4.704316	0.000038	chr3	141560204	-	NaN	NaN
cg20522387	0.063699	4.669317	0.000042	chr3	61548172	+	PTPRG	NaN
cg01911708	0.042075	4.645972	0.000045	chr20	61886082	-	NKAIN4;FLJ16779	NaN
cg24058407	-0.044793	-4.612566	0.000050	chr20	57428282	-	GNAS;GNAS;GNAS;GNAS;GNAS	Unclassified

Table 1: Internal Feature Selection Identified DMPs



	chr	start	end	width	no.cpgs	min_smoothed_fdr	Stouffer	HMFDR	Fisher	maxdiff	meandiff	overlapping.genes
1	chr18	3452302	3452443	142	4	3.769599e-06	1.0	0.999414	1.0	0.009128	0.005686	TGIF1
2	chr19	58715251	58715677	427	4	2.390215e-06	1.0	0.999414	1.0	-0.007691	-0.005865	ZNF274
3	chr6	139012618	139013146	529	5	6.915122e-06	1.0	0.999414	1.0	-0.034612	-0.023345	NHSL1
4	chr13	23309774	23310675	902	9	5.487448e-07	1.0	0.999414	1.0	-0.012499	-0.010393	NaN
5	chr5	23507030	23507656	627	7	5.487448e-07	1.0	0.999414	1.0	-0.013183	-0.009304	PRDM9
6	chr11	5959658	5960213	556	6	6.432050e-06	1.0	0.999414	1.0	-0.016023	-0.009530	AC025016.1, TRIM5

Table 2: Internal Feature Selection Identified DMRs

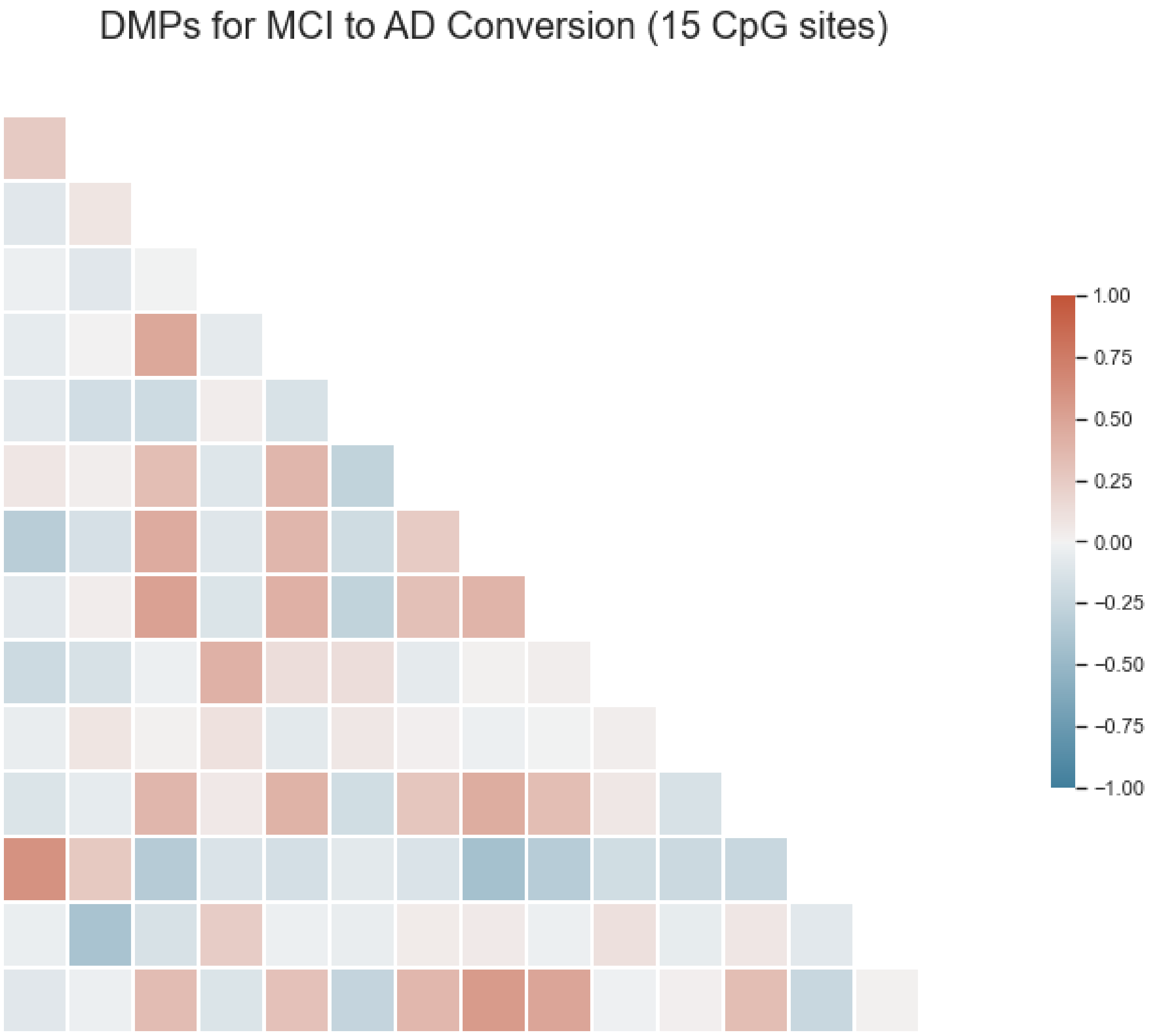


Figure 3: Correlation structure of Internal Feature Selection Identified DMP-derived CpG sites.

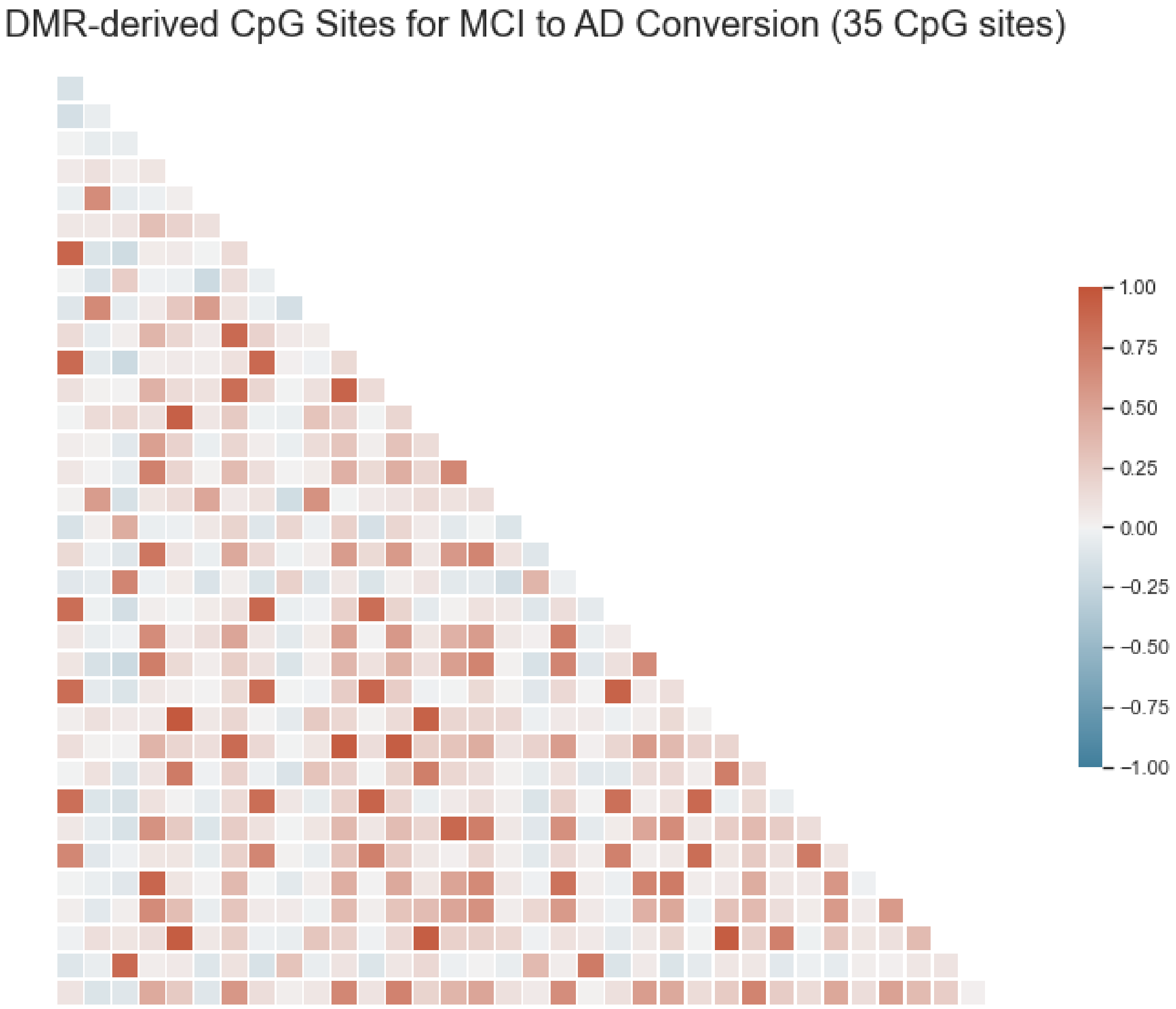


Figure 4: Correlation structure of Internal Feature Selection Identified DMR-derived CpG sites.

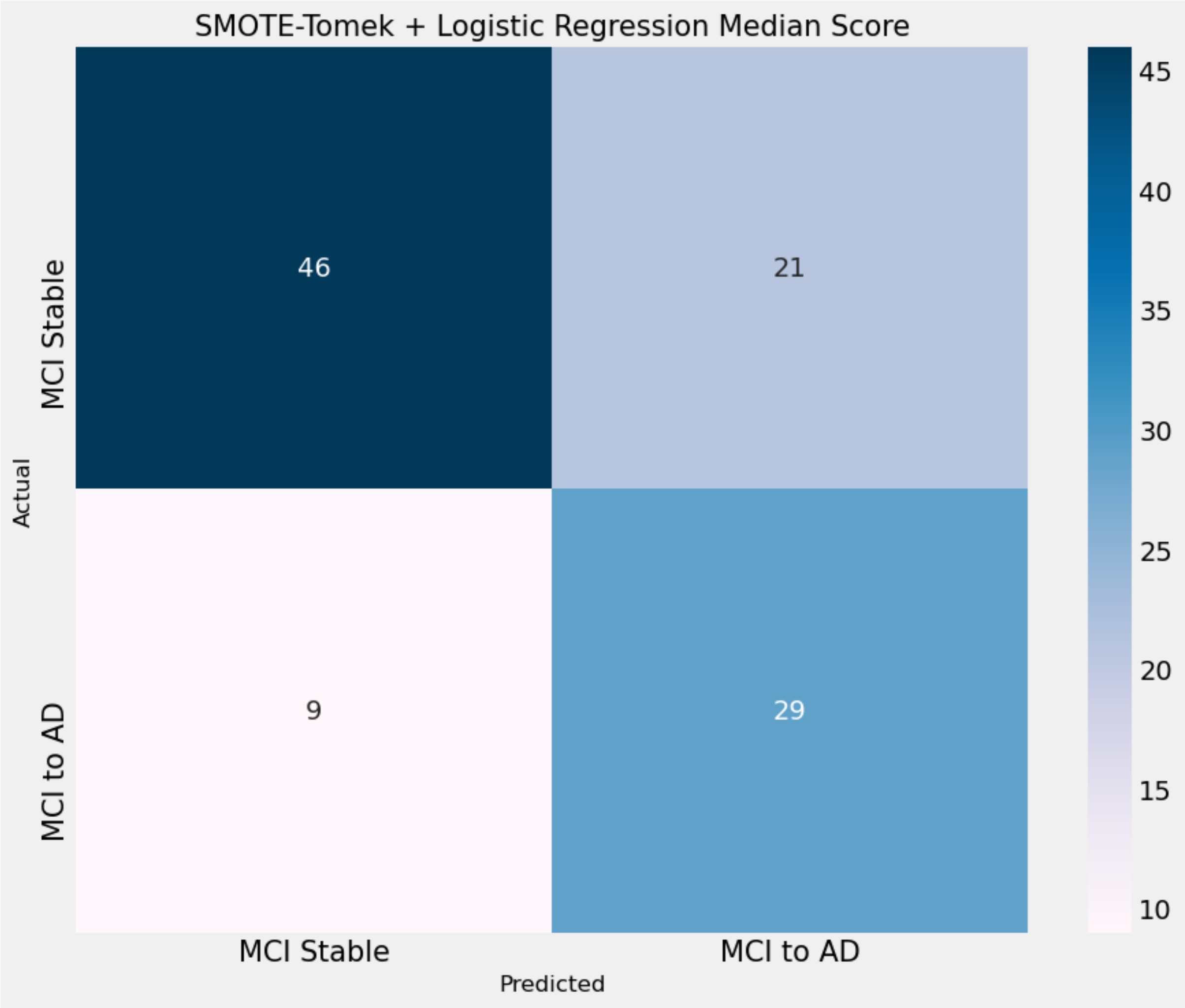


Figure 5: Confusion Matrix for Top Model (Gene-driven features; SMOTE-Tomek + Logistic Regression Median Score). A summation of TP, FP, TN, FN across 5 outer folds was taken.

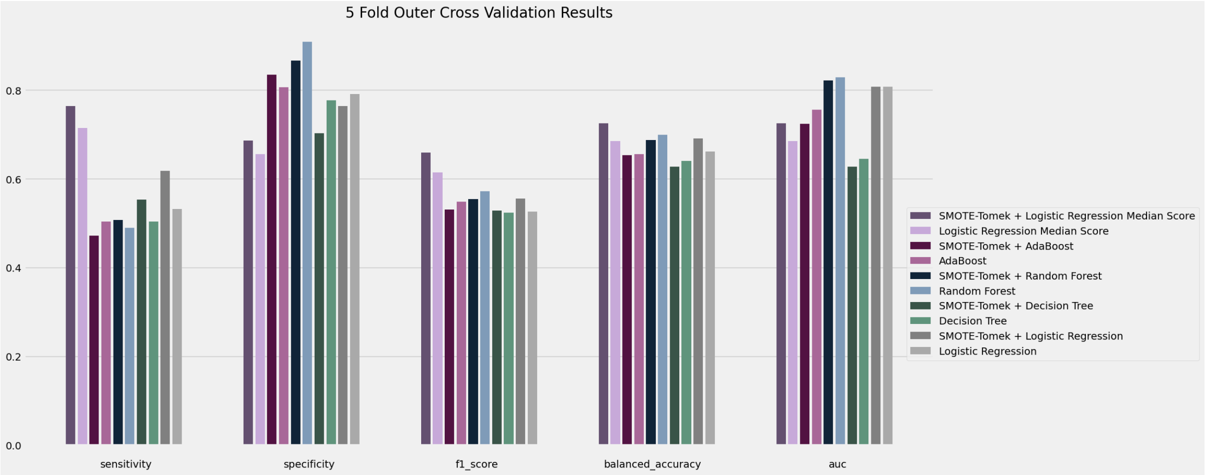
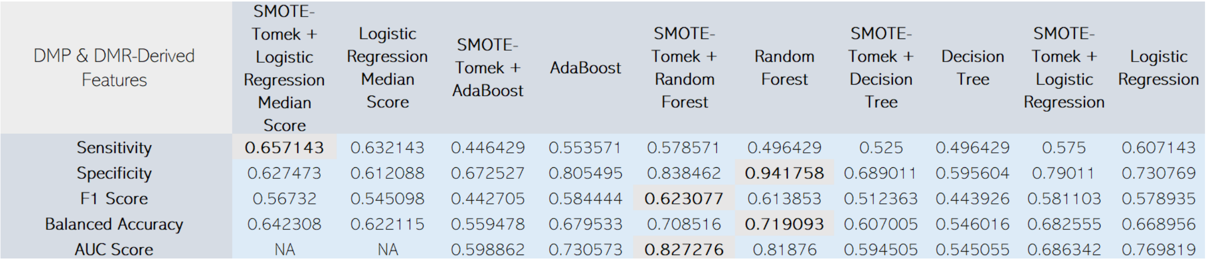


Figure 6: Gene-derived Features 5x3 Nested CV Visual Model Performance Comparison



DMP & DMR-Derived Features	SMOTE-Tomek + Logistic Regression Median Score	Logistic Regression Median Score	SMOTE-Tomek + AdaBoost	AdaBoost	SMOTE-Tomek + Random Forest	Random Forest	SMOTE-Tomek + Decision Tree	Decision Tree	SMOTE-Tomek + Logistic Regression	Logistic Regression
Sensitivity	0.657143	0.632143	0.446429	0.553571	0.578571	0.496429	0.525	0.496429	0.575	0.607143
Specificity	0.627473	0.612088	0.672527	0.805495	0.838462	0.941758	0.689011	0.595604	0.79011	0.730769
F1 Score	0.56732	0.545098	0.442705	0.584444	0.623077	0.613853	0.512363	0.443926	0.581103	0.578935
Balanced Accuracy	0.642308	0.622115	0.559478	0.679533	0.708516	0.719093	0.607005	0.546016	0.682555	0.668956
AUC Score	NA	NA	0.598862	0.730573	0.827276	0.81876	0.594505	0.545055	0.686342	0.769819

Table 3: Methylation Derived Features 5x3 Nested CV Model Performance Comparison

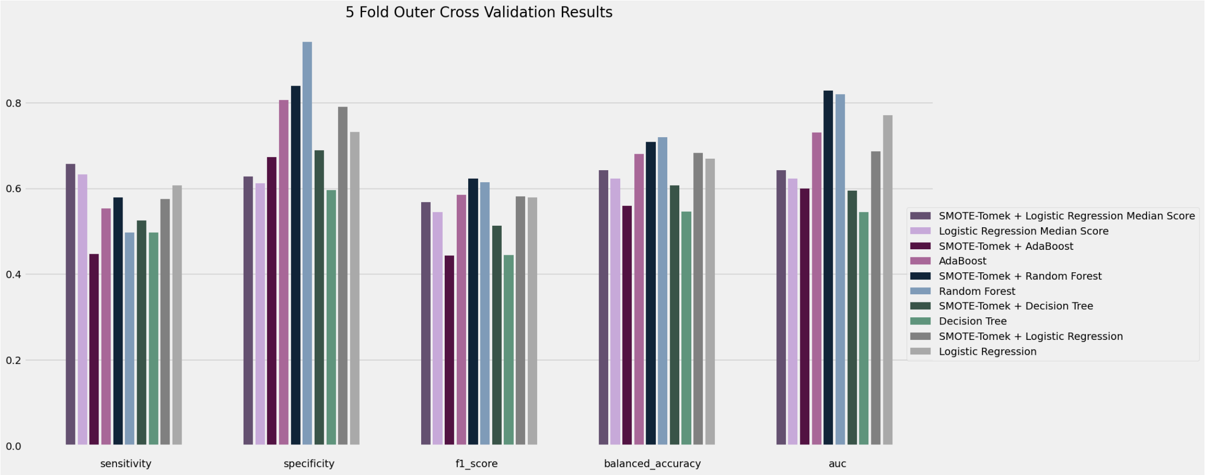
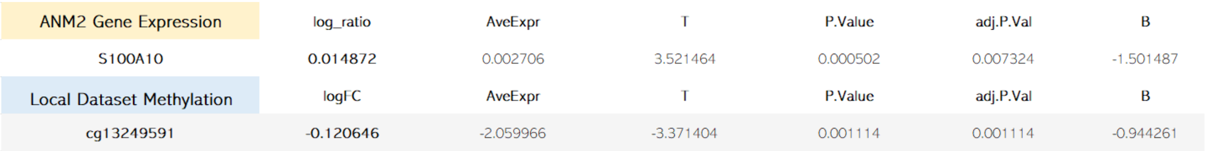
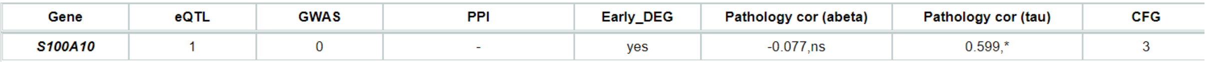


Figure 7: Methylation Derived Features 5x3 Nested CV Visual Model Performance Comparison



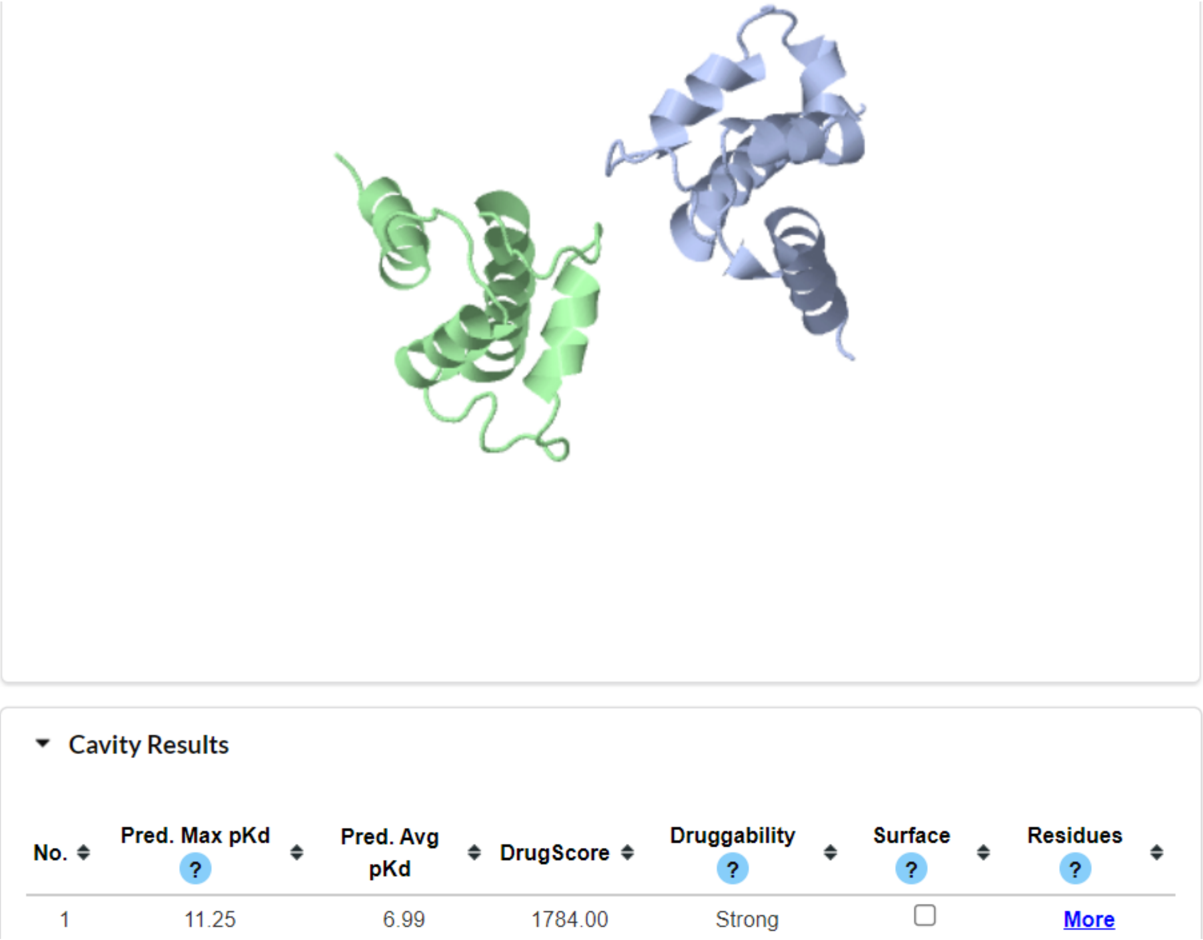
ANM2 Gene Expression	log_ratio	AveExpr	T	P.Value	adj.P.Val	B
S100A10	0.014872	0.002706	3.521464	0.000502	0.007324	-1.501487
Local Dataset Methylation	logFC	AveExpr	T	P.Value	adj.P.Val	B
cg13249591	-0.120646	-2.059966	-3.371404	0.001114	0.001114	-0.944261

Table 4: Comparing Methylation and Gene Expression. This is not a perfect correlation example as Lee et. al did not specifically investigate MCI to AD conversion. However, in the context of understanding AD as a continuum, this may provide value.



Gene	eQTL	GWAS	PPI	Early_DEG	Pathology cor (abeta)	Pathology cor (tau)	CFG
S100A10	1	0	-	yes	-0.077,ns	0.599,*	3

Table 5: CFG rank results from AlzGene



▾ Cavity Results

No.	Pred. Max pKd	Pred. Avg pKd	DrugScore	Druggability	Surface	Residues
1	11.25	6.99	1784.00	Strong	□	More

Figure 8: CavityPlus Denotes Potential Highly druggable cavity on S100A10 protein. For this analysis, the PDBe id 1a4p used, for both chains A and B.

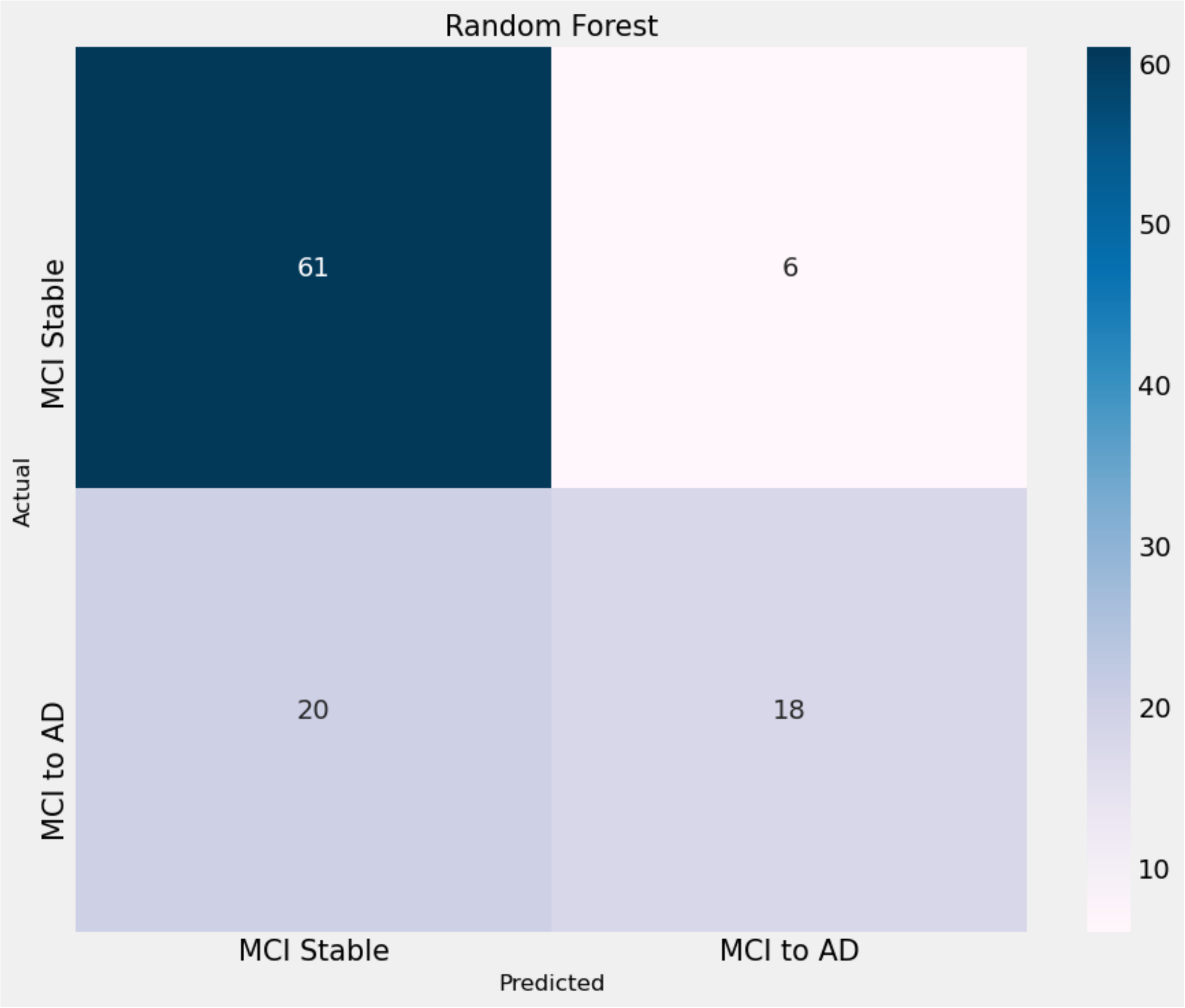


Figure 9: If Top Model had been simply chosen with AUC (Gene-driven features; Random Forest Model), would have resulted in skew towards majority class. A summation of TP, FP, TN, FN across 5 outer folds was taken.

Github-hosted tables, including the Final 31 CpG Panel are found here, as mentioned in the paper.

References

[1] Giulia Abate, Marika Vezzoli, Letizia Polito, Antonio Guaita, Diego Albani, Moira Marizzoni, Emirena Garrafa, Alessandra Marengoni, Gianluigi Forloni, Giovanni B. Frisoni, and et al. A conformation variant of p53 combined with machine learning identifies alzheimer disease in preclinical and prodromal stages. *Journal of Personalized Medicine*, 11(1):14, 2020.

[2] M. Altuna-Azkargorta and M. Mendioroz-Iriarte. Blood biomarkers in alzheimer's disease. *Neurología (English Edition)*, 36(9):704–710, 2021.

[3] Matthias Arnold, Kwangsik Nho, Alexandra Kueider-Paisley, Tyler Massaro, Kevin Huynh, Barbara Brauner, Siamak MahmoudianDehkordi, Gregory Louie, M. Arthur Moseley, J. Will Thompson, and et al. Sex and apoe 4 genotype modify the alzheimer's disease serum metabolome. *Nature Communications*, 11(1), 2020.

[4] Martin J. Aryee, Andrew E. Jaffe, Hector Corrada-Bravo, Christine Ladd-Acosta, Andrew P. Feinberg, Kasper D. Hansen, and Rafael A. Irizarry. Minfi: A flexible and comprehensive bioconductor package for the analysis of infinium dna methylation microarrays. *Bioinformatics*, 30(10):1363–1369, 2014.

[5] Zahra Bagheri-Hosseinabadi, Mohadese Abbasi, Mahmood Kahnooji, Zainab Ghorbani, and Mitra Abbasifard. The prognostic value of s100a calcium binding protein family members in predicting severe forms of covid-19. *Inflammation Research*, 71(3):369–376, 2022.

[6] Sarah K. Baker, Zu-Lin Chen, Erin H. Norris, Alexey S. Revenko, A. Robert MacLeod, and Sidney Strickland. Blood-derived plasminogen drives brain inflammation and plaque deposition in a mouse model of alzheimer's disease. *Proceedings of the National Academy of Sciences*, 115(41), 2018.

[7] Gustavo E. Batista, Ronaldo C. Prati, and Maria Carolina Monard. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*, 6(1):20–29, 2004.

[8] Juan Felipe Beltrán, Brandon Malik Wahba, Nicole Hose, Dennis Shasha, and Richard P. Kline. Inexpensive, non-invasive biomarkers predict alzheimer transition using machine learning analysis of the alzheimer's disease neuroimaging (adni) database. *PLOS ONE*, 15(7), 2020.

[9] Miles C Benton, Alice Johnstone, David Eccles, Brennan Harmon, Mark T Hayes, Rod A Lea, Lyn Griffiths, Eric P Hoffman, Richard S Stubbs, Donia Macartney-

Coxson, and et al. An analysis of dna methylation in human adipose tissue reveals differential modification of obesity genes before and after gastric bypass and weight loss. *Genome Biology*, 16(1), 2015.

[10] Saroj Kumar Biswas, Arpita Nath Boruah, Rajib Saha, Ravi Shankar Raj, Manomita Chakraborty, and Monali Bordoloi. Early detection of parkinson disease using stacking ensemble method. *Computer Methods in Biomechanics and Biomedical Engineering*, page 1–13, 2022.

[11] Jonathan Blackman, Marta Swirski, James Clynes, Sam Harding, Yue Leng, and Elizabeth Coulthard. Pharmacological and non[U+2010]pharmacological interventions to enhance sleep in mild cognitive impairment and mild alzheimer's disease: A systematic review. *Journal of Sleep Research*, 30(4), 2020.

[12] K. Blennow and H. Zetterberg. Biomarkers for alzheimer's disease: Current status and prospects for the future. *Journal of Internal Medicine*, 284(6):643–663, 2018.

[13] Mariana Boroni, Alessandra Zonari, Carolina Reis de Oliveira, Kallie Alkatib, Edgar Andres Ochoa Cruz, Lear E. Brace, and Juliana Lott de Carvalho. Highly accurate skin-specific methylome analysis algorithm as a platform to screen and validate therapeutics for healthy aging. *Clinical Epigenetics*, 12(1), 2020.

[14] Johannes Brägelmann and Justo Lorenzo Bermejo. A comparative analysis of cell-type adjustment methods for epigenome-wide association studies based on simulated and real data sets. *Briefings in Bioinformatics*, 20(6):2055–2065, 2018.

[15] Johannes Brägelmann and Justo Lorenzo Bermejo. A comparative analysis of cell-type adjustment methods for epigenome-wide association studies based on simulated and real data sets. *Briefings in Bioinformatics*, 20(6):2055–2065, 2018.

[16] André M. Carrington, Paul W. Fieguth, Hammad Qazi, Andreas Holzinger, Helen H. Chen, Franz Mayr, and Douglas G. Manuel. A new concordant partial auc and partial c statistic for imbalanced data in the evaluation of machine learning algorithms. *BMC Medical Informatics and Decision Making*, 20(1), 2020.

[17] Phasit Charoenkwan, Wararat Chiangjong, Chanin Nantasenamat, Md Mehedi Hasan, Balachandran Manavalan, and Watshara Shoombuatong. Stackil6: A stacking ensemble model for improving the prediction of il-6 inducing peptides. *Briefings in Bioinformatics*, 22(6), 2021.

[18] Angelos Chatzimparmpas, Rafael M. Martins, Kostiantyn Kucher, and Andreas Kerren. Stackgenvis: Alignment of data, algorithms, and models for stacking ensemble learning using performance metrics. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1547–1557, 2021.

[19] Jun Chen, Ehsan Behnam, Jinyan Huang, Miriam F. Moffatt, Daniel J. Schaid, Liming Liang, and Xihong Lin. Fast and robust adjustment of cell mixtures in epigenome-wide association studies with smartsva. *BMC Genomics*, 18(1), 2017.

[20] Mark J Chen. Xreactive_probes: Cross-reactive probes/loci in markgene/maxprobes: Methylation array cross-reactive probes, May 2019.

[21] Xi Chen and Hemant Ishwaran. Random forests for genomic data analysis. *Genomics*, 99(6):323–329, 2012.

[22] Yi-an Chen, Mathieu Lemire, Sanaa Choufani, Darci T. Butcher, Daria Grafodatskaya, Brent W. Zanke, Steven Gallinger, Thomas J. Hudson, and Rosanna Weksberg. Discovery of cross-reactive probes and polymorphic cpgs in the illumina infinium humanmethylation450 microarray. *Epigenetics*, 8(2):203–209, 2013.

[23] Davide Chicco, Niklas Tötsch, and Giuseppe Jurman. The matthews correlation coefficient (mcc) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. *BioData Mining*, 14(1), 2021.

[24] S. Davis and P. S. Meltzer. Geoquery: A bridge between the gene expression omnibus (geo) and bioconductor. *Bioinformatics*, 23(14):1846–1847, 2007.

[25] Nicolas De Jay, Simon Papillon-Cavanagh, Catharina Olsen, Nehme El-Hachem, Gianluca Bontempi, and Benjamin Haibe-Kains. Mrmre: An r package for parallelized mrmr ensemble feature selection. *Bioinformatics*, 29(18):2365–2368, 2013.

[26] Lloyd A. Demetrius, Anne Eckert, and Amandine Grimm. Sex differences in alzheimer's disease: Metabolic reprogramming and therapeutic intervention. *Trends in Endocrinology amp; Metabolism*, 32(12):963–979, 2021.

[27] Yao Deng, Hao Wan, Jianbo Tian, Xiang Cheng, Meilin Rao, Jiaoyuan Li, Hongli Zhang, Ming Zhang, Yimin Cai, Zequn Lu, and et al. Cpg-methylation-based risk score predicts progression in colorectal cancer. *Epigenomics*, 12(7):605–615, 2020.

[28] Danielle Denisko and Michael M. Hoffman. Classification and interaction in random forests. *Proceedings of the National Academy of Sciences*, 115(8):1690–1692, 2018.

[29] Min Dong, Zengli Yang, Xingfang Li, Zhenxiang Zhang, and Ankang Yin. Screening of methylation gene sites as prognostic signature in lung adenocarcinoma. *Yonsei Medical Journal*, 61(12):1013, 2020.

[30] Pan Du, Xiao Zhang, Chiang-Ching Huang, Nadereh Jafari, Warren A Kibbe, Lifang Hou, and Simon M Lin. Comparison of beta-value and m-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics*, 11(1), 2010.

[31] R. Duara, D.A. Loewenstein, G. Lizarraga, M. Adjouadi, W.W. Barker, M.T. Greig-Custo, M. Rosselli, A. Penate, Y.F. Shea, R. Behar, and et al. Effect of age, ethnicity, sex, cognitive status and apoe genotype on amyloid load and the threshold for amyloid positivity. *NeuroImage: Clinical*, 22:101800, 2019.

[32] Yasser EL-Manzalawy, Tsung-Yu Hsieh, Manu Shivakumar, Dokyoon Kim, and Vasant Honavar. Min-redundancy and max-relevance multi-view feature selection for predicting ovarian cancer survival using multi-omics data. *BMC Medical Genomics*, 11(S3), 2018.

[33] Wei Fang, Zhen-Zong Fa, Qun Xie, Gui-Zhen Wang, Jiu Yi, Chao Zhang, Guang-Xun Meng, Ju-Lin Gu, and Wan-Qing Liao. Complex roles of annexin a2 in host blood-brain barrier invasion by cryptococcus neoformans. *CNS Neuroscience amp; Therapeutics*, 23(4):291–300, 2017.

[34] Yuemei Feng, Guanzhang Li, Zhongfang Shi, Xu Yan, Zhiliang Wang, Haoyu Jiang, Ye Chen, Renpeng Li, You Zhai, Yuanhao Chang, and et al. A novel methylation signature predicts radiotherapy sensitivity in glioma. *Scientific Reports*, 10(1), 2020.

[35] Peter Daniel Fransquet, Paul Lacaze, Richard Saffery, James Phung, Emily Parker, Raj Shah, Anne Murray, Robyn L. Woods, and Joanne Ryan. Blood dna methylation signatures to detect dementia prior to overt clinical symptoms. *Alzheimer's amp; Dementia: Diagnosis, Assessment amp; Disease Monitoring*, 12(1), 2020.

[36] Sharon L Freshour, Susanna Kiwala, Kelsy C Cotto, Adam C Coffman, Joshua F McMichael, Jonathan J Song, Malachi Griffith, Obinbsp;L Griffith, and Alex H Wagner. Integration of the drug–gene interaction database (dgidb 4.0) with open crowdsource efforts. *Nucleic Acids Research*, 49(D1), 2020.

[37] Tamas Fulop, Jacek M. Witkowski, Karine Bourgade, Abdelouahed Khalil, Echarki Zerif, Anis Larbi, Katsuiku Hirokawa, Graham Pawelec, Christian Bocti, Guy Lacombe, and et al. Can an infection hypothesis explain the beta amyloid hypothesis of alzheimer's disease? *Frontiers in Aging Neuroscience*, 10, 2018.

[38] Alessandra Gallo, Laure-Elise Pillet, and Romain Verpillot. New frontiers in alzheimer's disease diagnostic: Monoamines and their derivatives in biological fluids. *Experimental Gerontology*, 152:111452, 2021.

[39] Shu Gong, Weijian Ye, Tiankai Liu, Shaofen Jian, and Wenhua Liu. The development of three-dna methylation signature as a novel prognostic biomarker in patients with colorectal cancer. *BioMed Research International*, 2020:1–11, 2020.

[40] Viivi Halla-aho and Harri Lähdesmäki. Luxus: Dna methylation analysis us-

ing generalized linear mixed model with spatial correlation. *Bioinformatics*, 36(17):4535–4543, 2020.

[41] Kasper Hansen. Illuminahumanmethylation450kanno.ilmn12.hg19: Annotation for illumina's 450k methylation arrays. r package version 0.6.0. 2016.

[42] Xiang-Yong Hao, An-Qiang Li, Hao Shi, Tian-Kang Guo, Yan-Fei Shen, Yuan Deng, Li-Tian Wang, Tao Wang, and Hui Cai. A novel dna methylation-based model that effectively predicts prognosis in hepatocellular carcinoma. *Bioscience Reports*, 41(3), 2021.

[43] Jonathan A. Heiss and Allan C. Just. Improved filtering of dna methylation microarray data by detection p values and its impact on downstream analyses. *Clinical Epigenetics*, 11(1), 2019.

[44] Peter Henneman, Arjan Bouman, Adri Mul, Lia Knegt, Anne-Marie van der Kevie-Kersemaekers, Nitash Zwaveling-Soonawala, Hanne E. Meijers-Heijboer, A. S. van Trotsenburg, and Marcel M. Mannens. Widespread domain-like perturbations of dna methylation in whole blood of down syndrome neonates. *PLOS ONE*, 13(3), 2018.

[45] Yanting Huang, Xiaobo Sun, Huige Jiang, Shaojun Yu, Chloe Robins, Matthew J. Armstrong, Ronghua Li, Zhen Mei, Xiaochuan Shi, Ekaterina Sergeevna Gerasimov, and et al. A machine learning approach to brain epigenetic analysis reveals kinases associated with alzheimer's disease. *Nature Communications*, 12(1), 2021.

[46] Zhang Hui, Yuan Zhijun, Yan Yushan, Chen Liping, Zhou Yiying, Zhang Difan, Choi Tony Chunglit, and Cui Wei. The combination of acyclovir and dexamethasone protects against alzheimer's disease-related cognitive impairments in mice. *Psychopharmacology*, 237(6):1851–1860, 2020.

[47] Hao Jia, Ben Huang, Le Kang, Hao Lai, Jun Li, Chunsheng Wang, and Yongxin Sun. Preoperative and intraoperative risk factors of postoperative stroke in total

aortic arch replacement and stent elephant trunk implantation. *eClinicalMedicine*, 47:101416, 2022.

[48] Philipp Jurmeister, Michael Bockmayr, Philipp Seegerer, Teresa Bockmayr, Denise Treue, Grégoire Montavon, Claudia Vollbrecht, Alexander Arnold, Daniel Teichmann, Keno Bressem, and et al. Machine learning analysis of dna methylation profiles distinguishes primary lung squamous cell carcinomas from head and neck metastases. *Science Translational Medicine*, 11(509), 2019.

[49] Sanaz Khalili, Javad Faradmal, Hossein Mahjub, Babak Moeini, and Khadijeh Ezzati-Rastegar. Overcoming the problems caused by collinearity in mixed-effects logistic model: Determining the contribution of various types of violence on depression in pregnant women. *BMC Medical Research Methodology*, 21(1), 2021.

[50] Carl Kingsford and Steven L Salzberg. What are decision trees? *Nature Biotechnology*, 26(9):1011–1013, 2008.

[51] Nobuyuki Kobayashi, Shunichiro Shinagawa, Hidehito Niimura, Hisashi Kida, Tomoyuki Nagata, Kenji Tagai, Kazuya Shimada, Naomi Oka, Ryo Shikimoto, Yoshihiro Noda, and et al. Increased blood coasy dna methylation levels a potential biomarker for early pathology of alzheimer's disease. *Scientific Reports*, 10(1), 2020.

[52] Han Koh, SangJoon Lee, Hyo Lee, Jae-Woong Min, Takeshi Iwatsubo, Charlotte Teunissen, Hyun-Jeong Cho, and Jin-Hyeob Ryu. Targeting microrna-485-3p blocks alzheimer's disease progression. *International Journal of Molecular Sciences*, 22(23):13136, 2021.

[53] Mikko Konki, Maia Malonzo, Ida K. Karlsson, Noora Lindgren, Bishwa Ghimire, Johannes Smolander, Noora M. Scheinin, Miina Ollikainen, Asta Laiho, Laura L. Elo, and et al. Peripheral blood dna methylation differences in twin pairs discordant for alzheimer's disease. *Clinical Epigenetics*, 11(1), 2019.

[54] Taesic Lee and Hyunju Lee. Prediction of alzheimer's disease using blood gene expression data. *Scientific Reports*, 10(1), 2020.

[55] Jeffrey T Leek and John D Storey. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genetics*, 3(9), 2007.

[56] Guillaume Lemaître, Fernando Nogueira, and Christos K. Aridas. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18(17):1–5, 2017.

[57] Joshua J. Levy, Alexander J. Titus, Curtis L. Petersen, Youdinghuan Chen, Lucas A. Salas, and Brock C. Christensen. Methylnet: An automated and modular deep learning approach for dna methylation analysis. *BMC Bioinformatics*, 21(1), 2020.

[58] Dong-hai Li, Xiao-hui Du, Ming Liu, and Rui Zhang. A 10-gene-methylation-based signature for prognosis prediction of colorectal cancer. *Cancer Genetics*, 252-253:80–86, 2021.

[59] Hai-Tao Li, Shao-Xun Yuan, Jian-Sheng Wu, Yu Gu, and Xiao Sun. Predicting conversion from mci to ad combining multi-modality data and based on molecular subtype. *Brain Sciences*, 11(6):674, 2021.

[60] Qingqin S. Li, Aparna Vasanthakumar, Justin W. Davis, Kenneth B. Idler, Kwangsik Nho, Jeffrey F. Waring, and Andrew J. Saykin. Association of peripheral blood dna methylation level with alzheimer's disease progression. *Clinical Epigenetics*, 13(1), 2021.

[61] Ziyi Li, Xiaoqian Jiang, Yizhuo Wang, and Yejin Kim. Applied machine learning in alzheimer's disease research: Omics, imaging, and clinical data. *Emerging Topics in Life Sciences*, 5(6):765–777, 2021.

[62] Eliana Lima, Peers Davies, Jasmeet Kaler, Fiona Lovatt, and Martin Green. Variable selection for inferential models with relatively high-dimensional data: Between

method heterogeneity and covariate stability as adjuncts to robust selection. *Scientific Reports*, 10(1), 2020.

[63] Rui-Ming Liu. Aging, cellular senescence, and alzheimer's disease. *International Journal of Molecular Sciences*, 23(4):1989, 2022.

[64] Simon Lovestone, Paul Francis, Iwona Kloszewska, Patrizia Mecocci, Andrew Simmons, Hilkka Soininen, Christian Spenger, Magda Tsolaki, Bruno Vellas, Lars-Olof Wahlund, and et al. Addneuromed-the european collaboration for the discovery of novel biomarkers for alzheimer's disease. *Annals of the New York Academy of Sciences*, 1180(1):36–46, 2009.

[65] Jovana Maksimovic, Belinda Phipson, and Alicia Oshlack. A cross-package bioconductor workflow for analysing methylation array data. *F1000Research*, 5:1281, 2016.

[66] Georgina Mansell, Tyler J. Gorrie-Stone, Yanchun Bao, Meena Kumari, Leonard S. Schalkwyk, Jonathan Mill, and Eilis Hannon. Guidance for dna methylation studies: Statistical insights from the illumina epic array. *BMC Genomics*, 20(1), 2019.

[67] Kevin McGregor, Sasha Bernatsky, Ines Colmegna, Marie Hudson, Tomi Pastinen, Aurélie Labbe, and Celia M.T. Greenwood. An evaluation of methods correcting for cell-type heterogeneity in dna methylation studies. *Genome Biology*, 17(1), 2016.

[68] Erika L Moen, Edward Litwin, Stephen Arnovitz, Xu Zhang, Wei Zhang, M Eileen Dolan, and Lucy A Godley. Characterization of cpg sites that escape methylation on the inactive human x-chromosome. *Epigenetics*, 10(9):810–818, 2015.

[69] Erik S Musiek and David M Holtzman. Three dimensions of the amyloid hypothesis: Time, space and 'wingmen'. *Nature Neuroscience*, 18(6):800–806, 2015.

[70] Fatema Yasmin Nisa, Md. Atiar Rahman, Md. Amjad Hossen, Mohammad Forhad Khan, Md. Asif Khan, Mumtahina Majid, Farjana Sultana, and Md. Areeful Haque. Role of neurotoxicants in the pathogenesis of alzheimer's disease: A mechanistic insight. *Annals of Medicine*, 53(1):1479–1504, 2021.

[71] Bodo Parady. Innate immune and fungal model of alzheimer's disease. *Journal of Alzheimer's Disease Reports*, 2(1):139–152, 2018.

[72] Saeid Parvandeh, Hung-Wen Yeh, Martin P Paulus, and Brett A McKinney. Consensus features nested cross-validation. *Bioinformatics*, 36(10):3093–3098, 2020.

[73] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.*, 12(null):2825–2830, nov 2011.

[74] Hanchuan Peng, Fuhui Long, and C. Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1226–1238, 2005.

[75] Yi-Hao Peng, Chih-Yu Chen, Ching-Hua Su, Chih-Hsin Muo, Kuan-Fei Chen, Wei-Chih Liao, and Chia-Hung Kao. Increased risk of dementia among patients with pulmonary tuberculosis. *American Journal of Alzheimer's Disease amp; Other Dementiasr*, 30(6):629–634, 2015.

[76] Timothy J Peters, Michael J Buckley, Aaron L Statham, Ruth Pidsley, Katherine Samaras, Reginald V Lord, Susan J Clark, and Peter L Molloy. De novo identification of differentially methylated regions in the human genome. *Epigenetics amp; Chromatin*, 8(1), 2015.

[77] Belinda Phipson, Jovana Maksimovic, and Alicia Oshlack. Missmethyl: An r package for analyzing data from illumina's humanmethylation450 platform. *Bioinformatics*, 32(2):286–288, 2015.

[78] Dennis Pischel, Jörn H. Buchbinder, Kai Sundmacher, Inna N. Lavrik, and Robert J. Flassig. A guide to automated apoptosis detection: How to make sense of imaging flow cytometry data. *PLOS ONE*, 13(5), 2018.

[79] E. M. Price and Wendy P. Robinson. Adjusting for batch effects in dna methylation microarray data, a lesson learned. *Frontiers in Genetics*, 9, 2018.

[80] Raúl Fernández Pérez, Juan José Alba-Linares, Juan Ramón Tejedor, Agustín Fernández Fernández, Miguel Calero, Aurora Román-Domínguez, Consuelo Borrás, José Viña, Jesús Ávila, Miguel Medina, and et al. Blood dna methylation patterns in older adults with evolving dementia. *The Journals of Gerontology: Series A*, 2022.

[81] Talal Jamil Qazi, Zhenzhen Quan, Asif Mir, and Hong Qing. Epigenetics in alzheimer's disease: Perspective of dna methylation. *Molecular Neurobiology*, 55(2):1026–1044, 2017.

[82] Milos Radovic, Mohamed Ghalwash, Nenad Filipovic, and Zoran Obradovic. Minimum redundancy maximum relevance feature selection approach for temporal gene expression data. *BMC Bioinformatics*, 18(1), 2017.

[83] Sebastian Rauschert, Phillip E. Melton, Anni Heiskala, Ville Karhunen, Graham Burdge, Jeffrey M. Craig, Keith M. Godfrey, Karen Lillycrop, Trevor A. Mori, Lawrence J. Beilin, and et al. Machine learning-based dna methylation score for fetal exposure to maternal smoking: Development and validation in samples collected from adolescents and adults. *Environmental Health Perspectives*, 128(9):097003, 2020.

[84] Sebastian Rauschert, Phillip E. Melton, Anni Heiskala, Ville Karhunen, Graham Burdge, Jeffrey M. Craig, Keith M. Godfrey, Karen Lillycrop, Trevor A. Mori, Lawrence J. Beilin, and et al. Machine learning-based dna methylation score for fetal exposure to maternal smoking: Development and validation in samples collected from adolescents and adults. *Environmental Health Perspectives*, 128(9):097003, 2020.

[85] Matthew E. Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, Wei

Shi, and Gordon K. Smyth. Limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic Acids Research*, 43(7), 2015.

[86] Janou A.Y. Roubroeks, Adam R. Smith, Rebecca G. Smith, Ehsan Pishva, Zina Ibrahim, Martina Sattlecker, Eilis J. Hannon, Iwona Kłoszewska, Patrizia Mecocci, Hilkka Soininen, and et al. An epigenome-wide association study of alzheimer's disease blood highlights robust dna hypermethylation in the hoxb6 gene. *Neurobiology of Aging*, 95:26–45, 2020.

[87] Takaya Saito and Marc Rehmsmeier. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, 10(3), 2015.

[88] Claudia Sala, Pietro Di Lena, Danielle Fernandes Durso, Andrea Prodi, Gastone Castellani, and Christine Nardini. Evaluation of pre-processing on the meta-analysis of dna methylation data from the illumina humanmethylation450 beadchip platform. *PLOS ONE*, 15(3), 2020.

[89] Nahid Sarahian, Hosna Sarvazad, Elham Sajadi, Nasrin Rahnejat, and Narges Eskandari Roozbahani. Investigation of common risk factors between polycystic ovary syndrome and alzheimer's disease: A narrative review. *Reproductive Health*, 18(1), 2021.

[90] Alessia Sarica, Antonio Cerasa, and Aldo Quattrone. Random forest algorithm for the classification of neuroimaging data in alzheimer's disease: A systematic review. *Frontiers in Aging Neuroscience*, 9, 2017.

[91] Dustin Scheinost, Stephanie Noble, Corey Horien, Abigail S. Greene, Evelyn MR. Lake, Mehraveh Salehi, Siyuan Gao, Xilin Shen, David O'Connor, Daniel S. Barron, and et al. Ten simple rules for predictive modeling of individual differences in neuroimaging. *NeuroImage*, 193:35–45, 2019.

[92] Madhava Vishwanath Shervegar and Ganesh V. Bhat. Heart sound classification using gaussian mixture model. *Porto Biomedical Journal*, 3(1), 2018.

[93] Mingchao Shi, Chunrong Li, Xiaoping Tian, Fengna Chu, and Jie Zhu. Can control infections slow down the progression of alzheimer's disease? talking about the role of infections in alzheimer's disease. *Frontiers in Aging Neuroscience*, 13, 2021.

[94] Daichi Shigemizu, Taiki Mori, Shintaro Akiyama, Sayuri Higaki, Hiroshi Watanabe, Takashi Sakurai, Shumpei Niida, and Kouichi Ozaki. Identification of potential blood biomarkers for early diagnosis of alzheimer's disease through rna sequencing analysis. *Alzheimer's Research amp; Therapy*, 12(1), 2020.

[95] Daichi Shigemizu, Taiki Mori, Shintaro Akiyama, Sayuri Higaki, Hiroshi Watanabe, Takashi Sakurai, Shumpei Niida, and Kouichi Ozaki. Identification of potential blood biomarkers for early diagnosis of alzheimer's disease through rna sequencing analysis. *Alzheimer's Research amp; Therapy*, 12(1), 2020.

[96] Vikash Singh, Michael Pencina, Andrew J. Einstein, Joanna X. Liang, Daniel S. Berman, and Piotr Slomka. Impact of train/test sample regimen on performance estimate stability of machine learning in cardiovascular imaging. *Scientific Reports*, 11(1), 2021.

[97] Annette Spooner, Emily Chen, Arcot Sowmya, Perminder Sachdev, Nicole A. Kochan, Julian Trollor, and Henry Brodaty. A comparison of machine learning methods for survival analysis of high-dimensional clinical data for dementia prediction. *Scientific Reports*, 10(1), 2020.

[98] Daniel Stamate, Wajdi Alghambdi, Jeremy Ogg, Richard Hoile, and Fionn Murtagh. A machine learning framework for predicting dementia and mild cognitive impairment. *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2018.

[99] Jamal Stie and Deborah Fox. Blood–brain barrier invasion by cryptococcus neoformans is enhanced by functional interactions with plasmin. *Microbiology*, 158(1):240–258, 2012.

[100] P SVENNINGSSON and P GREENGARD. P11 (s100a10) — an inducible adaptor protein that modulates neuronal functions. *Current Opinion in Pharmacology*, 7(1):27–32, 2007.

[101] Junpei Takeishi, Yasuko Tatewaki, Taizen Nakase, Yumi Takano, Naoki Tomita, Shuzo Yamamoto, Tatsushi Mutoh, and Yasuyuki Taki. Alzheimer's disease and type 2 diabetes mellitus: The use of mct oil and a ketogenic diet. *International Journal of Molecular Sciences*, 22(22):12310, 2021.

[102] Jiayi Tang, Alex Henderson, and Peter Gardner. Exploring adaboost and random forests machine learning approaches for infrared pathology on unbalanced data sets. *The Analyst*, 146(19):5880–5891, 2021.

[103] Mei-Yun Tang, Fredric A. Gorin, and Pamela J. Lein. Review of evidence implicating the plasminogen activator system in blood-brain barrier dysfunction associated with alzheimer's disease. *Ageing and Neurodegenerative Diseases*, 2022.

[104] Julia R. Taylor, Joseph G. Skeate, and W. Martin Kast. Annexin a2 in virus infection. *Frontiers in Microbiology*, 9, 2018.

[105] Andrius Vabalas, Emma Gowen, Ellen Poliakoff, and Alexander J. Casson. Machine learning algorithm validation with a limited sample size. *PLOS ONE*, 14(11), 2019.

[106] Mohammad Amin Valizade Hasanloei, Razieh Sheikhpour, Mehdi Agha Sarram, Elnaz Sheikhpour, and Hamdollah Sharifi. A combined fisher and laplacian score for feature selection in qsar based drug design using compounds with known and unknown activities. *Journal of Computer-Aided Molecular Design*, 32(2):375–384, 2017.

[107] Yogatheesan Varatharajah, Vijay K. Ramanan, Ravishankar Iyer, and Prashanthi Vemuri. Predicting short-term mci-to-ad progression using imaging, csf, genetic factors, cognitive resilience, and demographics. *Scientific Reports*, 9(1), 2019.

[108] Alexander Vezhnevets and Olga Barinova. Avoiding boosting overfitting by removing confusing samples. *Machine Learning: ECML 2007*, page 430–441.

[109] Hannah Walgrave, Lujia Zhou, Bart De Strooper, and Evgenia Salta. The promise of microrna-based therapies in alzheimer's disease: Challenges and perspectives. *Molecular Neurodegeneration*, 16(1), 2021.

[110] Xuelong Wang, Bin Zhou, Yuxin Xia, Jianxin Zuo, Yanchao Liu, Xin Bi, Xiong Luo, and Chengwei Zhang. A methylation-based nomogram for predicting survival in patients with lung adenocarcinoma. *BMC Cancer*, 21(1), 2021.

[111] Xueli Wei, Le Zhang, and Yi Zeng. Dna methylation in alzheimer's disease: In brain and peripheral blood. *Mechanisms of Ageing and Development*, 191:111319, 2020.

[112] Xueli Wei, Le Zhang, and Yi Zeng. Dna methylation in alzheimer's disease: In brain and peripheral blood. *Mechanisms of Ageing and Development*, 191:111319, 2020.

[113] C S Wilhelm-Benartzi, D C Koestler, M R Karagas, J M Flanagan, B C Christensen, K T Kelsey, C J Marsit, E A Houseman, and R Brown. Review of processing and analysis methods for dna methylation array data. *British Journal of Cancer*, 109(6):1394–1402, 2013.

[114] Min Xu, Deng[U+2010]Feng Zhang, Rongcan Luo, Yong Wu, Hejiang Zhou, Li[U+2010]Li Kong, Rui Bi, and Yong[U+2010]Gang Yao. A systematic integrated analysis of brain expression profiles reveals yap1 and other prioritized hub genes as important upstream regulators in alzheimer's disease. *Alzheimer's Dementia*, 14(2):215–229, 2017.

[115] Youjun Xu, Shiwei Wang, Qiwan Hu, Shuaishi Gao, Xiaomin Ma, Weilin Zhang, Yihang Shen, Fangjin Chen, Luhua Lai, Jianfeng Pei, and et al. Cavityplus: A web server for protein cavity detection with pharmacophore modelling, allosteric site identification and covalent ligand binding ability prediction. *Nucleic Acids Research*, 46(W1), 2018.

[116] Weishuang Xue, Jinwei Li, Kailei Fu, and Weiyu Teng. Differential expression of mrnas in peripheral blood related to prodrome and progression of alzheimer's disease. *BioMed Research International*, 2020:1–10, 2020.

[117] J. Yang, Y.L. Liu, C.S. Feng, and G.Q. Zhu. Applying the fisher score to identify alzheimer's disease-related genes. *Genetics and Molecular Research*, 15(2), 2016.

[118] Hao Yu and Jie Wu. Amyloid-: A double agent in alzheimer's disease? *Biomedicine amp; Pharmacotherapy*, 139:111575, 2021.

[119] Yaxia Yuan, Jianfeng Pei, and Luhua Lai. Binding site detection and druggability prediction of protein targets for structure- based drug design. *Current Pharmaceutical Design*, 19(12):2326–2333, 2013.

[120] Lanbo Zhao, Sijia Ma, Linconghua Wang, Yiran Wang, Xue Feng, Dongxin Liang, Lu Han, Min Li, and Qiling Li. A polygenic methylation prediction model associated with response to chemotherapy in epithelial ovarian cancer. *Molecular Therapy - Oncolytics*, 20:545–555, 2021.

[121] Lanbo Zhao, Sijia Ma, Linconghua Wang, Yiran Wang, Xue Feng, Dongxin Liang, Lu Han, Min Li, and Qiling Li. A polygenic methylation prediction model associated with response to chemotherapy in epithelial ovarian cancer. *Molecular Therapy - Oncolytics*, 20:545–555, 2021.